DIGIECOQUARRY

INNOVATIVE DIGITAL SUSTAINABLE
AGGREGATES SYSTEMS

Horizon 2020 research
and innovation programme
(Nº 101003750)

Deliverable D4.1.

Report on IQS ICT requirement analysis

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement Nº 101003750

Deliverable report

Deliverable No.	D4.5	Work Package No.	WP4	Task/s No.	Task 4.1
Work Package Title		Development of an integrated IoT/BIM/AI platform for smart quarrying [KTA4]			
Linked Task/s Title		ICT requirements analysis and assets inventory			
Status		Draft Final	(Draft/Draft Final/Final)		
Dissemination level		PU	(PU-Public, PP, RE-Restricted, CO-Confidential)		
Due date deliverable		2022-07-31	Submission date		2022-07-28
Deliverable version		DIGIECOQUARRY_D4.1_Report_IQS_ICT_requirement_analysis_1.0_Final.docx			

Document Contributors

Deliverable responsible	**AKKA**
Contributors	Organisation
DALET Benoît, DIALLO Abdoul-Gadiri, GERMENIS Evangelos, MADANI Radwane, MARTY Paul, YAR Anne-Gaëlle, ZOUGARI Sadeq	AKKA
DÖPPENSCHMITT Simon, BROECKMANN Frank	DHP
Pierre Plaza, Jorge Rico, César Pérez, Javier Gavilanes	SIGMA
Diego Laza	abaut
Paulo Romero	ANEFA
Petrus Van Staden, Parisa Doubra	MINTEK
Tuomo Pirinen	SANDVIK
Juan Navarro Miguel	MAXAM
Pablo Gómez-C. Martín, Sadik Serdar Tekin Mevlüt Tuna	APP (APP Consultoría)
Pablo Segarra, José A. Sanchidrián	UPM-M
Jesse Backman	METSO
José Luis Blanco, José Eugenio Ortiz	UPM-AI
Reviewers	Organisation
Asim Jafa, Fernando Maria Beitia Gomez de Segura, Juan Navarro Miguel, Paulo Jose Costa Couceiro, Vicente José Huelamo	MAXAM
Michel Zablocki, Lara Maëlla Bikanda	VICAT
Lorena Viladés	ANEFA

Document History

Version	Date	Comment
1.0_Draft1	2022-02-10	Creation of the document
1.0_Draft2	2022-07-08	Updated version for internal and external review
1.0_Final Draft	2022-07-27	Final draft after peer reviews, document ready for submission
1.0_Final	2022-07-28	Final document after second review, document ready for submission

Disclaimer

This document reflects only the author's view. Responsibility for the information and views expressed therein lies entirely with the authors. The European Commission are not responsible for any use that may be made of the information it contains.

Table of contents

List of Abbreviations

Abbreviation	Description
AI	Artificial Intelligence
API	Application Programming Interface
AWS	Amazon Web Services
BI	Business Intelligence
BIM	Building Information Modelling
BMT	Business Management tools
CDE	Common Data Environment
CDMP	Centralized DEQ Data Management Platform
CPU	Central Processing Unit
DEQ	DIGIECOQUARRY
DIU	Data Interface Unit
ELK	ElasticSearch, Logstash, Kibana
ES	Expert System
ETL	Extract Transform Load
GA	Grant Agreement
HA	High Availability
HDD	Hard Disk Drive
HMI	Human-Machine Interface
HTTP	Hyper Text Transfer Protocol
ICT	Information and communication technology
IoT	Internet of Things
IQS	Intelligent Quarrying System
KPI	Key Performance Indicator
KTA	Key Technology Area
LAN	Local Area Network
LDAP	Light Directory Access Protocol
ML	Machine Learning
MWD	Measurement While Drilling
N/A or NA	not applicable
Nb	Number

Abbreviation	Description
OS	Operating System
Paas	Platform as a service
PSD	Particle Size Distribution
RAM	Random Access Memory
RDBMS	Relational DataBase Management System
REST-API	Representational State Transfer-API
SFTP	Secure File Transfer Protocol
SQL	Structured Query Language
SSD	Solid State Drive
VM	Virtual Machine
VNET	Velocity Networking Execution Technology
WAF	Web Application Firewall
WP	Work Package
Unit	Description
€ / $	Euro / Dollar
Go = Gb	Gibabyte
Ko = Kb	Kilobyte
Mo = Mb	Megabyte
Mn	Minutes
s	Seconds
To = Tb	Terabyte
File Format	Description
.csv (CSV)	Comma-separated values, delimited text file
.jsn (JSON)	JavaScript Object Notation, Data interchange format
.xml (XML)	Extensible Markup Language

1 Executive Summary

This document reports the results of the IQS ICT requirements analysis done by the partners involved in the task 4.1 (ICT requirements analysis and assets inventory) Those results are the main inputs for the development of an integrated IoT/BIM/AI platform for smart quarrying (KTA4) that will be done in the frame of the WP4.

Firstly, the WP1's deliverables and the D3.1 (List and characterization of key data inputs) were deeply analyzed to produce an exhaustive ICT assets inventory, known at this stage of the DigiEcoQuarry project, for all the pilot sites. These inventories list the expert systems and the interfaces, give a data contents summary, and highlight the data format and the data sharing within each pilot site and for all involved partners.

Secondly, their analysis, completed by several exchanges and workshops between partners, also enable the creation of the data flow diagrams for each pilot site. These diagrams permit to identify the necessary configurations of the interfaces to build to connect the IQS with the pilot sites and partners expert systems.

Main activity of this task was also the realization of a benchmark study allowing the selection of the best components and tools that will be used to build the IQS. This document gives the conclusions of the benchmark (in appendix, the whole study is also available)

Finally, all the intended components that will be used for the data lake, the IoT, the data warehouse platforms and for the business management tools are listed, costed, and presented here. The sharing of first dataset examples between the partners enabled the realization of first prototypes. Thanks to these prototypes, certain risks could be eliminated, the choice of components and tools could be confirmed, and a global IQS integration could be defined.

Through the sharing of these dataset examples, it has also been possible to create a first version of data models, by quarrying process, that seem to be relevant for the aggregates industry. These data models are also presented within this document.

2 Introduction

2.1 Concept/Approach

The D4.1 deliverable is the main output of the Task 4.1, ICT requirements analysis and assets inventory, run in the frame of the WP4, Development of an integrated IoT/BIM/AI platform for smart quarrying [KTA4] led by AKKA, and involving the following other partners: ANEFA, Sandvik, Metso, Maxam, ITK, MUL, Chalmers, UPM-M, Abaut GmbH DH&P, ROCTIM, SIGMA, UPM-AI, Ma-estro SRL, ARCO and APP Consultoría.

Within this Task 4.1, each technological partner had the opportunity to present in more details its key technology area and their related tools to all the project stakeholders. Several bilateral workshops have been organized with the pilot sites and between the technological partners to go deeply in the details of all the ICT requirements described within WP1's deliverables. These workshops allowed the partners to gradually build the inventory of the existing ICT assets of each pilot site and to define what could be deployed, and how, on the quarries, to fulfil their digitalisation needs. A benchmark has also been performed to select the best digitalisation tools (data lake, IoT platform elements and data warehouse) by considering the state of the art, defining evaluation criteria, and identifying potential solutions. All these, workshops conclusions, benchmark results and potential solutions are presented in the next sections of this deliverable.

2.2 Deliverable objectives

The objectives of the D4.1 deliverable are to describe:

- The solutions to be deployed in the quarries: networks, devices, tools, and architectures
- The data flows between these solutions
- The related/proposed data models
- The needed interfaces
- The additional services related to:
    - The IoT platform and the data lakes,
    - The data warehouse and the AI system,
    - The BIM systems,
    - The reporting and management tools

and to explain how to integrate these.

2.3 Intended audience

The dissemination level of this deliverable is public.

This deliverable is a key input for all the other tasks to be done in the frame of the WP2: Selection and development of innovative aggregates processing techniques [KTA1, KTA2], WP3: Development of sensors, automation, and process control [KTA3] and WP4: Development of an integrated IoT/ BIM/AI platform for smart quarrying [KTA4].

3 ICT requirements analysis

3.1 Networks and devices deployed in the quarries

3.1.1 HANSON

3.1.1.1 Inventory of the existing ICT assets

The following table provides a high-level view of the expert systems, interfaces, contents summary, format, data sharing, and partners involved within this site. It enables the creation of pilot site's data flow. Please refer to D3.1 to have a detailed view of the data.

System Expert	**Description of the function**	**Interface type provided**	**Content**	**Format**	**Shared data through**	**Shared data with**
HANSON Expert system: COPA, AOM/IoT system	Quarry management system	Manual upload	Historical data Production data Maintenance data General information data Specific data (amount of material at the bypass of the crusher)	xls, pdf	Data Lake	BMT SIGMA APP ABAUT MUL UPM-M MAXAM SANDVIK
SANDVIK's cloud platform	Data measurement during the drilling process	Sandvik OEM cloud with an API, Manual upload, and download	MWD Signals	IREDES (xml) Json, csv	Data Lake	MAXAM, SANDVIK, MUL, HANSON, UPM-M, ABAUT SIGMA: Hawkeye APP
MAXAM's Blast Design software, RIOBLAST	blast design optimization automatic assessment of rock structure. Explosive performance assessment Borehole condition and resulting advance control of the blast results including rock damage assessment	Manual	Reports	csv, xls	Data Lake	MAXAM, SANDVIK, MUL, HANSON, UPM-M, ABAUT SIGMA: Hawkeye APP
UPM-M	Quality distributions using UAV-made block models Rock mass characterization techniques.	Manual	SHARED INPUTS • UAV photogrammetric acquisition • Internal hole wall video	Standard/Proprietary	Data Lake	MAXAM, SANDVIK, MUL, HANSON, UPM-M, ABAUT SIGMA: Hawkeye APP

System Expert	Description of the function	Interface type provided	Content	Format	Shared data through	Shared data with
			• Log of vibration signal • Detonation pressure SHARED OUTPUTS • Cloud points • 3D geo-ref model • Seismic propagation velocities • Seismic quality factors of the rock mass • Fracturing index			
MUL	implementation of a drill to mill concept. cost/efficiency analysis in order to optimize the blasting procedure used	Manual upload and download	• Particle size distribution • Muck pile characteristics • Quality (rock type, hardness) • Experimental setup (layout, explosives, delay time)	Json, csv	Data Lake	MAXAM, SANDVIK, MUL, HANSON, UPM-M, ABAUT SIGMA: Hawkeye APP
ABAUT	Product mass flow Fleet performance Reports Implementation of drill to mill concept	Manual and automatic upload/download	• Work time of Pecker • Production [sum of tonnage, tons/h] per machine and locations • Geofence • Cycle times • Duration loading/hauling/unloading/idling • Number of cycles • Haulage distance • Number of passes/scoops for loading a truck • Loading performance • Recognition of environment using cameras	Standard/Proprietary	Data shared in Abaut expert system and in	MAXAM, SANDVIK, MUL, HANSON, UPM-M, ABAUT SIGMA: Hawkeye APP

System Expert	Description of the function	Interface type provided	Content	Format	Shared data through	Shared data with
			• Recognition of activities using cameras			
BMT	Generate, store and share reports and dashboard	manual upload	• Dynamic and static view of data and KPIs shared	pdf, xls	Data Lake	HANSON

3.1.1.1.1 Data flow

The following diagram depicts the data flows between the partners or systems within this pilot site.

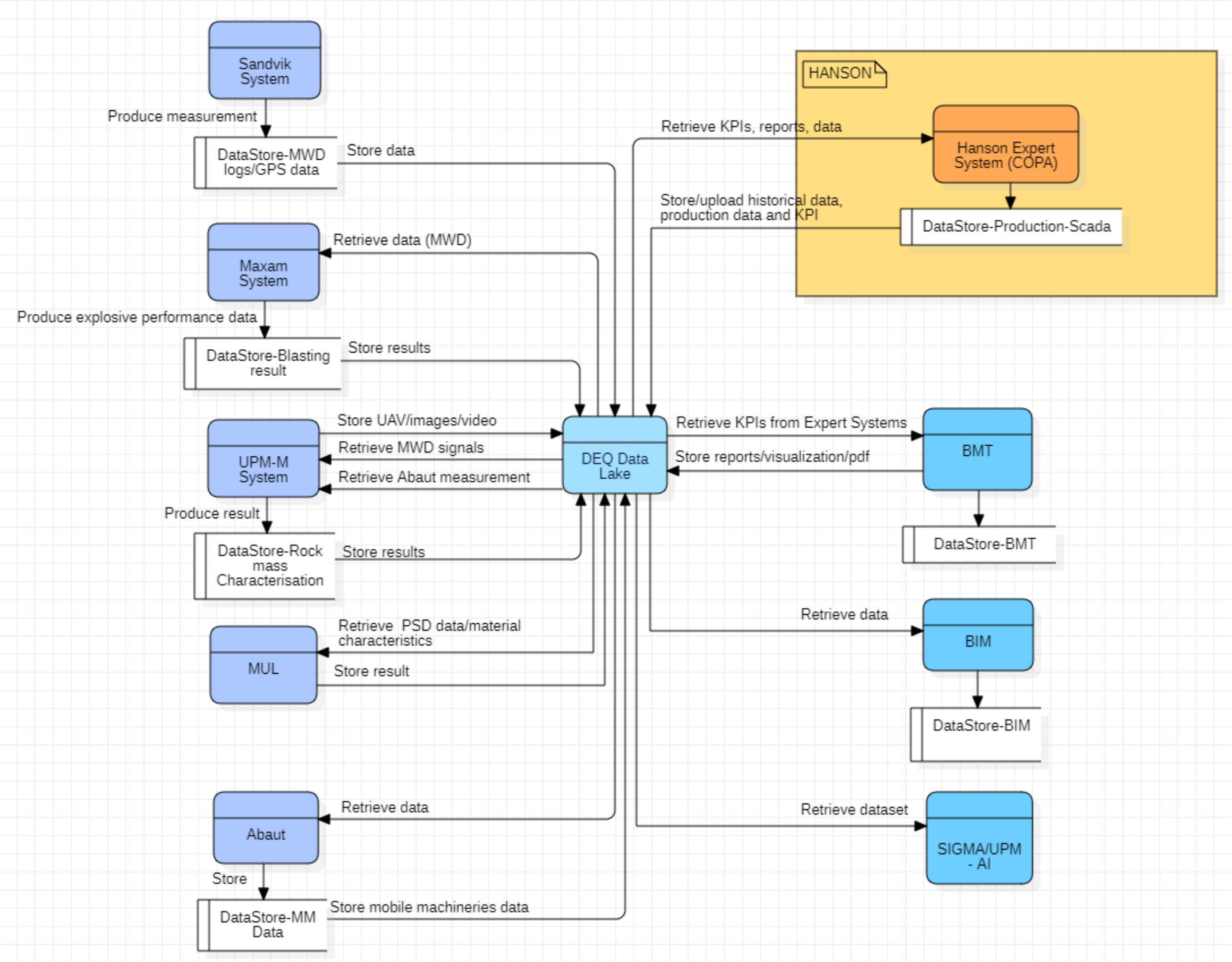

Sandvik System
Produce measurement
DataStore-MWD logs/GPS data
Store data
Maxam System
Retrieve data (MWD)
Produce explosive performance data
DataStore-Blasting result
Store results
UPM-M System
Store UAV/images/video
Retrieve MWD signals
Retrieve Abaut measurement
Produce result
DataStore-Rock mass Characterisation
Store results
MUL
Retrieve PSD data/material characteristics
Store result
Abaut
Retrieve data
Store
DataStore-MM Data
Store mobile machineries data
DEQ Data Lake
HANSON
Retrieve KPIs, reports, data
Hanson Expert System (COPA)
Store/upload historical data, production data and KPI
DataStore-Production-Scada
Retrieve KPIs from Expert Systems
Store reports/visualization/pdf
BMT
DataStore-BMT
Retrieve data
BIM
DataStore-BIM
Retrieve dataset
SIGMA/UPM - AI

Figure 1: Hanson 's Data Flow Diagram

Hanson is the reference pilot site for KTA1 (improved extraction, rock mass characterisation and control) Within Hanson's data lake, partners working on KTA1 will exchange data and results during several periods or test campaigns. After each blasting operation, Sandvik will collect and share MWD information (reports…TBD). Furthermore images, videos and logs of vibration will also be stored and shared MUL and UPM-M. Hanson will also contribute by storing data related to the primary crushing process of this blasted material. To that end Maxam will retrieve this information to produce and then store analysis reports related to the blasting process as well as optimized blasting parameters for future blasting operations. MUL will also store assessment reports related to the vibrations due to the blasting operations.

Hanson is also the reference pilot site for KTA3.2 (monitoring sensors and analysing tools both for Mobile Machinery in Loading &Transport and for the recognition of workers). Abaut, using Hanson's general information, will store KPIs related to the mobile machineries: usages, cycles, transportation times, distances, loading performance and transported tons.

Sigma/UPM-AI will retrieve dataset from Hanson to run their Hawkeye tool (used for aggregate quality and grain size determinations) The business management tool will retrieve KPIs to propose Business management dashboards. APP will also retrieve data from the data lake for their BIM solution. Hanson will take advantage of its data lake by retrieving KPIs, reports and processed data which will bring added value for the management of the quarry.

3.1.1.2 Advanced rock mass characterisation (KTA1.1)

3.1.1.2.1 MAXAM

A new methodology to assess rock mass quality from drill-monitoring data to guide blasting in open pit operations. Two rock description indexes will be derived directly from Measurement While Drilling (MWD) data collected by Sandvik drill. Principal component analysis will be used to combine MWD information. For that, corrections of the MWD parameters to minimize external influences other than the rock mass will be applied.

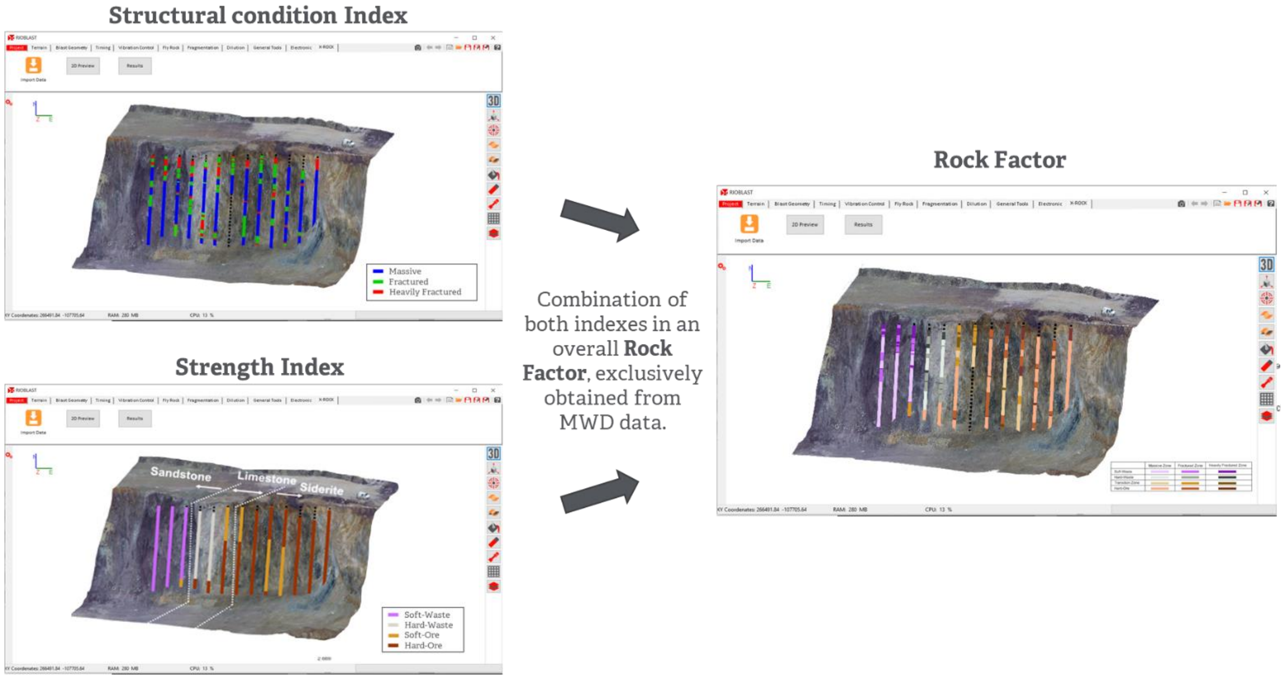
The first index is a Structural factor that classifies the rock mass condition in three classes (massive, fractured and heavily fractured). From it, a Structural Block model has been developed to simplify the recognition of rock classes. Video recording or Televiewer measurements (together with UPM) of the inner wall of the blastholes will be used to calibrate the results obtained.

The second index is a Strength factor, based on the combination of MWD parameters, that has been assessed from the analysis of the rock type description and strength properties from geology reports.

Finally, the Structural Block model is combined with the Strength factor to create the X-Rock model. This model, exclusively obtained from drill monitoring data, can provide an automatic assessment of rock structure, strength to be used as a Rock Factor.

The mathematical model of the X-Rock is implemented in MAXAM's Blast Design software, RIOBLAST, and is customized and calibrated for each drill/quarry/mine; it filters and normalizes automatically the MWD data to remove external influences different than the rock. Figure 2 shows an example of the model.

The output of the X-Rock model can be exported in *.csv or *.xlsx format to be imported into the DEQ Data Lake.



Combination of both indexes in an overall **Rock Factor**, exclusively obtained from MWD data.

Figure 2: Rock characterization form X-Rock model

3.1.1.2.2 SANDVIK

Drill plans, Quality reports and MWD data to/from drill rigs will be exchanged in IREDES format. The IREDES format is an XML container allowing easy access to the data and a flexible data payload depending on the data logged on the drill.

IREDES files can be transferred using Sandvik cloud platforms with manual upload and download. In addition, APIs for automatic retrieval of files can be made available. IREDES files can also be transferred manually using USB flash drives.

IREDES information can be parsed from further use. In addition, many drill & blast planning SW allow exports of data in converted formats, including CSV and XLS.

Equipment utilization data will be available as Excel files in Sandvik cloud environment – an API extension to download and automatically retrieve in CSV or JSON formats will be developed.

Custom data loggers, e.g., for CAN bus data should not be integrated directly to higher level systems, but first parsed to a standard format suitable for integration into databases.

Main inputs for drilling execution are 1) drill plans in IREDES format (manual import and conversion from other formats is possible) and 2) surface models in LandXML format (conversion from DWG and DXF is possible).

Data to and from the drill rig(s) will be transferred through Sandvik OEM cloud with an API to external data lakes.

3.1.1.2.3 UPM-M

Structural rock conditions (jointing, cavities, etc.) and lithology changes will be investigated from in-borehole images and/or photogrammetric models of the highwall faces. The measurements will be processed and analyzed with MATLAB, ShapeMetrix 3D and associated softwares from 3GSM, and CloudCompare. If televiewer is finally used to log the blastholes, WellCAD software from ALT will be also employed. The discontinuities characteristics, like orientation, spacing between discontinuities, fracture length, from these softwares will be an input to calculate the In-situ Block Size Distributions (IBSD) with Fracman suit or Matlab; for the latter non-parametric distributions will be used.

The drilling data recorded while drilling or Measurement-While drilling (MWD) will be analyzed using MATLAB algorithms and scripts; direct measurements of the rock mass will be used to calibrate the model. The purpose is to detect automatically clay patches and fractures from drilling data. For more details on the input data refer to Deliverable 3.1.

3.1.1.3 Productive and efficient drilling technology (KTA1.2)

Drilling productivity, performance and settings follow-up using the data logging and transfer means described in section 3.1.1.2.2. Focus is on the MWD data and drilling production KPIs.

3.1.1.4 Better explosives characterisation (KTA1.3)

3.1.1.4.1 MAXAM

The Selective Energy combines a series of innovative and technological components designed to deliver in each borehole the right quantity and distribution of the explosive's energy according to the properties of the rock. In order to carry out it, it is firstly necessary to consider the geomechanical properties of the rock mass within the blasting, as from the X-Rock model, in order to adjust the explosive density to match the energy released by the detonation. Thus, MAXAM's innovative explosives application technology (Smart RIOFLEX), combined with the geomechanical and geotechnical characterization of the rock, allows the optimization of the blast outcomes, such as fragmentation, rock micro-fissuration (reduction of the rock grindability indexes) and control of slope damage in the buffer and contour rows.

Smart RIOFLEX allows a wide range of densities (0.6 g/cm3 - 1.35 g/cm3) to be achieved, making it possible to adapt the energy available in the detonation process more selectively. The selectivity process is normally developed by adapting and varying the density of the explosive (and thus its energy) along the borehole itself according to specific loading profiles or as per the type of rock defined by the X-Rock (geotechnical and hardness domains, as exemplified in Figure 2**¡Error! No se encuentra el origen de la referencia.**). Figure 3 shows an example of selective energy application to rock type. For that, RIOBLAST includes a new modulus that allows to assign an explosive density to match the rock condition along the blasthole (as from the X-Rock), considering the drill pattern, with the goal to obtain a specific fragmentation size that will optimize both digging and comminution rates.

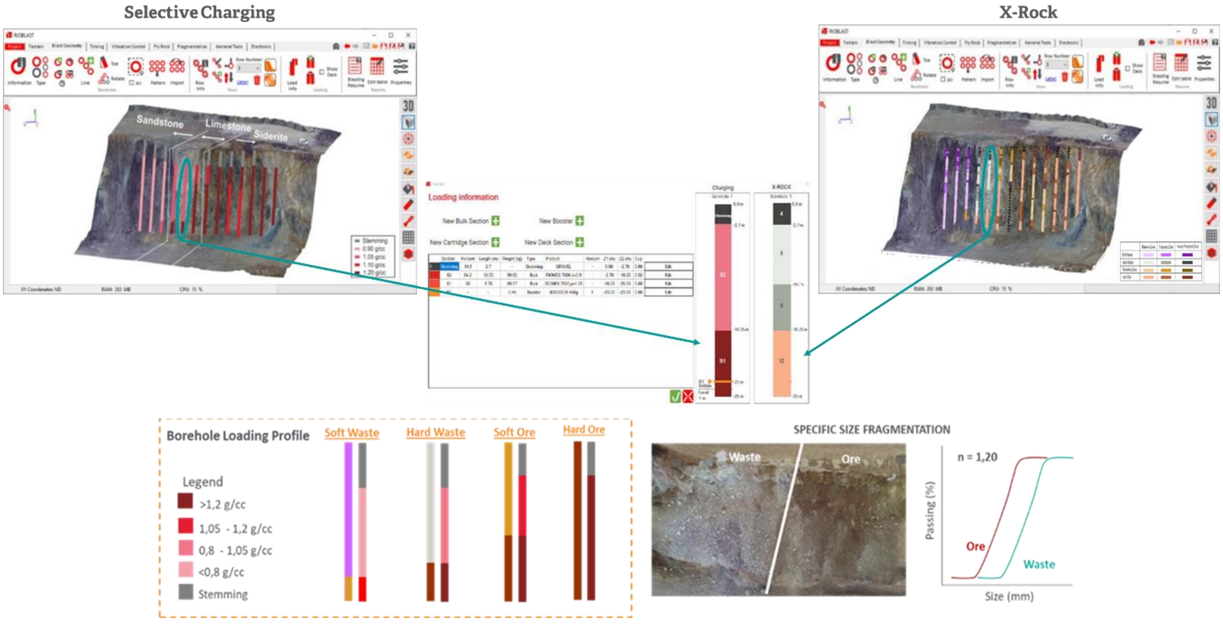


Figure 3: Example of the adjustment of density of the explosive according to the characteristics of the rock obtained with the X-Rock

Information of the explosive amount (kg), densities (g/cm3) and energy (Kj/kg) along of the hole, together with the borehole geometry and conditions will be later exported in either *.cvs, *.xlsx or *.XML format files to be imported into DEQ – Data Lake.

3.1.1.4.2 UPM-M

Measurements of velocity of detonation (VOD) will be downloaded with the Datatrap software manufactured by MREL. MATLAB will be used for the determination of the velocity of VOD and the calculation of detonation pressure from pressure-time histories. For more details on the input data refer to Deliverable 3.1.

3.1.1.5 Blasting Results Control through on-site novel technologies (KTA1.4)

3.1.1.5.1 MAXAM

Systematic quality control processes during drilling and blasting operations must be carried out by collecting and digitalizing field data of the different variable/stages that have an impact in blasting results. For that MAXAM's Digital Tools will be implemented developed on site and customized for quarries:

1. **RIOBLAST**

A 3D blast design and simulation software specially developed to help blasters and engineers to add value in their daily works thanks to its simplified and intuitive interface, offering the possibility of designing, analyzing, and simulating different blasting configurations according to real rock characteristics.

User-friendly interface
Built-in 3D design environment
Multi-density loading possibilities
Electronic timing design
Advanced simulation modules
Integration with BLAST CENTER

***Figure 4**. RIOBLAST Software for blast design and simulation.*

2. **MAXAM BLAST CENTER**

A cloud-based hub that enables the full digitalization, management, traceability, and analysis of blasting services. It integrates the most novel MAXAM's digital tools such as RIOBLAST, X-Logger and X-Truck, to ensure a reliable integration of different technological components of MAXAM's X-Energy solutions.

Integrates Fleet Management system

Blast-related data handling, storing, sharing and analyzing

Create Personalized Reports

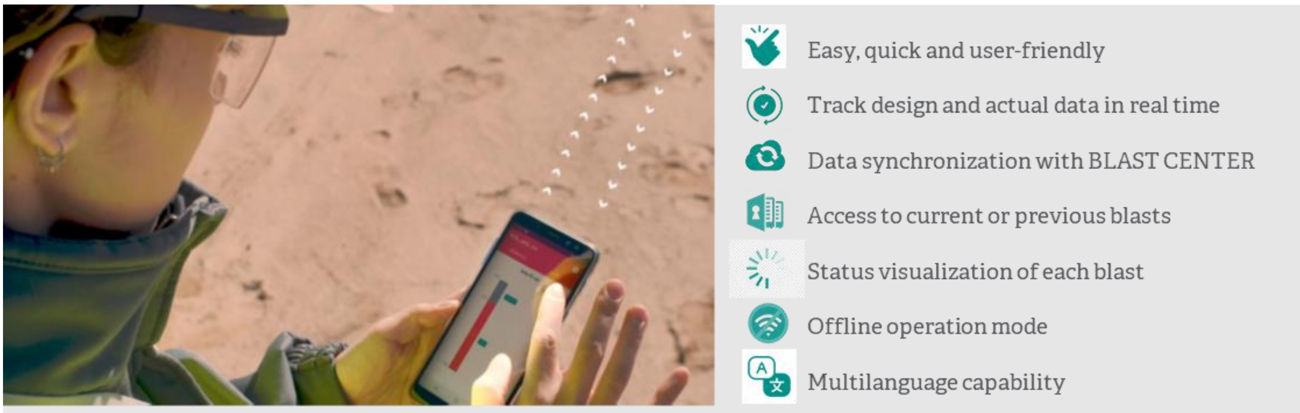
Data Integration with customer's systems

Continuous evolution (import third-party data, adapt to customer's needs)

***Figure 5**. MAXAM Blast Center is a digital platform.*

3. X-LOGGER

A secure, efficient, and user-friendly portable device application to collect, verify and update drilling and blasting data on the bench. As a fundamental part of MAXAM Digital Tools, X-LOGGER brings the opportunity to easily digitalize vital information for a sustainable blasting optimization program. When using X-Logger, all the information is transferred in real time to Blast Center, this even includes a new borehole that has been created in the field and was not in the original blast plan. The system allows for multiple devices operating simultaneously, including off-line mode communications.



Easy, quick and user-friendly

Track design and actual data in real time

Data synchronization with BLAST CENTER

Access to current or previous blasts

Status visualization of each blast

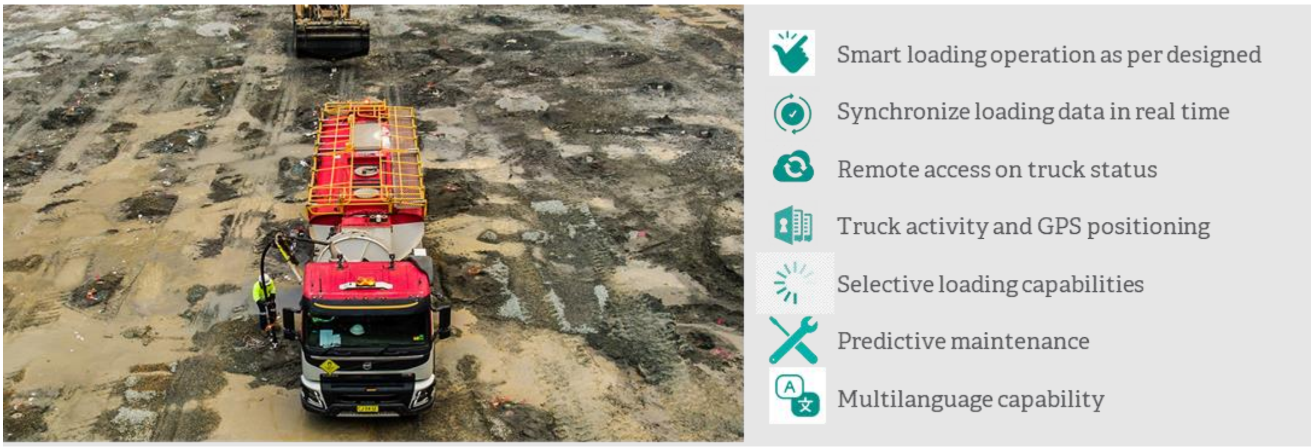
Offline operation mode

Multilanguage capability

Figure 6. X-LOGGER, app to actual data collection.

4. X-TRUCK

Is the new generation of MAXAM's fully digitalized Mobile Sensitizing Unit (MSU). As part of our digital capabilities for optimizing loading operations with real-time data exchange and transparence, X-TRUCK integration with MAXAM digital environment via Blast Center allows designed loading plans to be accurately executed in the bench, and actual as-loaded data be remotely reported in real time. The possibility to collect information from the truck is not mandatory to develop the QA/QC program; however, it can be discuss the adaptation of the truck used during the trials for this capability.



Smart loading operation as per designed

Synchronize loading data in real time

Remote access on truck status

Truck activity and GPS positioning

Selective loading capabilities

Predictive maintenance

Multilanguage capability

***Figure 7**. X-Truck. Digitalization and integration of MSUs.*

Systematic quality control processes in blasting operations ensures the correct compliance with the international standards of the explosives application, optimizing the use of mining resources to achieve the target results. Once each blast has a detailed loading plan, with precise specifications regarding to the quantity and density of the explosive, powder factor, charge and stemming length, among others, MAXAM will keep a control of the main quality and performance indicators to ensure the compliance of the blasting specifications. This information will be later exported in either *.cvs, *.xlsx or *.XML format files to be imported into DEQ – Data Lake and to correlate with blasting results to optimize fragmentation and digging and comminution performance.

3.1.1.5.2 UPM-M

A comprehensive list of the measurements that will be made before and after the blast is included in Deliverable D3.1. Images collected from drone flights at different stages (e.g., before the blast, immediately after the blast, and after mucking) it will be processed with BlastMetrix UAV module (3GSM) to develop the 3D models before and after the blasts. The coordinates of the actual borehole path and the resulting point clouds from the 3D models will be analyzed with BlastMetrix software (3GSM), quarry X (Geo-Koncept) and the open-source software, Cloudcompare. From them, the blast characteristics, like volume of rock broken by the blast, drilling pattern, hole deviation, bench height and subdrill length, among others will be calculated.

The 3D models of the muck piles will be analyzed with a fragmentation analysis software, e.g., Split desktop (Split Engineering) or Fragmenter (3GSM), to obtain the size distributions curves. The amount of material at the bypass of the crusher provided by the belt scale at that location will be used to calibrate the size distributions.

Measurements from geophones in the near field will be analyzed and processed with MATLAB to calibrate the semi-analytical full-field solution model. For this, detonation pressure measurements will be an input to simulate the shock pressure acting on the borehole walls.

3.1.1.6 Drill-to-mill (D2M) concept implementation (KTA1.5)

3.1.1.6.1 MAXAM

Overall (drill to mill – D2M) assessment will be defined. Based on the rock characterization and drilling QAQC, the best blast configuration (explosive type, characteristics, and timing) will be customized to optimize rock fragmentation (homogeneous and desirable particle size), muck pile digging efficiency indicators and comminution performance data, all of them integrated in an overall cost/efficiency analysis to define the blasting that optimizes the operation. For that, close collaboration with the development of mobile machinery sensors for digging and hauling to extract information

and relevant parameters will be key. This will be used to define new blasts during a second blasting campaigns in Hanson where the model and methodology developed will be validated.
The output of this section is still to be define but information will be exported in any format to be imported into DEQ-Data Lake.

3.1.1.6.2 SANDVIK

Drill rigs will be used as data source for processing and throughput modelling. MWD data input with refined and specific drill rig output uses the means defined in Section 3.1.1.2.2.

3.1.1.6.3 ABAUT - Mass flow and loading – hauling – dumping activities

Abaut will install Abaut Edge [sensor device consisting in sensor, antenna, and power supply] and Abaut mView [camera system for mobile machinery consisting of a camera, a holder for the cabin and a power supply] in the mobile machinery of Hanson in Valdilecha. More details about the use and components will be described in 3.1.1.7. The information that will be provided for the D2M Concept will be:

- Cycle times: Split according to the machines' activities. Loading, unloading, driving loaded, driving unloaded, idling loaded, idling unloaded. All the times of the cycle activity will be in seconds.
- Loading / unloading positions and material flow: Loading and unloading positions according to its location in the quarry in the cartographic system WGS84. In addition to the coordinate system, a geofence system will be created for also having the corresponding working name of the area [e.g., Bench 1, Crusher, etc.]. The information shared will be used to create a full mass flow reporting and monitoring system that can track each single truck, loading machine and tonnage from its origin until the dumping point.

3.1.1.6.4 UPM-M

The fragmentation energy-fan principles will be employed to predict fragmentation from blasting using rock mass properties (i.e., number of natural fines, spacing between fractures, orientation of discontinuities with respect the highwall face, IBSD distributions), blast characteristics (i.e., drilling data, explosive energy per hole, and timing), and size distributions from blasting. This will provide a tool to control fragmentation and define the optimum drilling parameters to optimize downstream key performance indicators, like mucking efficiency, energy consumption at the crusher, and amount of product fractions with higher prices. For this, minimization routines programed in MATLAB will be used.

3.1.1.7 Monitoring sensors and analysing tools both for Mobile Machinery in Loading &Transport and for the recognition of workers (KTA3.2)

abaut, as explained in 3.1.1.6.3, will install the patented abaut Edge sensor system and mView system in the mobile fleet of Hanson. It consists of sensor Edge device, antenna, and power supply. The system can be installed in any mobile heavy machine without the need of any retrofit kits and is independent of the age, brand, and model of the mobile equipment.

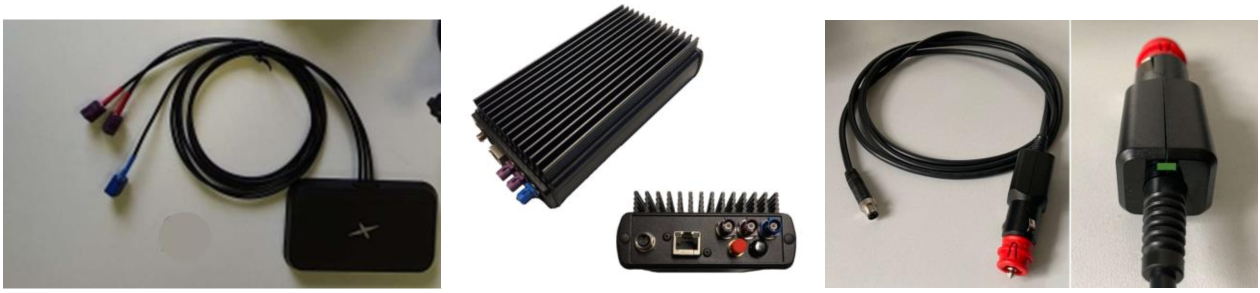

[1]

Figure 8: abaut Edge sensor system

The camera system, Abaut mView, will be installed in the cabin of the mobile machine providing a similar view as the one the operator has. Abaut mView is powered over ethernet [PoE] and is installed at the front wind-shield with the special holder for this purpose.

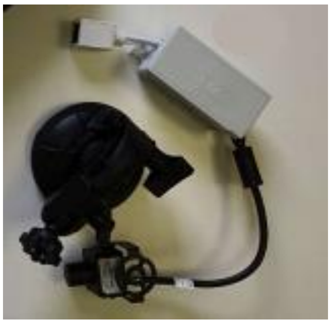
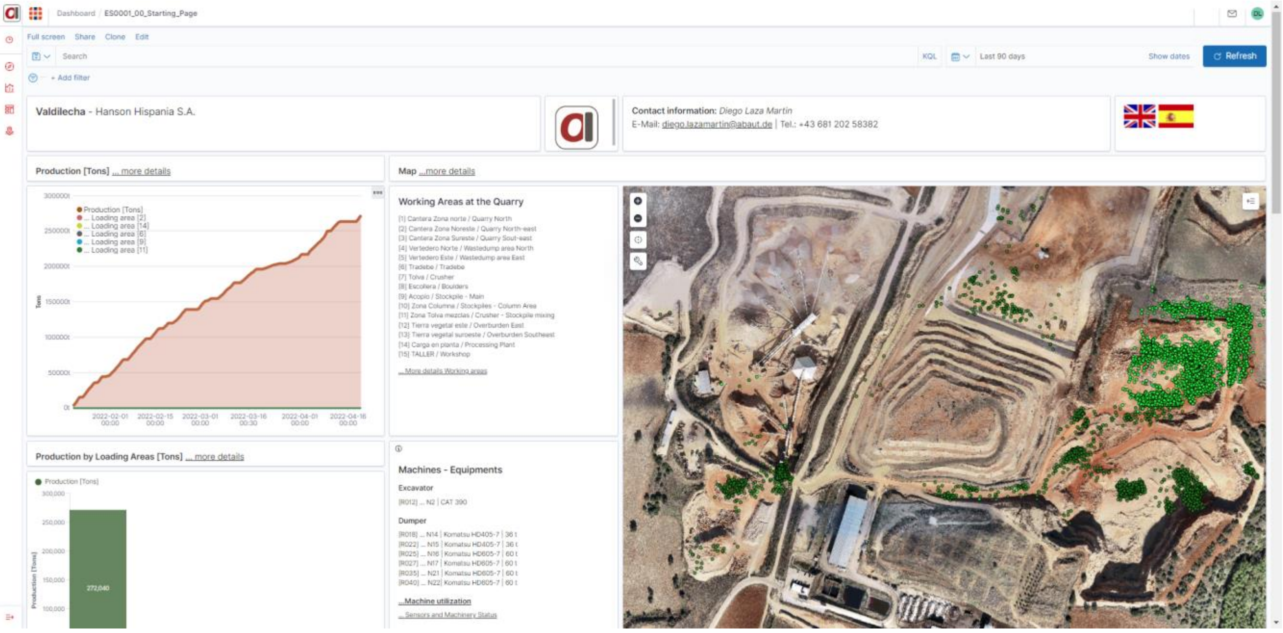


Figure 9: abaut mView

The data sent by the camera and sensor is automatically analyzed by the expert system of Abaut and will provide analytics and re-create the quarry activities in the digital twin of Abaut analytics, the cloud base Analytics that Abaut develops as Expert System Analytical tool.



Valdilecha - Hanson Hispania S.A.

Contact information: *Diego Laza Martin*
E-Mail: diego.lazamartin@abaut.de | Tel.: +43 681 202 58382

Production [Tons] ... more details

Map ...more details

Working Areas at the Quarry

[1] Cantera Zona norte / Quarry North
[2] Cantera Zona Noreste / Quarry North-east
[3] Cantera Zona Sureste / Quarry Sout-east
[4] Vertedero Norte / Wastedump area North
[5] Vertedero Este / Wastedump area East
[6] Tradebe / Tradebe
[7] Tolva / Crusher
[8] Escollera / Boulders
[9] Acopio / Stockpile - Main
[10] Zona Columna / Stockpiles - Column Area
[11] Zona Tolva mezclas / Crusher - Stockpile mixing
[12] Tierra vegetal este / Overburden East
[13] Tierra vegetal suroeste / Overburden Southeast
[14] Carga en planta / Processing Plant
[15] TALLER / Workshop

...More details Working areas

Production by Loading Areas [Tons] ... more details

Machines - Equipments

Excavator

[R012] ... N2 | CAT 390

Dumper

[R018] ... N14 | Komatsu HD405-7 | 36 t
[R022] ... N15 | Komatsu HD405-7 | 36 t
[R025] ... N16 | Komatsu HD605-7 | 60 t
[R027] ... N17 | Komatsu HD605-7 | 60 t
[R035] ... N21 | Komatsu HD605-7 | 60 t
[R040] ... N22 | Komatsu HD605-7 | 60 t

...Machine utilization

... Sensors and Machinery Status

Figure 10: abaut Analytics

Analytics, consists in 4 modules

- Production Mass flow: Origin and destination of the internal quarry logistics and production
- Analysis of Inefficiencies: Possibility to identify the different bottlenecks at the transport routes and idling times of all the fleet by duration, location, and mobile machine. The camera data will help to understand the working environment and take actions according to the affections detected

- Fleet Performance: Detailed analysis of cycle times, productivity, availability, and machinery utilization
- Reporting: Module dedicated to reporting and data sharing

The access to the first 3 modules will be done via User – Password credentials via web application. The last module, Reporting, can also integrate a VPN option, that allows to automatically send some predefined reports [e.g., Internal Logistics – Transports] directly to the data lake, IQS or to the reporting system of the quarry.

For more details regarding the complete Input-Output feature list, please go to the Appendix section included at D3.1 Definition of requirements and characteristics of the data inputs.

In the image below, internal data flow of Abaut' s expert system, is possible to observe Abaut' s data flow system, starting from the generation of the data via Abaut Edge & mView, the analysis of the data generated and the integration in the IQS/data lake system of DigiEcoQuarry:

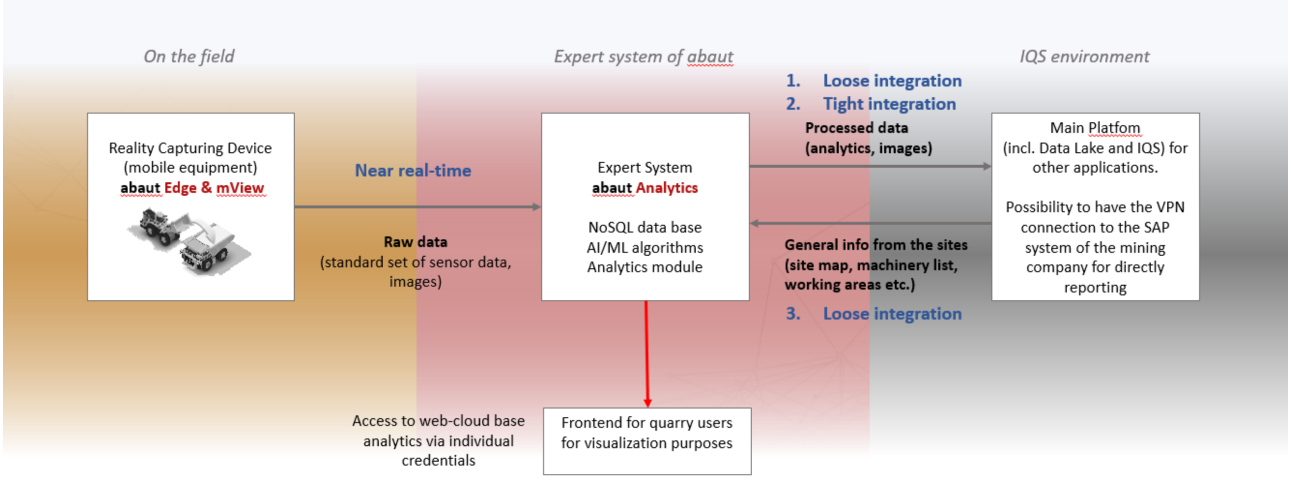

On the field
Expert system of abaut
IQS environment
Reality Capturing Device (mobile equipment) abaut Edge & mView
Near real-time
Raw data (standard set of sensor data, images)
Expert System abaut Analytics
NoSQL data base
AI/ML algorithms
Analytics module
1. Loose integration
2. Tight integration
Processed data (analytics, images)
Main Platfom (incl. Data Lake and IQS) for other applications.
Possibility to have the VPN connection to the SAP system of the mining company for directly reporting
General info from the sites (site map, machinery list, working areas etc.)
3. Loose integration
Access to web-cloud base analytics via individual credentials
Frontend for quarry users for visualization purposes

Figure 11: Abaut 's Expert System internal data flow

3.1.2 VICAT

3.1.2.1 Inventory of the existing ICT assets

The following table provides a high-level view of the expert systems, interfaces, contents summary, format, data sharing, and partners involved within this site. It enables the creation of pilot site's data flow. Please refer to D3.1 to have a detailed view of the data.

System Expert	Description of the function	Interface type provided	Content	Format	Shared data through	Shared data with
Vicat Expert Systems and reporting tools	Quarry management system: Store and Upload data	API/Manual upload and download	Historical data Production data Water consumption data General information data Documentation for Metaquarry	xls	Data Lake	BMT SIGMA: Metaquarry APP ABAUT: abaut Analytics METSO ARCO

System Expert	Description of the function	Interface type provided	Content	Format	Shared data through	Shared data with
Arco Expert system	Store weighting data	API	weighting data	json	Data Lake	VICAT BMT APP
Maestro/ QProd (Not yet agreed at this stage of the project, to be confirmed)	Store production data	REST API	Production data	Json	Data Lake	VICAT BMT APP
Metso /Metrics	Store production and environment KPIs	Rest API Manual upload	Noise data Production data Running hours data Fuel consumption (effective and non-effective)	csv	Data Lake	VICAT BMT APP
Abaut Analytics	Store data.	API	Recognition of activities results data.	csv	Data Lake	VICAT BMT APP SIGMA: Metaquarry
BMT	Generate, store and share reports and dashboard	manual upload	Dynamic and static view of data and KPIs shared	pdf, xls	Data Lake	VICAT

3.1.2.1.1 Data flow

The following diagram depicts the data flows between the partners or systems within this pilot site.

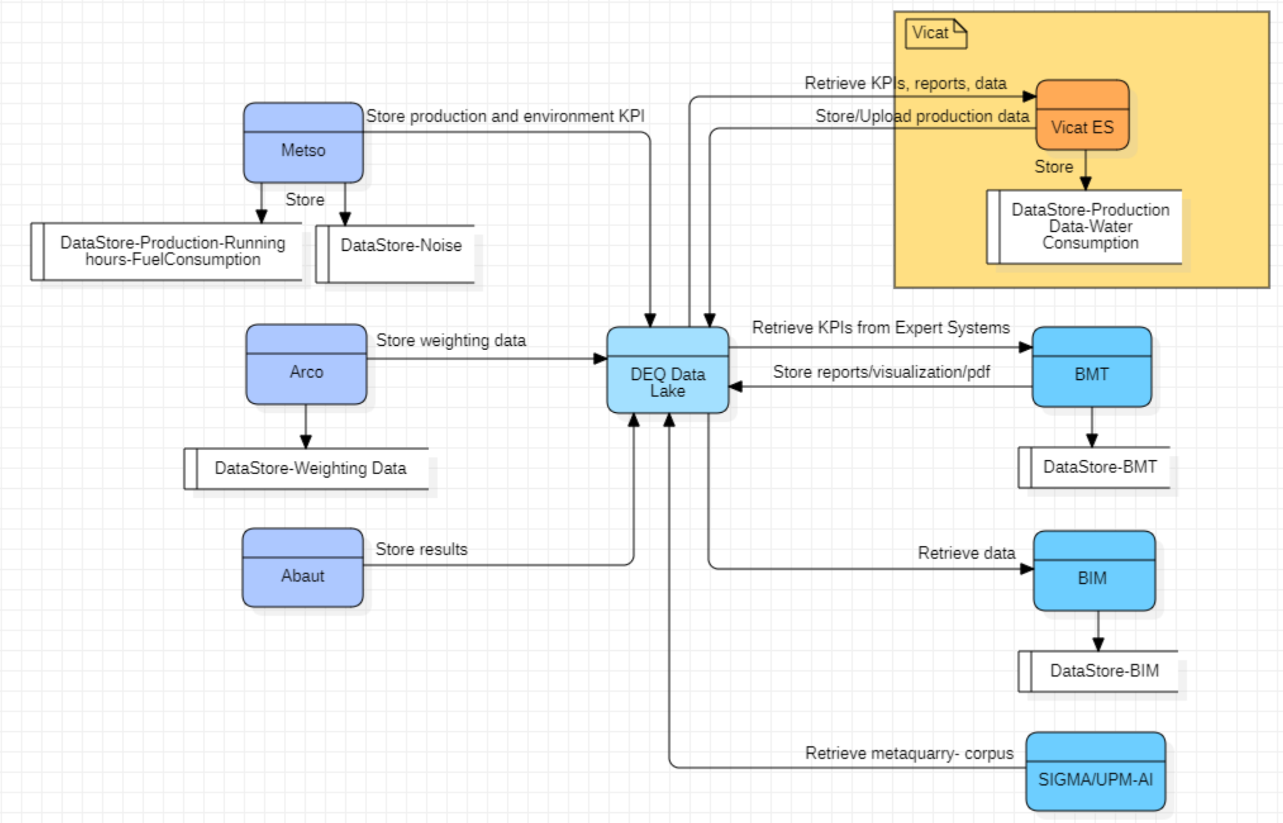

Vicat
Retrieve KPIs, reports, data
Store/Upload production data
Vicat ES
Store
DataStore-Production Data-Water Consumption
Metso
Store production and environment KPI
Store
DataStore-Production-Running hours-FuelConsumption
DataStore-Noise
Arco
Store weighting data
DEQ Data Lake
Retrieve KPIs from Expert Systems
Store reports/visualization/pdf
BMT
DataStore-Weighting Data
DataStore-BMT
Abaut
Store results
Retrieve data
BIM
DataStore-BIM
Retrieve metaquarry- corpus
SIGMA/UPM-AI

Figure 12: Vicat’s Data Flow Diagram

Within its data lake, Vicat will store historical data and production data related to the production of material and water consumption on a daily or monthly basis. Additional data, such as production KPIs data, from Vicat’s scada system and from ARCO’s weighting system, specific production data and environmental KPIs data coming from Metso’s expert system will also be stored. This data will be available, according to their rights, and usable by external partners. As such, Abaut will retrieve Vicat’s general information. Abaut will store in return, recognition of activities results data.

Sigma/UPM-AI will retrieve dataset from Vicat to run their Metaquarry tool (NLP information and document search engine) The business management tool will retrieve KPIs to propose Business management dashboards. APP will also retrieve data from the data lake for their BIM solution. Vicat will take advantage of its data lake by retrieving KPIs, reports and processed data which will bring added value for the management of the quarry.

3.1.2.1.2 Data model

After a first analysis of Treatment plant production provided by Vicat and Hanson. The following diagram shows the first version of the data model deducted from data provided by Vicat.

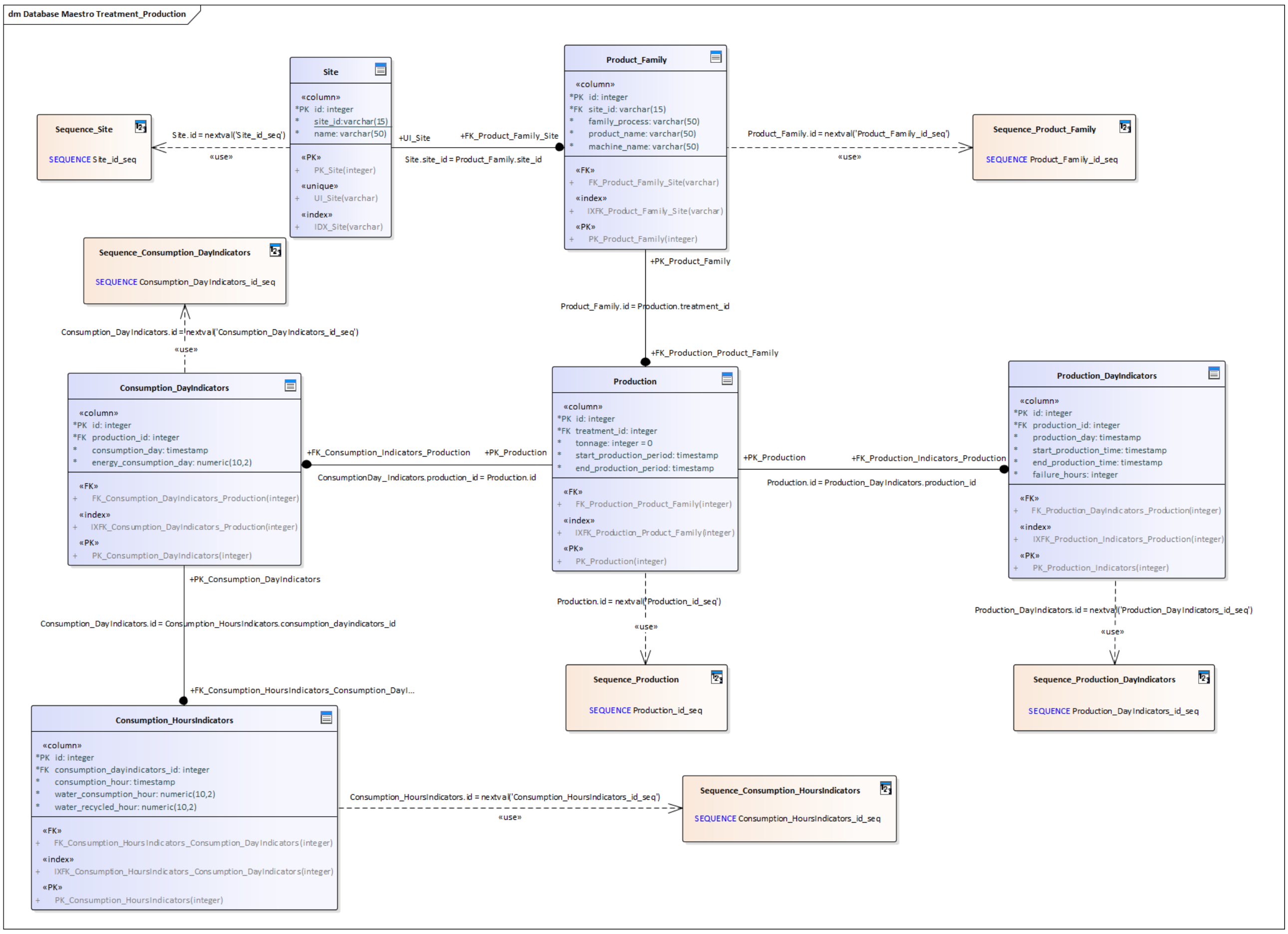
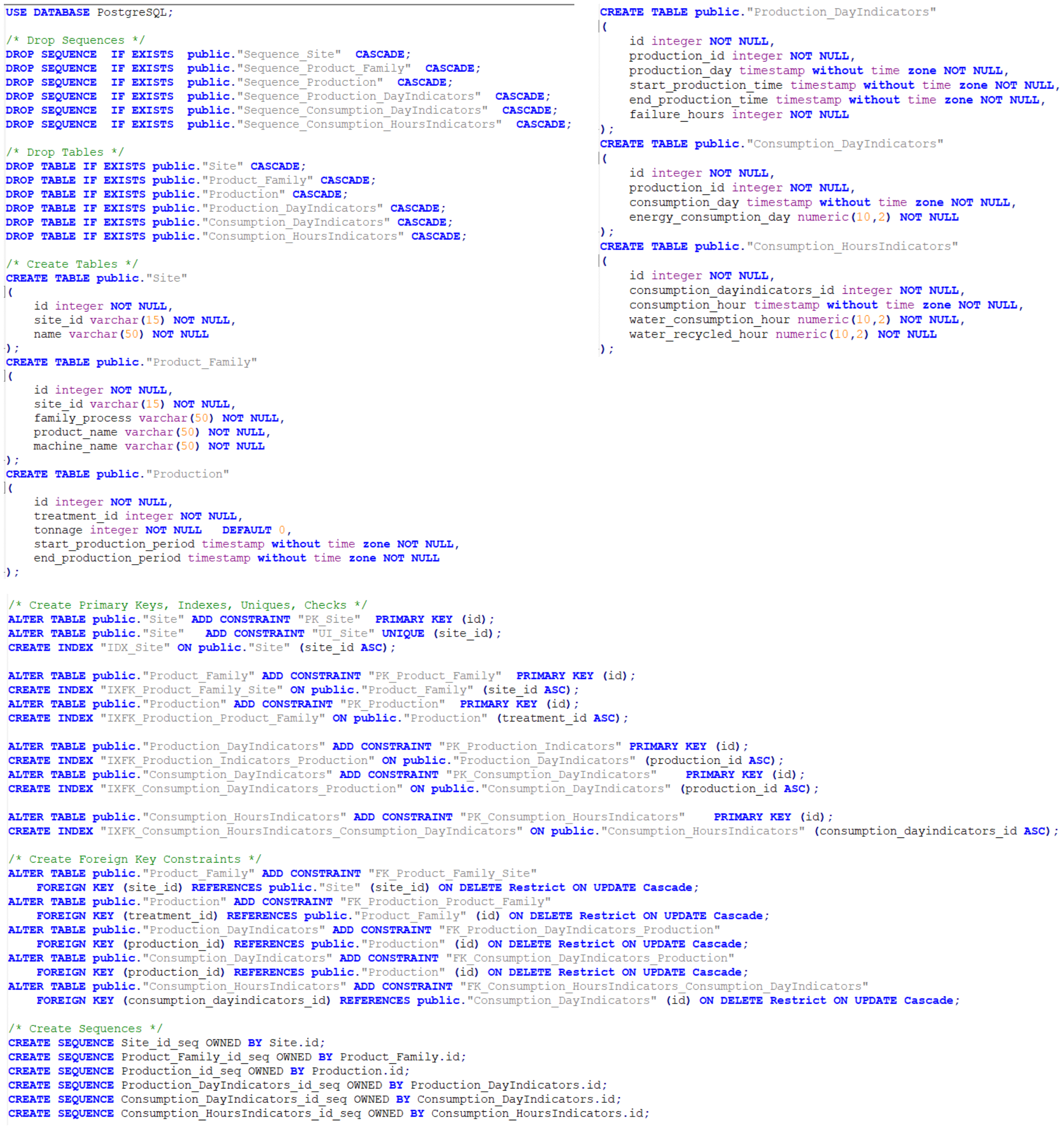

dm Database Maestro Treatment_Production
Sequence_Site
SEQUENCE Site_id_seq
Site.id = nextval('Site_id_seq')
«use»
Site
«column»
*PK id: integer
* site_id:varchar(15)
* name: varchar(50)
«PK»
+ PK_Site(integer)
«unique»
+ UI_Site(varchar)
«index»
+ IDX_Site(varchar)
+UI_Site
+FK_Product_Family_Site
Site.site_id = Product_Family.site_id
Product_Family
«column»
*PK id: integer
*FK site_id: varchar(15)
* family_process: varchar(50)
* product_name: varchar(50)
* machine_name: varchar(50)
«FK»
+ FK_Product_Family_Site(varchar)
«index»
+ IXFK_Product_Family_Site(varchar)
«PK»
+ PK_Product_Family(integer)
Product_Family.id = nextval('Product_Family_id_seq')
«use»
Sequence_Product_Family
SEQUENCE Product_Family_id_seq
+PK_Product_Family
Product_Family.id = Production.treatment_id
+FK_Production_Product_Family
Sequence_Consumption_DayIndicators
SEQUENCE Consumption_DayIndicators_id_seq
Consumption_DayIndicators.id = nextval('Consumption_DayIndicators_id_seq')
«use»
Consumption_DayIndicators
«column»
*PK id: integer
*FK production_id: integer
* consumption_day: timestamp
* energy_consumption_day: numeric(10,2)
«FK»
+ FK_Consumption_DayIndicators_Production(integer)
«index»
+ IXFK_Consumption_DayIndicators_Production(integer)
«PK»
+ PK_Consumption_DayIndicators(integer)
+FK_Consumption_Indicators_Production
+PK_Production
ConsumptionDay_Indicators.production_id = Production.id
Production
«column»
*PK id: integer
*FK treatment_id: integer
* tonnage: integer = 0
* start_production_period: timestamp
* end_production_period: timestamp
«FK»
+ FK_Production_Product_Family(integer)
«index»
+ IXFK_Production_Product_Family(integer)
«PK»
+ PK_Production(integer)
+PK_Production
+FK_Production_Indicators_Production
Production.id = Production_DayIndicators.production_id
Production_DayIndicators
«column»
*PK id: integer
*FK production_id: integer
* production_day: timestamp
* start_production_time: timestamp
* end_production_time: timestamp
* failure_hours: integer
«FK»
+ FK_Production_DayIndicators_Production(integer)
«index»
+ IXFK_Production_Indicators_Production(integer)
«PK»
+ PK_Production_Indicators(integer)
+PK_Consumption_DayIndicators
Consumption_DayIndicators.id = Consumption_HoursIndicators.consumption_dayindicators_id
+FK_Consumption_HoursIndicators_Consumption_DayI...
Production.id = nextval('Production_id_seq')
«use»
Sequence_Production
SEQUENCE Production_id_seq
Production_DayIndicators.id = nextval('Production_DayIndicators_id_seq')
«use»
Sequence_Production_DayIndicators
SEQUENCE Production_DayIndicators_id_seq
Consumption_HoursIndicators
«column»
*PK id: integer
*FK consumption_dayindicators_id: integer
* consumption_hour: timestamp
* water_consumption_hour: numeric(10,2)
* water_recycled_hour: numeric(10,2)
«FK»
+ FK_Consumption_HoursIndicators_Consumption_DayIndicators(integer)
«index»
+ IXFK_Consumption_HoursIndicators_Consumption_DayIndicators(integer)
«PK»
+ PK_Consumption_HoursIndicators(integer)
Consumption_HoursIndicators.id = nextval('Consumption_HoursIndicators_id_seq')
«use»
Sequence_Consumption_HoursIndicators
SEQUENCE Consumption_HoursIndicators_id_seq

Figure 13: "Treatment / Production" Data Model used in the Data Lake to be compliant with VICAT / MAESTRO Data Structures

The following figure shows the database creation scripts:



```sql
USE DATABASE PostgreSQL;

/* Drop Sequences */
DROP SEQUENCE  IF EXISTS  public."Sequence_Site"  CASCADE;
DROP SEQUENCE  IF EXISTS  public."Sequence_Product_Family"  CASCADE;
DROP SEQUENCE  IF EXISTS  public."Sequence_Production"  CASCADE;
DROP SEQUENCE  IF EXISTS  public."Sequence_Production_DayIndicators"  CASCADE;
DROP SEQUENCE  IF EXISTS  public."Sequence_Consumption_DayIndicators"  CASCADE;
DROP SEQUENCE  IF EXISTS  public."Sequence_Consumption_HoursIndicators"  CASCADE;

/* Drop Tables */
DROP TABLE IF EXISTS public."Site" CASCADE;
DROP TABLE IF EXISTS public."Product_Family" CASCADE;
DROP TABLE IF EXISTS public."Production" CASCADE;
DROP TABLE IF EXISTS public."Production_DayIndicators" CASCADE;
DROP TABLE IF EXISTS public."Consumption_DayIndicators" CASCADE;
DROP TABLE IF EXISTS public."Consumption_HoursIndicators" CASCADE;

/* Create Tables */
CREATE TABLE public."Site"
(
    id integer NOT NULL,
    site_id varchar(15) NOT NULL,
    name varchar(50) NOT NULL
);
CREATE TABLE public."Product_Family"
(
    id integer NOT NULL,
    site_id varchar(15) NOT NULL,
    family_process varchar(50) NOT NULL,
    product_name varchar(50) NOT NULL,
    machine_name varchar(50) NOT NULL
);
CREATE TABLE public."Production"
(
    id integer NOT NULL,
    treatment_id integer NOT NULL,
    tonnage integer NOT NULL   DEFAULT 0,
    start_production_period timestamp without time zone NOT NULL,
    end_production_period timestamp without time zone NOT NULL
);
CREATE TABLE public."Production_DayIndicators"
(
    id integer NOT NULL,
    production_id integer NOT NULL,
    production_day timestamp without time zone NOT NULL,
    start_production_time timestamp without time zone NOT NULL,
    end_production_time timestamp without time zone NOT NULL,
    failure_hours integer NOT NULL
);
CREATE TABLE public."Consumption_DayIndicators"
(
    id integer NOT NULL,
    production_id integer NOT NULL,
    consumption_day timestamp without time zone NOT NULL,
    energy_consumption_day numeric(10,2) NOT NULL
);
CREATE TABLE public."Consumption_HoursIndicators"
(
    id integer NOT NULL,
    consumption_dayindicators_id integer NOT NULL,
    consumption_hour timestamp without time zone NOT NULL,
    water_consumption_hour numeric(10,2) NOT NULL,
    water_recycled_hour numeric(10,2) NOT NULL
);

/* Create Primary Keys, Indexes, Uniques, Checks */
ALTER TABLE public."Site" ADD CONSTRAINT "PK_Site"  PRIMARY KEY (id);
ALTER TABLE public."Site"   ADD CONSTRAINT "UI_Site" UNIQUE (site_id);
CREATE INDEX "IDX_Site" ON public."Site" (site_id ASC);

ALTER TABLE public."Product_Family" ADD CONSTRAINT "PK_Product_Family"  PRIMARY KEY (id);
CREATE INDEX "IXFK_Product_Family_Site" ON public."Product_Family" (site_id ASC);
ALTER TABLE public."Production" ADD CONSTRAINT "PK_Production"  PRIMARY KEY (id);
CREATE INDEX "IXFK_Production_Product_Family" ON public."Production" (treatment_id ASC);

ALTER TABLE public."Production_DayIndicators" ADD CONSTRAINT "PK_Production_Indicators" PRIMARY KEY (id);
CREATE INDEX "IXFK_Production_Indicators_Production" ON public."Production_DayIndicators" (production_id ASC);
ALTER TABLE public."Consumption_DayIndicators" ADD CONSTRAINT "PK_Consumption_DayIndicators"    PRIMARY KEY (id);
CREATE INDEX "IXFK_Consumption_DayIndicators_Production" ON public."Consumption_DayIndicators" (production_id ASC);

ALTER TABLE public."Consumption_HoursIndicators" ADD CONSTRAINT "PK_Consumption_HoursIndicators"    PRIMARY KEY (id);
CREATE INDEX "IXFK_Consumption_HoursIndicators_Consumption_DayIndicators" ON public."Consumption_HoursIndicators" (consumption_dayindicators_id ASC);

/* Create Foreign Key Constraints */
ALTER TABLE public."Product_Family" ADD CONSTRAINT "FK_Product_Family_Site"
    FOREIGN KEY (site_id) REFERENCES public."Site" (site_id) ON DELETE Restrict ON UPDATE Cascade;
ALTER TABLE public."Production" ADD CONSTRAINT "FK_Production_Product_Family"
    FOREIGN KEY (treatment_id) REFERENCES public."Product_Family" (id) ON DELETE Restrict ON UPDATE Cascade;
ALTER TABLE public."Production_DayIndicators" ADD CONSTRAINT "FK_Production_DayIndicators_Production"
    FOREIGN KEY (production_id) REFERENCES public."Production" (id) ON DELETE Restrict ON UPDATE Cascade;
ALTER TABLE public."Consumption_DayIndicators" ADD CONSTRAINT "FK_Consumption_DayIndicators_Production"
    FOREIGN KEY (production_id) REFERENCES public."Production" (id) ON DELETE Restrict ON UPDATE Cascade;
ALTER TABLE public."Consumption_HoursIndicators" ADD CONSTRAINT "FK_Consumption_HoursIndicators_Consumption_DayIndicators"
    FOREIGN KEY (consumption_dayindicators_id) REFERENCES public."Consumption_DayIndicators" (id) ON DELETE Restrict ON UPDATE Cascade;

/* Create Sequences */
CREATE SEQUENCE Site_id_seq OWNED BY Site.id;
CREATE SEQUENCE Product_Family_id_seq OWNED BY Product_Family.id;
CREATE SEQUENCE Production_id_seq OWNED BY Production.id;
CREATE SEQUENCE Production_DayIndicators_id_seq OWNED BY Production_DayIndicators.id;
CREATE SEQUENCE Consumption_DayIndicators_id_seq OWNED BY Consumption_DayIndicators.id;
CREATE SEQUENCE Consumption_HoursIndicators_id_seq OWNED BY Consumption_HoursIndicators.id;
```

3.1.2.2 Innovative mobile crusher (KTA2.1)

Metso Outotec will deliver a LT1213SE track mounted mobile horizontal impact crusher with innovative features to VICAT pilot site. In addition to “traditional” data and information of the machine status e.g., engine power and crusher speed, new noise sensors will be installed, and thus new noise data will be available. Available noise data will be: measured a-weighted sound pressure (e.g., 15-minute averages), measured a-weighted sound pressure for fast-idle operating (machine running, no load) and measured background noise (machine turned off). For the DEQ, a single “point of contact” method of sharing data will be implemented in the machine. Through this new solution (on-board computer), all shared data will be transmitted to the DEQ data lake system. This data can then be utilized in different ways for both offline and online analysis. The data shared to the IQS will be treated onboard the machine to match the required format. The shared data is separate from data utilized by the automation system of the machine. The data flow from the on-board computer is described in the following figure.

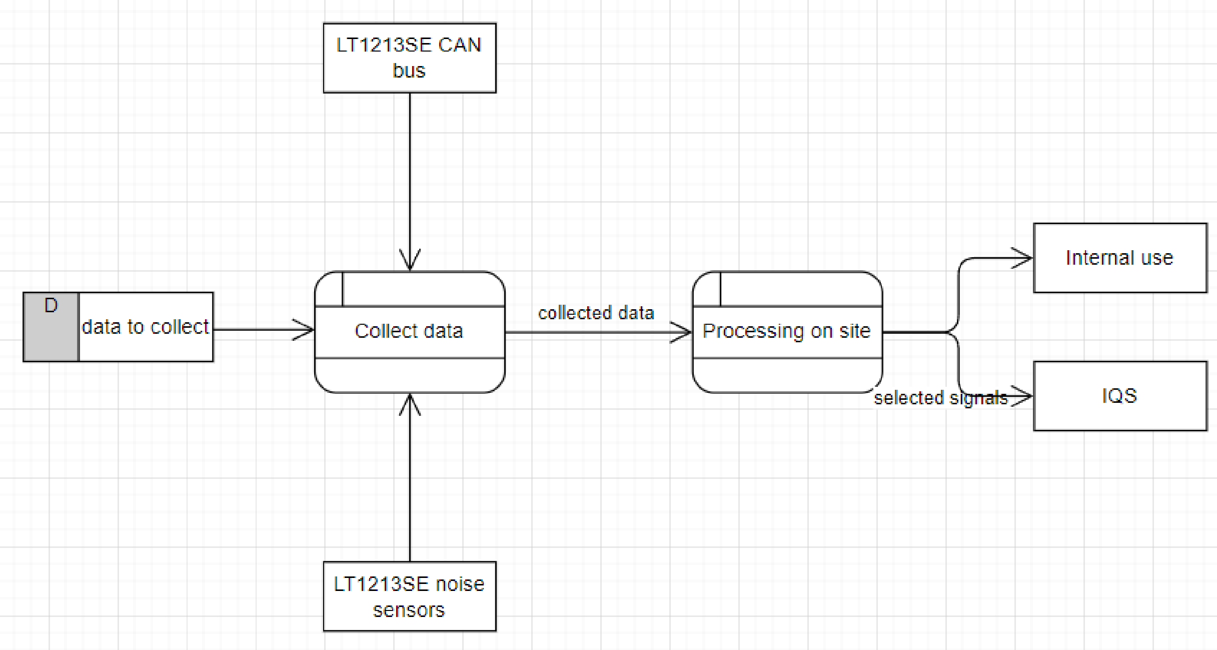


Figure 14: Metso: Outotec LT1213SE data flow to IQS

3.1.2.3 Devices for automation of treatment plants and storage facilities (KTA3.1)

Arco Weighing system can weigh material in flow through a bridge installed on a conveyor belt.

AP-DEQ-07: Integrated weighing

Control of production at the different points of transport by conveyor belts. Performs static weighing measurement and belt speed measurement.

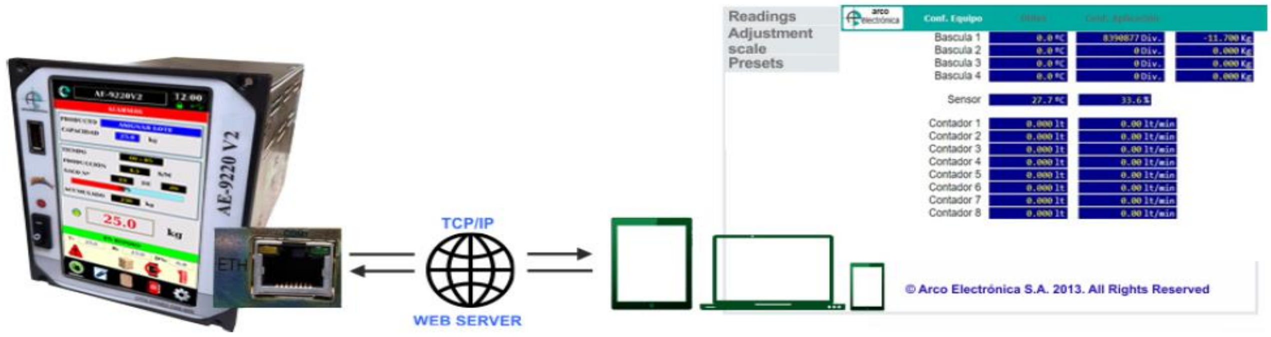
This equipment has:

Web server for configuration from any device.

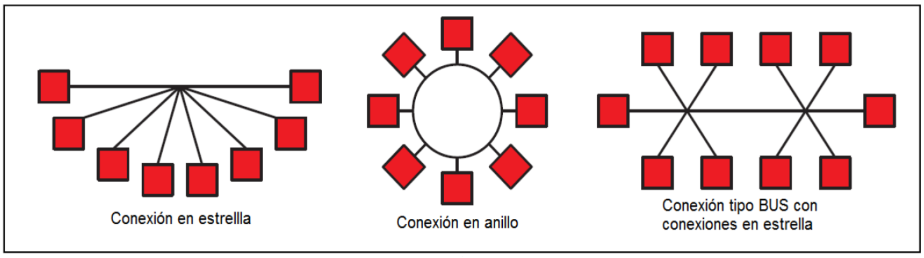
Real-time display.

Calibration and adjustment of equipment.

We get the information regarding the quantity of material we are processing in real time and the material produced in a determinate time, and we store the data to be consulting from other devices in any place.

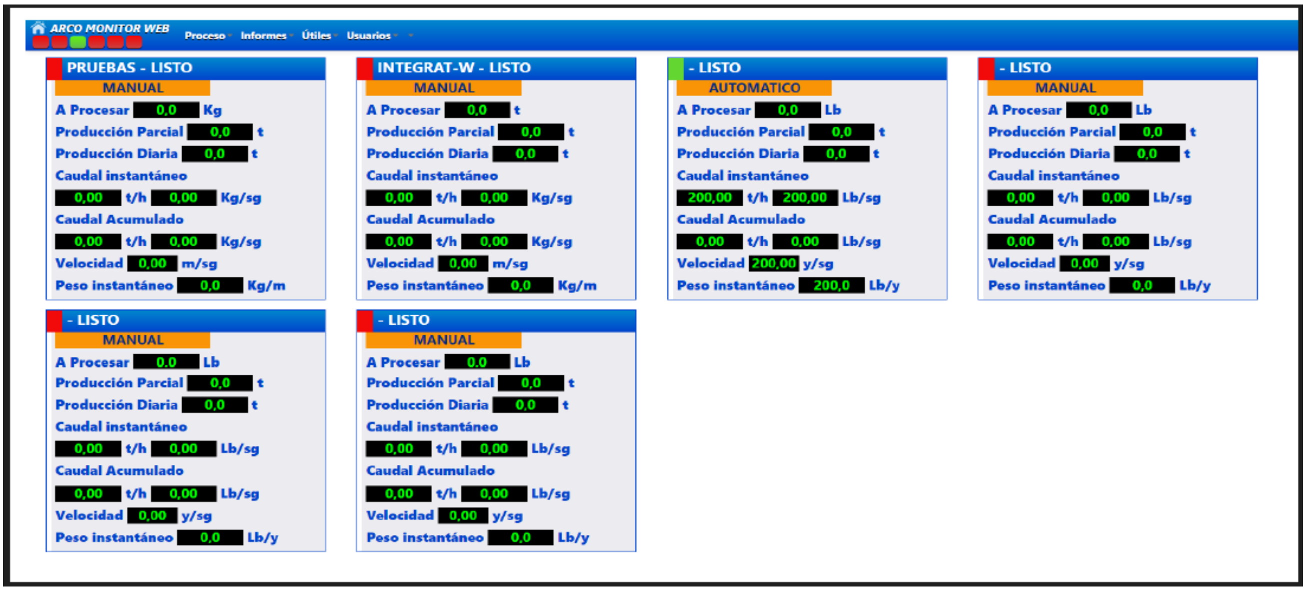


It is possible to connect multiple devices to work together in the same process.



It is advisable using the system connected to the computer to increase the performance, display the data in an extensive format, and enjoin the Arco Monitor serves.

ARCO MONITOR WEB



PRUEBAS - LISTO
MANUAL
A Procesar 0,0 Kg
Producción Parcial 0,0 t
Producción Diaria 0,0 t
Caudal instantáneo
0,00 t/h 0,00 Kg/sg
Caudal Acumulado
0,00 t/h 0,00 Kg/sg
Velocidad 0,00 m/sg
Peso instantáneo 0,0 Kg/m

INTEGRAT-W - LISTO
MANUAL
A Procesar 0,0 t
Producción Parcial 0,0 t
Producción Diaria 0,0 t
Caudal instantáneo
0,00 t/h 0,00 Kg/sg
Caudal Acumulado
0,00 t/h 0,00 Kg/sg
Velocidad 0,00 m/sg
Peso instantáneo 0,0 Kg/m

- LISTO
AUTOMATICO
A Procesar 0,0 Lb
Producción Parcial 0,0 t
Producción Diaria 0,0 t
Caudal instantáneo
200,00 t/h 200,00 Lb/sg
Caudal Acumulado
0,00 t/h 0,00 Lb/sg
Velocidad 200,00 y/sg
Peso instantáneo 200,0 Lb/y

- LISTO
MANUAL
A Procesar 0,0 Lb
Producción Parcial 0,0 t
Producción Diaria 0,0 t
Caudal instantáneo
0,00 t/h 0,00 Lb/sg
Caudal Acumulado
0,00 t/h 0,00 Lb/sg
Velocidad 0,00 y/sg
Peso instantáneo 0,0 Lb/y

- LISTO
MANUAL
A Procesar 0.0 Lb
Producción Parcial 0,0 t
Producción Diaria 0,0 t
Caudal instantáneo
0,00 t/h 0,00 Lb/sg
Caudal Acumulado
0,00 t/h 0,00 Lb/sg
Velocidad 0,00 y/sg
Peso instantáneo 0,0 Lb/y

- LISTO
MANUAL
A Procesar 0.0 Lb
Producción Parcial 0,0 t
Producción Diaria 0,0 t
Caudal instantáneo
0,00 t/h 0,00 Lb/sg
Caudal Acumulado
0,00 t/h 0,00 Lb/sg
Velocidad 0,00 y/sg
Peso instantáneo 0,0 Lb/y

The Arco Mineral Platinum application provides access to production data and status of weighing equipment integrated through WebServices. In order to access the WebServices, the Industrial PC where the application is installed must have an Internet connection.

Arco system will provide access to instant data and to production data using two APIs.

The following rules give an example of how to connect and access accumulated production data per day:

The date format is YYYY-MM-DD

A GET request to https://demo.arcoelectronica.es:8090/resources/produccion/2022-05-21 must be performed (the date is added as a parameter) and using a Token: mSbRLMWNmu7/WSU71xCMomUbIAjWI0XOYwvGrNByg44

A json response like the following is returned in the following format:

```
        [{
    "DEVICE": "PI-1",
    "DATE": "2022-05-21",
    "TURN1_AUT": 1344.22,
    "TURN2_AUT": 5405.41,
    "TURN3_AUT": 3419.2,
    "TURN1_MAN": 314.22,
    "TURN2_MAN": 425.41,
    "TURN3_MAN": 439.2,
    "TOTAL": 11347.66
        },
   {"DEVICE": "PI-2",
    "DATE": "2022-05-21",
    "TURN1_AUT": 344.22,
    "TURN2_AUT": 405.41,
    "TURN3_AUT": 419.2,
    "TURN1_MAN": 14.22,
    "TURN2_MAN": 25.41,
    "TURN3_MAN": 39.2,
    "TOTAL": 1247.66
        }]
```

Description of the fields:

DEVICE	Integrated weighing identifier
DATE	Production date
TURN1_AUT	Tons accumulated in turn 1 with the team in automatic
TURN2_AUT	Tons accumulated in turn 2 with the team in automatic
TURN3_AUT	Tons accumulated in turn 3 with the team in automatic
TURN1_MAN	Accumulated tons in shift 1 with the equipment in manual
TURN2_MAN	Accumulated tons in shift 2 with the equipment in manual
TURN3_MAN	Accumulated tons in turn 3 with the equipment in manual
TOTAL	Total tons of the day

3.1.2.4 Monitoring sensors and analysing tools both for Mobile Machinery in Loading &Transport and for the recognition of workers (KTA3.2)

Abaut will install abaut mView module in Vicat for analyzing the type of material that has been sent to the plant on each truck.

The idea is to identify in near-real-time the type of material transport and if this material contains any kind of pollutants [e.g., plastics, woods or steel bars between the material that is going to be processed] can affect the different processes at the processing plant.

At VICAT, the camera module mView will transmit the data generated to the expert system of abaut. The information will be analyzed and then displayed in Analytics for its analysis and decision-making step. This step is still under development and right now is only possible to offer a preliminary data flow system that will be tested at VICAT

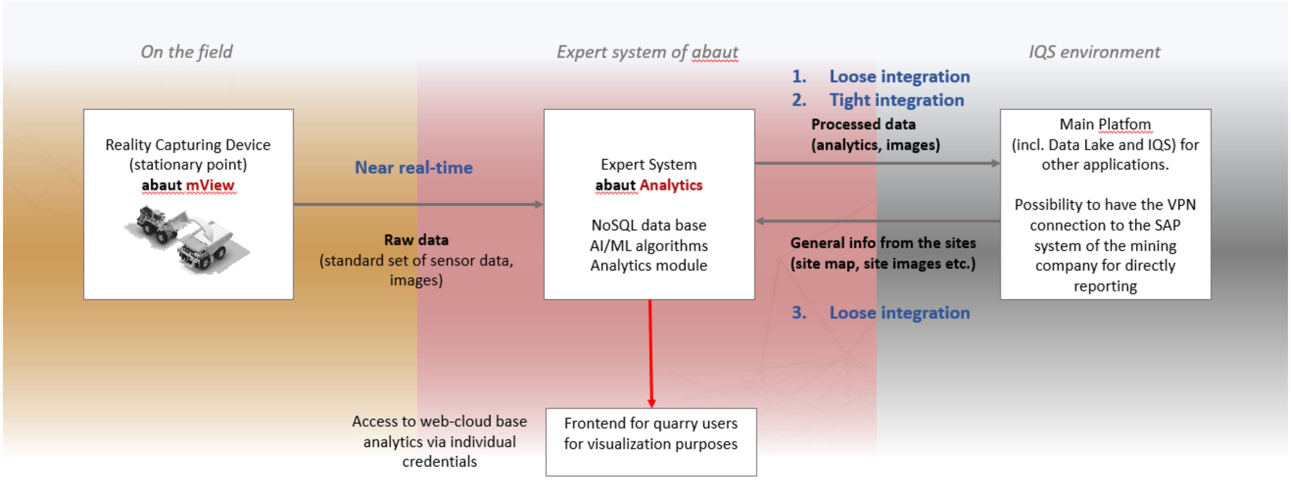

On the field
Expert system of abaut
IQS environment
Reality Capturing Device
(stationary point)
abaut mView
Near real-time
Raw data
(standard set of sensor data, images)
Expert System
abaut Analytics
NoSQL data base
AI/ML algorithms
Analytics module
1. Loose integration
2. Tight integration
Processed data
(analytics, images)
General info from the sites
(site map, site images etc.)
3. Loose integration
Main Platfom
(incl. Data Lake and IQS) for other applications.
Possibility to have the VPN connection to the SAP system of the mining company for directly reporting
Access to web-cloud base analytics via individual credentials
Frontend for quarry users for visualization purposes

Figure 15: Abaut 's data flow system in VICAT

The idea is that the user at the quarry, will connect to the expert system, abaut Analytics in order to visualize the results of the data image using their personal account that will be created for this purpose. The system will work as follows:

- The truck will arrive at the processing plant and the system will take pictures from the top of the truck e.g., every 2 seconds [or any period of time] in order to get enough picture data for detecting the pollutants at the surface of the pile. The format of each picture taken will be a JPG and they will be transmitted to the data base of abaut.
- Once the images are received at the data, the AI-ML algorithm will analyze the picture in order to detect, if it exists, certain pollutants at the materials [e.g., plastics, woods or any other residues]
- The results of the analysis will be then displayed at the web cloud front end system in order to be visualized
- The responsible person of the quarry can login at any time to see the results of the computation of the images

The format, refresh ratio of the images and the access period of time of the analysis is already not being define. The starting of this activity is planned for June 2023 so, the main actions are going to take place during the second half of 2022 and not at this early stage.

3.1.3 HOLCIM

3.1.3.1 Inventory of the existing ICT assets

The following table provides a high-level view of the expert systems, interfaces, contents summary, format, data sharing, and partners involved within this site. It enables the creation of pilot site's data flow. Please refer to D3.1 to have a detailed view of the data.

System Expert	Description of the function	Interface type provided	Content	Format	Shared data through	Shared data with
Holcim Expert System: SAP	Quarry management system (Store and Upload data)	Manual upload and download	Historical data Production data not covered by scada system General information data Datasets Images of quarry stockpiles	Xls, doc, pdf	Data Lake	MAESTRO BMT APP SIGMA
MAESTRO SCADA: Q-Production	Provide production data Maestro scada system enables access to Holcim production data from this site: https://demodeq.quarrycontrol.com to enable data visualization and comparison with actuals	REST API	Production data • Processed aggregates • Salable aggregates • Production rate index • REE • Fresh water	Json	Data Lake	HOLCIM BMT APP SIGMA
Abaut Analytics	Provide a risk map activities of workers in the surrounding of mobile machinery	Abaut Analytics web interface	Data sets Risk maps	Proprietary	Abaut analytics system	HOLCIM
MINTEK: IDEAS Andritz SW	Store optimization results	Manual upload	Studies, optimization results	xls	Data Lake	HOLCIM
BMT	Generate, store and share reports and dashboard	manual upload	Dynamic and static view of data and KPIs shared	pdf, xls	Data Lake	HOLCIM
Arco Expert system	Store weighting data	API	weighting data	json	Data lake	HOLCIM BMT

3.1.3.1.1 Data flow

The following diagram depicts the data flows between the partners or systems within this pilot site.

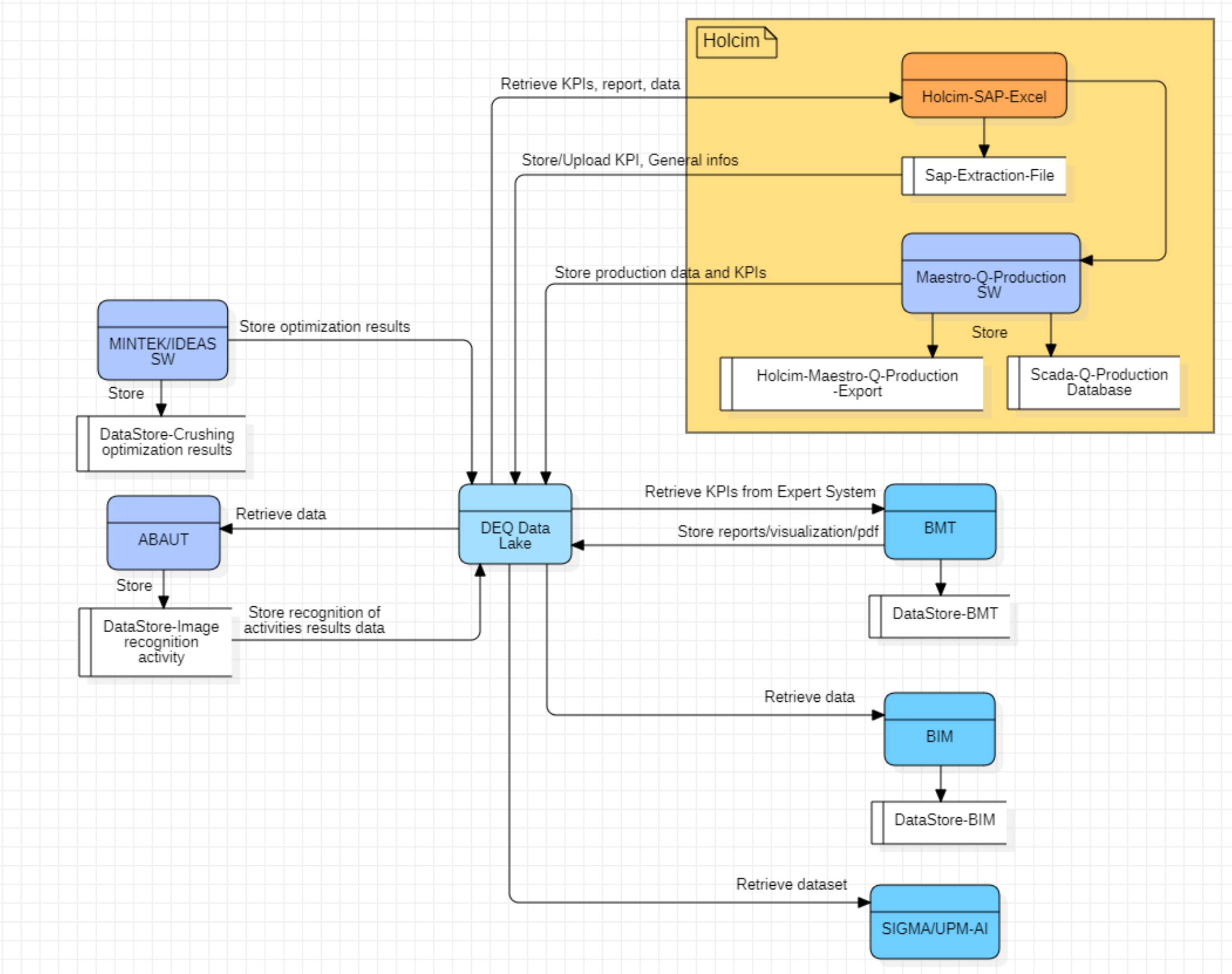

Holcim
Holcim-SAP-Excel
Sap-Extraction-File
Maestro-Q-Production SW
Store
Holcim-Maestro-Q-Production -Export
Scada-Q-Production Database
Retrieve KPIs, report, data
Store/Upload KPI, General infos
Store production data and KPIs
MINTEK/IDEAS SW
Store optimization results
Store
DataStore-Crushing optimization results
ABAUT
Retrieve data
Store
DataStore-Image recognition activity
Store recognition of activities results data
DEQ Data Lake
Retrieve KPIs from Expert System
Store reports/visualization/pdf
BMT
DataStore-BMT
Retrieve data
BIM
DataStore-BIM
Retrieve dataset
SIGMA/UPM-AI

Figure 16: Holcim's Data Flow Diagram

Within its data lake, Holcim will store general data and KPIs. The production data will come directly from the Scada system (Q-Production) via MAESTRO REST API in a JSON format; this data will be collected and stored daily. Additional data, such as optimization results from Mintek's SW will also be stored. This data will be available, according to their rights, and usable by external partners. As such, Abaut will retrieve Holcim's general information. Abaut will store in return, recognition of activities results data. Sigma/UPM-AI will retrieve dataset from Holcim to run AI services proposed by their Stockforecast tool (Stockpile volume calculation). The business management tool will retrieve KPIs to propose Business management dashboards. APP will also retrieve data from the data lake for their BIM solution. Holcim will take advantage of its data lake by retrieving KPIs, reports and processed data which will bring added value for the management of the quarry.

Note that at this stage of the project, the implementation of the Arco Weighing system in Holcim pilot site has still to be agreed. In case of agreement, the same implementation as defined in section 3.1.2.3 for Vicat is being considered for Holcim.

3.1.3.1.2 Data model

Holcim data model analysis is described jointly in Vicat data model section 3.1.2.1.2.

3.1.3.2 Models for crushing and screening optimization (KTA2.2)

Mintek is looking at building an online simulation tool that can mimic the real quarry in a pilot plant to develop and optimize its respective functions in real time. Holcim and Mintek are in communication for sample provision, plant flowsheet development and PSD measurements. Tasks of Mintek involve determination of both breakage and screening functions via experimental measurements on a pilot plant that resembles their flowsheet, validation of each function gained by Mintek's pilot plant and further simulation studies.

The data communication between Mintek and Holcim will take place through the Cloud and the intention is to store the data within the Data Lake and made available to the consortium parties interested. Optimization approach is performed using IDEAS Andritz (3rd party software). The input and outputs (data geolocation, flow of information, etc.) is exported to excel as a medium software. Excel has easy communication with IDEAS and a server computer will be placed at Mintek to conduct the simulation work.

In simulation studies, the breakage and work index functions will be incorporated to IDEAS simulation of the pilot plant and then will be calibrated against the experimental data. Once it is established that the simulation results are well comparable to the plant operational parameters, the simulation will be up scaled to real quarry flowsheet. Here the incorporation of breakage and screening functions will validate the simulation against real plant results. In each case, the model needs to be calibrated only if there were discrepancies faced. The data need to be transferred from IDEAS simulation to a local interface. This needs to follow a constant communicational link that will eventually play the transfer bridge role from IDEAS to global interface. Here, the link will direct the results to an excel file upon running of simulation.

Last stage involves development of client interface, internet communication and machine learning from global database. Here, the sole responsibility is to establish a two-way communication between client and software that new optimal plant parameters are suggested via minimizing various objective functions. This can eventually lead to extensive interpretations around maximum profits, minimum energy requirements, minimum cost associated with the plant running, size establishments and further details.

This will enable the engineers of quarry to suggest new parameters on their digital interface whereby the numbers will be sent and treated by the software connected to database and suggest the optimum conditions with minimized errors.

3.1.3.3 Devices for automation of treatment plants and storage facilities (KTA3.1)

Ma-estro will provide a web portal that permits to manage the data directly from a PLC or an industrial PC. Thanks to this system it's possible to handle many data as timing, production, consumes, alarms, maintenance, batches and so on. This kind of technology has high performance and flexibility, and quite easy to modify. There's the possibility to send commands to the plant and machines. At the begin, the signals come from sensors and reach the PLC. This tool elaborate instructions and communicate with and ethernet protocol with an industrial PC. Thanks to a modem or an internet network the PC send the data to Ma-estro's cloud and then the system replies data and instructions. The communication between this software and the Data Lake (AKKA) it is possible with API services.

A simplified form of authentication is used through a pre-shared token called qcwDeviceId. The token can be any string usually we use an MD5 hash in ASCII format for example:

D066EC6360FC1EAD2581AF031F2B39FD71B78FF751EAC409C8E26AB32909E204

This authentication mode allows access only to a reduced part of the API, for interconnection with ERP-type software or other third-party systems.

Reading of production data

Call: /qpcDB/qpcgetdatainfo [GET]

Returns an object in JSON format that contains information about the available data.

The required parameters for this call are:

qcwDeviceId (string): authentication token

idRisorsaConfigImpianto (string): identifies the system from which to take data the string is defined by MaEstro at the time of installation usually "Plant1" is used for the first system "Plant2" for the second and so on.

E.g.:

https://customername.quarrycontrol.com/qpcDB/qpcgetdatainfo?qcwDeviceId=12346&idRisorsaConfigImpianto=Plant1

The result is a JSON object.

The interesting parts of the object are the production data range: ProductionDateInterval and the flag that assure that the production database is correctly initialized DBInitialized = true.

Call: /qpcDB/qpcproduzionedettagliogiornate [GET]

Returns a string in JSON format that contains an array of objects, each one represents a single production data record for the selected plant as described below.

The required parameters for this call are:

qcwDeviceId (string): authentication token

idRisorsaConfigImpianto (string): identifies the plant from which to take data the string is defined by MaEstro at the time of installation usually "Plant1" is used for the first plant "Plant2" for the second and so on.

datareport (string): contains a date in YYYY-MM-dd format (ex: 2020-02-06) which specifies the day for which the data are desired.

The optional parameters for this call are:

TempoDal (string): contains a time in the format HH: mm: ss (ex: 09:11:33) which specifies the time of day from which you want the data.

TempoAl (string): contains a time in the format HH: mm: ss (ex: 09:11:33) which specifies the time of day to which you want the data.

E.g.:

https://customername.quarrycontrol.com/qpcDB/qpcproduzionedettagliogiornate?

qcwDeviceId=12346&idRisorsaConfigImpianto=Plant1&datareport=2020-02-06&tempoDal=09:00:00&tempoAl=10:00:00

Call: /qpcDB/qpcproduzionegetperiodo [GET]

Returns a string in JSON format that contains a single object that represents the summary of the production data of the selected plant over the required period as described below.

The data will be aggregated as appropriate. For example, for consumption and production, the data for the period will be sum. For other quantities such as absorptions and levels, you'll get the averages etc ...

The required parameters for this call are:

qcwDeviceId (string): authentication token

idRisorsaConfigImpianto (string): identifies the plant from which to take data the string is defined by MaEstro at the time of installation usually "Plant1" is used for the first plant "Plant2" for the second and so on.

dataDal (string): contains a date in YYYY-MM-dd format (ex: 2020-02-06) which specifies the starting day for data collection dataAl (string): contains a date in YYYY-MM-dd format (ex: 2020-02-06) which specifies the last day for data collection.

The optional parameters for this call are:

TempoDal (string): contains a time in the format HH: mm: ss (ex: 09:11:33) which specifies the time of day from which you want the data.

TempoAl (string): contains a time in the format HH: mm: ss (ex: 09:11:33) which specifies the time of day to which you want the data.

NB: The filter per hour is applied one by one on all the days involved, it is useful if you want to analyze how production varies in the different hours of the day or divide it into work shifts.

E.g.:

https://customername.quarrycontrol.com/qpcDB/qpcproduzionegetperiodo?qcwDeviceId=12346&idRisorsaConfigImpianto=Plant1&dataDal=2020-01-01&dataAl=2020-02- 06&tempoDal=06:00:00&tempoAl=14:00:00

Production data format

The result of the calls /qpcDB/qpcproduzioneedettagliogiornate and /qpcDB/qpcproductiongetperiodo is a string in JSON format that represents an array of production data records or an object with an aggregate version of them.

It is a simple object in which each property corresponds to a value of the production data. The Tempo property contains the date and time to which the data refers.

There are several data, some are simply calculated or cumulative values. The number of quantities and their correspondences is depending on the plant.

The scale and units of measurement depend on the measured quantity and must therefore be evaluated case-by-case.

Batch management

Call: /qbcDB/qbcbatchgetlist [GET]

Returns an array of objects in JSON format that contains the available batches.

The required parameters for this call are: qcwDeviceId (string): authentication token

filtrioIdRisorsaConfigImpianto (string): allows you to filter batches with the string that identifies the system, the string is defined by MaEstro at the time of installation usually "Plant1" is used for the first system "Plant2" for the second and so on.

filtroStato (integer): allows you to filter batches based on their status contains a numeric code (10 = All, 11 = Open and suspended, 0 = Open only, 1 = Only suspended, 2 = Only closed. It is optional by default and all)

E.g.:

https://customername.quarrycontrol.com/qbcDB/qbcbatchgetlist?qcwDeviceId=1234&filtroIdRisorsaConfigImpianto=Plant1&filtroStato=11

The data that can be interesting are:

Descrizione is the description of the batch.

DataDiRiferimento is an identification date of the batch.

ValoreObiettivo is the target value of the batch (the format, scale and unit of measurement are those specified in the Object VariabileObiettivo)[1].

[1] The target variable can be customized on the basis of the available data or their re-elaborations. It is possible to define multiple target variables, but a batch can always have only a single target.

Settings are objects that represent settings variables linked to the batch2.

Call: /qbcDB/qbcbatchinsert [POST]

Allows you to insert a new batch in the list.

The call is a POST which must have the request body in the www-form-urlencoded format. As parameters it requires:

qcwDeviceId (string): authentication token

idRisorsaConfigImpianto (string): identifies the plant, the string is defined by MaEstro at the time of installation usually "Plant1" is used for the first system "Plant2" for the second and so on.

stato (integer): batch status (0 = Open, 1 = Suspended, 2 = Closed. It is optional, the default is

Open)

descrizione (string): description of the batch

codice (string): batch code (optional)

prodotto (string): product description (optional)

dataDiRiferimento (string): reference date for the batch in ISO format (optional)

idAziendaCliente (integer): numeric id of the customer's company (optional)3

sorgenteObiettivo (string): source of the target variable. Currently the only supported is "QPC"

(optional)

nomeVariabileObiettivo (string): name of the target variable, the target variables available

depend on the system configuration. They are one or more variables that correspond to the production data seen above (optional)

valoreObiettivo (decimal number): unscaled target variable value (optional)

settings (JSON): variable object array in JSON format for the reference settings for batch4(optional)

Returns a JSON-formatted object that contains information about the batch just entered. The format is the same as the previous call only it contains a single object.

Call: /qbcDB/qbcbatchgetdefault [GET]

Returns a JSON-formatted object that contains the default batch (if available).

Whether or not to manage the predefined batches depends on the server configuration. Normally the choice of the batch to be used is made by the operator on the system, alternatively it is possible to define a default batch to be used selected on the web portal.

The required parameters for this call are:

qcwDeviceId (string): authentication token

idRisorsaConfigImpianto (string): identifies the plant, the string is defined by MaEstro at the time of installation usually "Plant1" is used for the first system "Plant2" for the second and so on.

E.g.:

2 As for the objective variables, the settings are also customizable according to need.

3 The company ID can be obtained with the call /coredata/qcwaziendagetlistshort described below.

4 The object for the settings is in the format:

[{"NomeVariabile":"setpoint00","Valore":120}, {"NomeVariabile":"setpoint01","Valore":450}]

The settings variables names depend on the installation, they are in any case fixed and are established at the time of the plant configuration.

https://customername.quarrycontrol.com/qbcDB/qbcbatchgetdefault?qcwDeviceId=1234&idRisorsaConfigImpianto=Plant1

Returns a JSON-formatted object that contains information about the newly entered batch. The format is the same as the call / qbcDB / qbcbatchgetlist only it contains a single object.

Call: /qbcDB/ qbcbatchsetdefault [POST]

Allows you to set the default batch. The call is a POST which must have the request body in the www-form-urlencoded format.

The required parameters for this call are:

qcwDeviceId (string): authentication token

idBatch (intero): numeric id of the batch to be set default dome. The id is the one found in the objects returned by the call /qbcDB/qbcbatchgetlist with the name IdBatch.

Returns a JSON-formatted object that contains true if the call was successful, false otherwise.

Call: coredata/qcwaziendagetlistshort [GET]

Returns an array of objects in JSON format that contains the companies available in the registry for

entering the batch.

The required parameters for this call are:

qcwDeviceId (string): authentication token

E.g.:

https://customername.quarrycontrol.com/coredata/qcwaziendagetlistshort?qcwDeviceId=1234

3.1.3.4 Monitoring sensors and analysing tools both for Mobile Machinery in Loading &Transport and for the recognition of workers (KTA3.2)

Abaut will install abaut mView module [see 3.1.1.7 for more information] in HOLCIM for recognizing works at the surroundings of the mobile machine or certain areas for creating and identifying in a risk map possible un-safe activity.

The idea is to identify workers in the surroundings of mobile machinery or, at the processing plant and integrate this detection and analysis in a risk map that can provide safety operational information in order to avoid accidents (when the material that is going to be processed) can affect the different processes at the processing plant.

At HOLCIM, the camera module mView will transmit the data generated to the expert system of abaut. The information will be analyzed and then displayed in Analytics for its analysis and decision-making step. This step is still under development and right now is only possible to offer a preliminary data flow system that will be tested at HOLCIM. The idea is to use the same data flow system as in VICAT due to the similarity of the activity (image recognition activity).

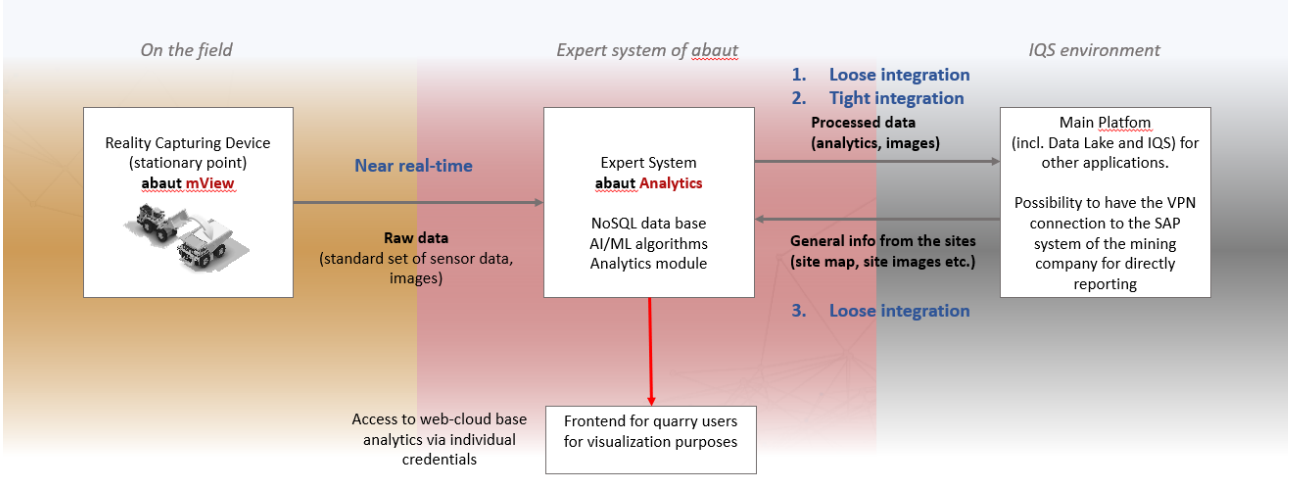

On the field
Expert system of abaut
IQS environment
Reality Capturing Device (stationary point)
abaut mView
Near real-time
Raw data (standard set of sensor data, images)
Expert System
abaut Analytics
NoSQL data base
AI/ML algorithms
Analytics module
1. Loose integration
2. Tight integration
Processed data (analytics, images)
General info from the sites (site map, site images etc.)
3. Loose integration
Main Platfom (incl. Data Lake and IQS) for other applications.
Possibility to have the VPN connection to the SAP system of the mining company for directly reporting
Access to web-cloud base analytics via individual credentials
Frontend for quarry users for visualization purposes

Figure 17: abaut' s data flow system in VICAT

The idea is that the user at the quarry, will connect to the expert system, abaut Analytics in order to visualize the results of the data image using a personal account that will be created for this purpose. The system will work as follows:

- The cameras installed in 2 mobile machines will take pictures during the working time and in order to get enough picture data for detecting the workers at the working area.
- Once the images are received at the data base, the AI-ML algorithm will analyze the picture in order to detect, the people and machines that are in that picture or series of pictures [e.g., worker, machine type, etc.]
- The results of the analysis will be then displayed at the web cloud front end system in order to be visualized and integrated in the risk map application that will be defined together with the rest of the members of WP5
- The responsible person of the quarry can login at any time to see the results of the computation of the images

The format, refresh ratio of the images and the access period of time of the analysis is already not defined. The start of this activity is planned for June 2023 so, the main actions are going to take place during the second half of 2022 and not at this early stage.

3.1.4 AGREPOR CIMPOR

3.1.4.1 Inventory of the existing ICT assets

The following table provides a high-level view of the expert systems, interfaces, contents summary, format, data sharing, and partners involved within this site. It enables the creation of pilot site's data flow. Please refer to D3.1 to have a detailed view of the data.

System Expert	Description of the function	Interface type provided	Content	Format	Shared data through	Shared data with
Cimpor Expert System: SAP	Store and Upload data	Manual upload and download	Historical data Production data General information data	xls	Data Lake	BMT APP SIGMA: StockForecast ABAUT: abaut Analytics

System Expert	Description of the function	Interface type provided	Content	Format	Shared data through	Shared data with
Cimpor Expert System: KoBotoolbox	Create and store data	API/Manual upload and download	Energy consumption and usage of their mobile machineries and crushers	xls	Data Lake	BMT SIGMA
Cimpor Expert System: Scada	Store production data	API	Production KPIs	Standard /Proprietary	Data Lake	BMT APP SIGMA: StockForecast ABAUT: abaut Analytics
ABAUT	Product mass flow Fleet performance Reports External transport/logistics performance	Manual and automatic upload / download	Same data as in Hanson pilot site	Standard /Proprietary	Data shared in Abaut expert system and in data lake	CIMPOR BMT
BMT	Generate, store and share reports and dashboard	manual upload	Dynamic and static view of data and KPIs shared	pdf, xls	Data Lake	CIMPOR

3.1.4.1.1 Data flow

The following diagram depicts the data flows between the partners or systems within this pilot site.

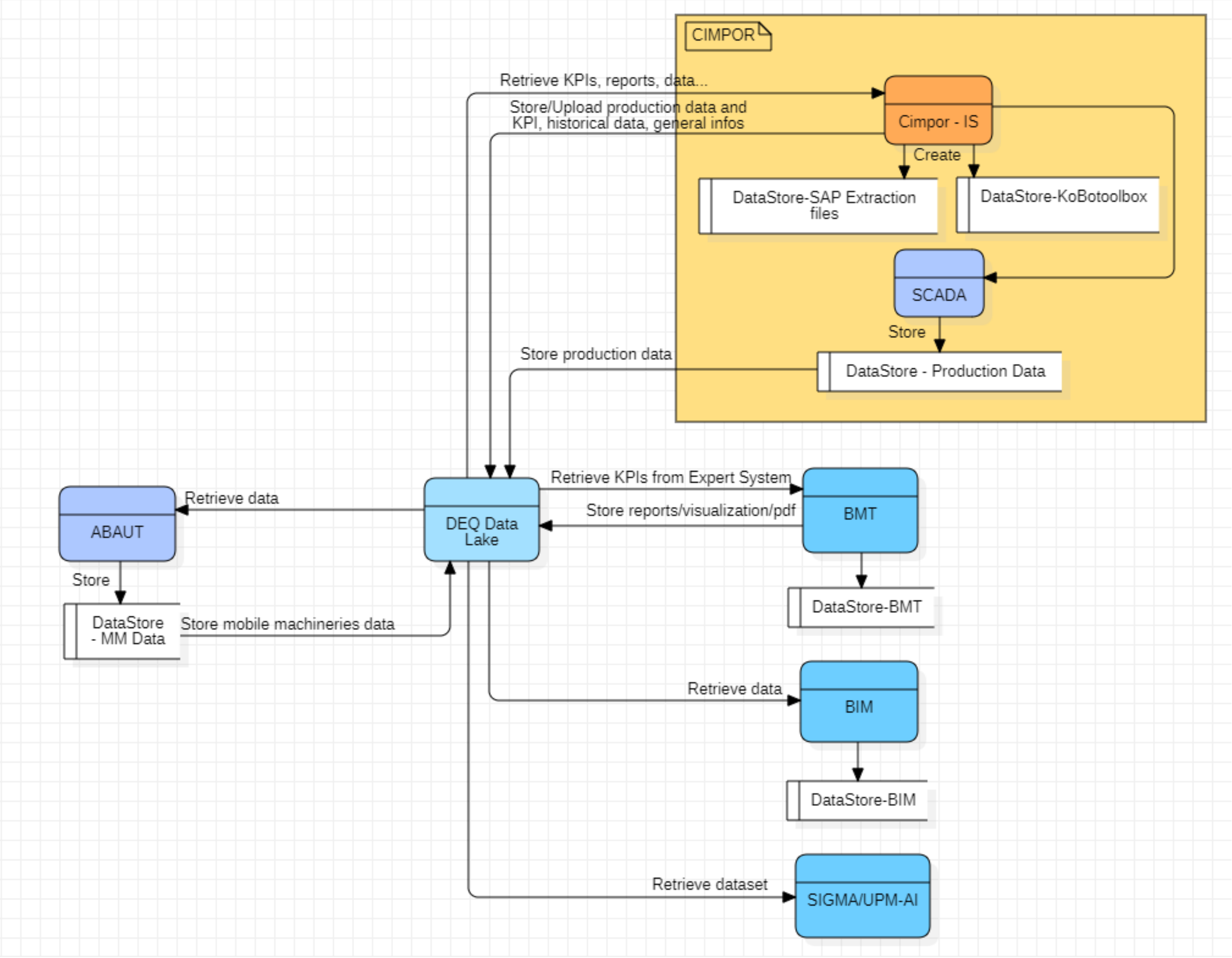

CIMPOR
Retrieve KPIs, reports, data...
Store/Upload production data and KPI, historical data, general infos
Cimpor - IS
Create
DataStore-SAP Extraction files
DataStore-KoBotoolbox
SCADA
Store
Store production data
DataStore - Production Data
ABAUT
Retrieve data
DEQ Data Lake
Retrieve KPIs from Expert System
Store reports/visualization/pdf
BMT
Store
DataStore - MM Data
Store mobile machineries data
DataStore-BMT
Retrieve data
BIM
DataStore-BIM
Retrieve dataset
SIGMA/UPM-AI

Figure 18: Agrepor Cimpor' s Data Flow Diagram

Within its data lake, Cimpor will store general information, historical data and production data related to the energy consumption and usage of their mobile machineries and crushers. In a near future, additional data, such as production KPIs data, from CIMPOR's scada system will also be stored. This data will be available, according to their rights, and usable by external partners. As such, Abaut will retrieve Cimpor' s general information; Abaut will store KPIs related to Cimpor' s mobile machineries and also reports related to the external transport. Sigma/UPM-AI will retrieve the necessary historical data from Cimpor to run AI services proposed by their Stockforecast tool (consumption and product forecasting). The business management tool will retrieve KPIs to propose Business management dashboards. APP will also retrieve data from the data lake for their BIM solution. Cimpor will take advantage of its data lake by retrieving KPIs, reports and processed data which will bring added value for the management of the quarry.

Note that at this stage of the project, the implementation of the Arco Weighing system in Cimpor pilot site has still to be agreed. In case of agreement, the same implementation as defined in section 3.1.2.3 for Vicat is being considered for Cimpor.

3.1.4.2 Monitoring sensors and analysing tools both for Mobile Machinery in Loading &Transport and for the recognition of workers (KTA3.2)

The activities of Abaut in CIMPOR are the same ones than in HANSON. This is due to the idea of replicate and compare the activities measuring the KPI's of both quarries under the same principles. The idea is also to create a digital model of the quarry under the same rationale [see point 3.1.1.7].

Something important to highlight is that the analysis of the outbound logistic will also be considered at this task. Abaut will deploy some light version of Abaut Edge in order to analyze the external logistic of the delivery transport service of CIMPOR. This activity will offer new insights regarding the performance of the external logistics, and it is interpretation inside DEQ.

This task has not started yet and is intended to start at the beginning of 2023. The concept has not been studied yet so. due to this reason is not possible to offer more information.

3.1.5 CSI

3.1.5.1 Inventory of the existing ICT assets

The following table provides a high-level view of the expert systems, interfaces, contents summary, format, data sharing, and partners involved within this site. It enables the creation of pilot site's data flow. Please refer to D3.1 to have a detailed view of the data.

System Expert	Description of the function	Interface type provided	Content	Format	Shared data through	Shared data with
CSI Expert System: SAP	Quarry management system: Store and Upload data	Manual upload and download	Historical data Production data and KPIs General information data	Xls, pdf	Data Lake	BMT APP SIGMA: Predictive Maintenance DH&P: SmartQuarry
Primary Crusher Controller (Primary Crusher, Belt scales)	Retrieve process parameters	PLC/API Software Exported files	Power consumption Crusher settings Engine hours Working hours Mass	Data sets via API, csv	DHP expert system, data lake on demand (manual upload files or API)	DHP expert system
Fuel consumption monitoring software (Pandora Soft)	Retrieve fuel data and engine hours	API	Fuel consumption per machine Engine hours per machine	Data sets via API, csv	DHP expert system, data lake on demand (manual upload files or API)	DHP expert system
DH&P Expert System: SmartQuarry	Fleet performance monitoring: Store KPIs/data	API	Mobile Machineries KPIs	json	Data Lake	BMT APP SIGMA: Predictive Maintenance CSI

System Expert	Description of the function	Interface type provided	Content	Format	Shared data through	Shared data with
BMT	Generate, store and share reports and dashboard	manual upload	Dynamic and static view of data and KPIs shared	pdf, xls	Data Lake	CSI

3.1.5.1.1 Data flow

The following diagram depicts the data flows between the partners or systems within this pilot site.

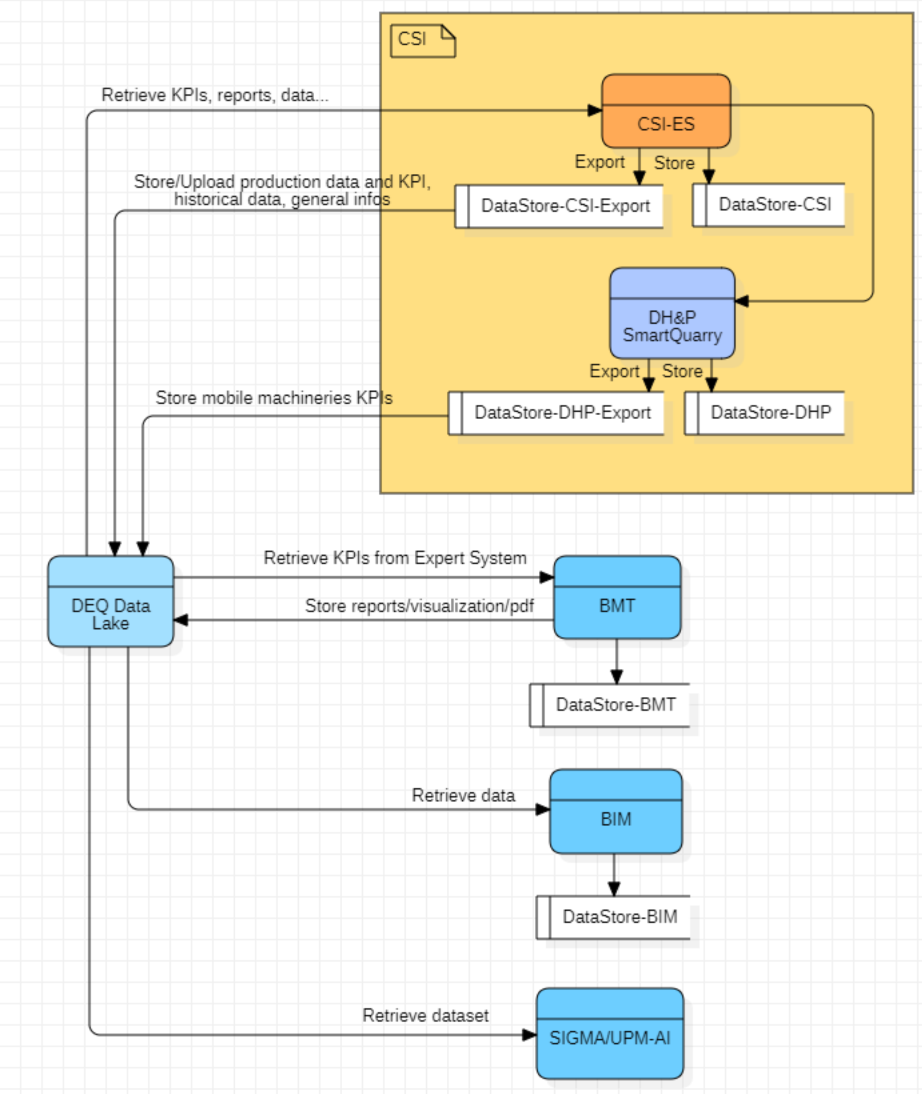


Figure 19: CSI's Data Flow Diagram

Within CSI data lake, CSI's mobile machineries KPIs/Data will be stored by DH&P. CSI will store general information, historical data and production data. This data will be available, according to their rights, and usable by external partners. As such, Sigma/UPM-AI will retrieve the necessary dataset from CSI to train AI models proposed by their Predictive

Maintenance tool. The business management tool will retrieve KPIs to propose Business management dashboards. APP will also retrieve data from the data lake for their BIM solution. CSI will take advantage of its data lake by retrieving KPIs, reports and processed data which will bring added value for the management of the quarry.

3.1.5.1.2 Data model

DH&P will implement in their expert system a data model that defines and correlates:

- Plants (here: single plant in expert system) + descriptions like type, country, customer
- Assets in plant + describing attributes and classifications (like type, OEM, model)
  - Optionally assets can be furthermore split up into sub-components.
- Parameter definitions including unit, interval…
- Asset parameters like position, fuel consumption, working hours etc., linked to a parameter definition and an asset
- Units with base unit and conversion formulas (e.g., m/s -> km/h)

Each artifact (asset, parameter) will be identifiable by a unique ID which can be used by external systems to query data via the API.

After the analysis of Mobile equipment data provided by DH&P, we created a first data model to be deployed within the IQS shown in the following diagram.

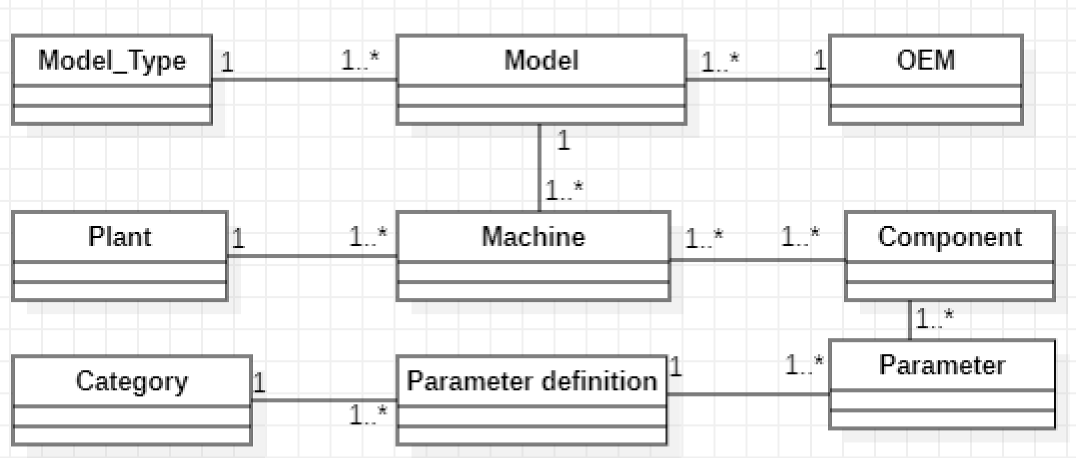


Figure 20: Mobile machinery data model

3.1.5.2 Mobile equipment & quarry geological deposit digitalisation & real-time modelling (KTA3.3)

An API will be provided to query:

- the available assets per plant
- parameters per asset
- parameter values by asset-parameter and time frame (see D3.1)

This model and API enable the data lake or 3rd parties to exchange the model definition and merge the data from different systems in the data lake and business management tools.

3.2 Data lakes

3.2.1 Results of the Benchmark for the best data lake tools

The full benchmark's study results done by AKKA are available in Appendix 7.1. Here below is a synthesis of the main results related to the data lake components.

Here is a global view of the components that will be used to build IQS data lakes:

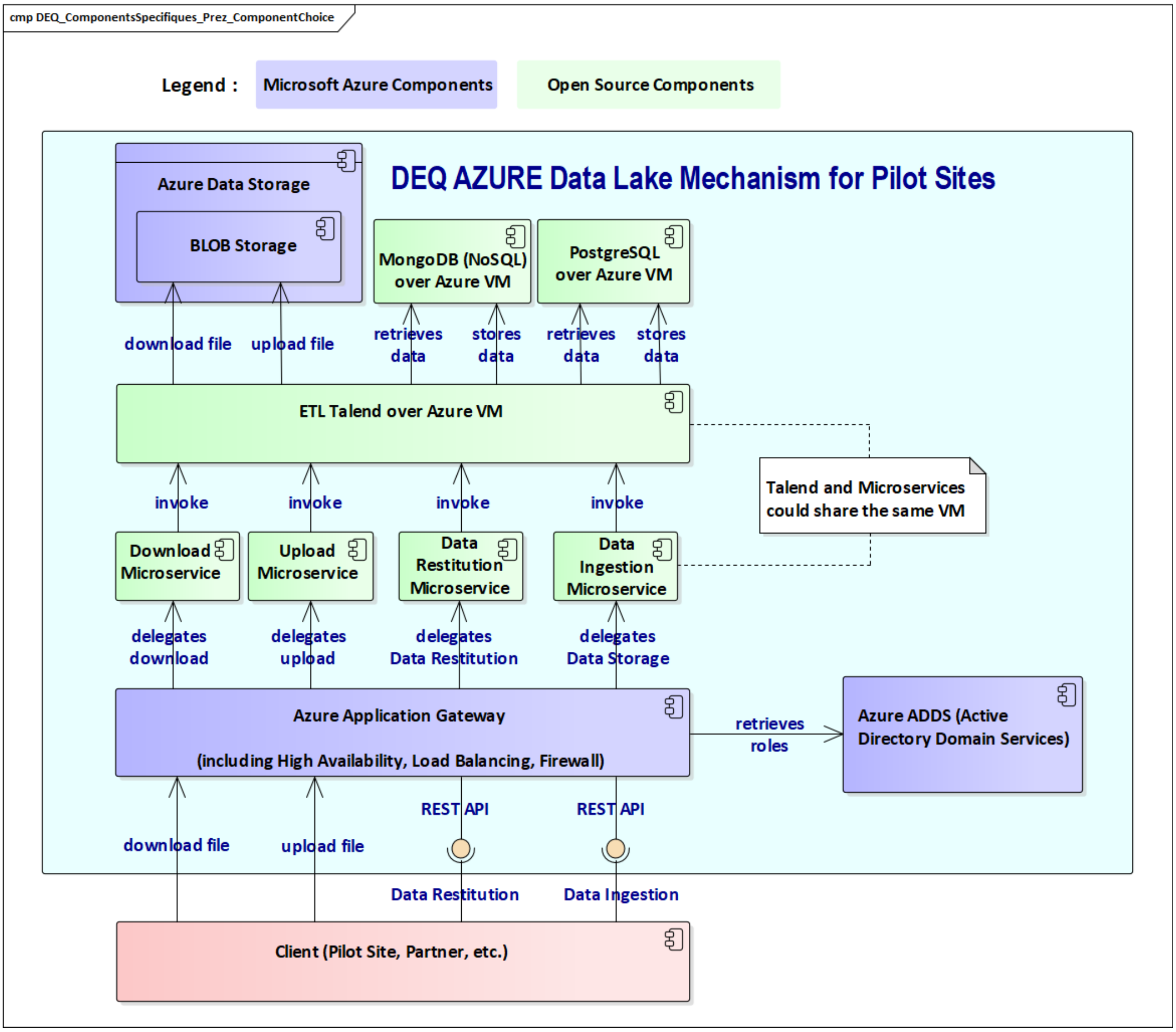

cmp DEQ_ComponentsSpecifiques_Prez_ComponentChoice
Legend :
Microsoft Azure Components
Open Source Components
DEQ AZURE Data Lake Mechanism for Pilot Sites
Azure Data Storage
BLOB Storage
MongoDB (NoSQL) over Azure VM
PostgreSQL over Azure VM
download file
upload file
retrieves data
stores data
retrieves data
stores data
ETL Talend over Azure VM
Talend and Microservices could share the same VM
invoke
invoke
invoke
invoke
Download Microservice
Upload Microservice
Data Restitution Microservice
Data Ingestion Microservice
delegates download
delegates upload
delegates Data Restitution
delegates Data Storage
Azure Application Gateway
(including High Availability, Load Balancing, Firewall)
retrieves roles
Azure ADDS (Active Directory Domain Services)
REST API
REST API
download file
upload file
Data Restitution
Data Ingestion
Client (Pilot Site, Partner, etc.)

Figure 21: Diagram of the components selected for the Data Lakes architecture

This architecture including many open-source components will significantly reduce the operating costs, will provide the most flexible architecture and the most reversible solution, but will generate additional development costs. Note that these additional costs would remain the same with Azure Logic Apps and App Service solution.

The global cost is estimated to be less than 500€ per month per each pilot site:

Azure Application Gateway	Azure ADDS	Talend + Microservice on Azure VM	Data Lake Storage Gen2	PostgreSQL	Deployment over Azure VM by Ansible	VM Creation & Monitoring	TOTAL
60€	70€	180€	50€	100€	Non-recurring cost	Embedded into Azure offer	**< 500 € / month**

Below, some details are given for each component:

- Description
- Metrics
- Costs

Component: Azure Application Gateway

Azure Application Gateway

(including High Availability, Load Balancing, Firewall)

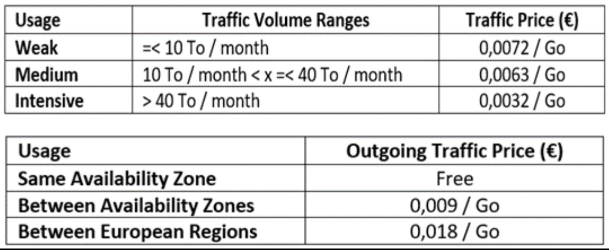
Description: This is the Frontal Gateway of DigiEcoQuarry Application. It exposes some specific REST API Web Services as:

- Data Ingestion
- Data Restitution

and some HTTP Requests as:

- upload files
- download files that have been uploaded

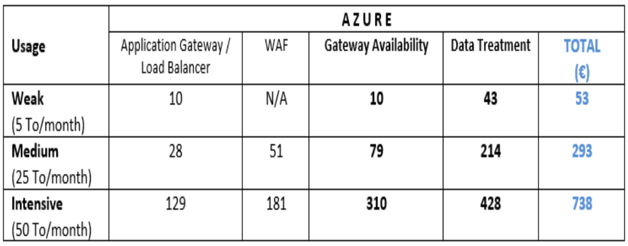
Metrics:



Usage	Traffic Volume Ranges	Traffic Price (€)
Weak	=< 10 To / month	0,0072 / Go
Medium	10 To / month < x =< 40 To / month	0,0063 / Go
Intensive	> 40 To / month	0,0032 / Go

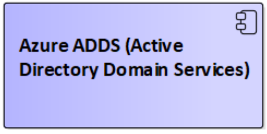
Usage	Outgoing Traffic Price (€)
Same Availability Zone	Free
Between Availability Zones	0,009 / Go
Between European Regions	0,018 / Go

Costs:



Usage	AZURE				
	Application Gateway / Load Balancer	WAF	Gateway Availability	Data Treatment	TOTAL (€)
Weak (5 To/month)	10	N/A	**10**	**43**	53
Medium (25 To/month)	28	51	**79**	**214**	293
Intensive (50 To/month)	129	181	**310**	**428**	738

Component: Azure Active Directory Domain Services for security and roles management



Azure ADDS (Active Directory Domain Services)

Description: In AD, Users, User Groups, Roles, Applicative Rights have been declared.

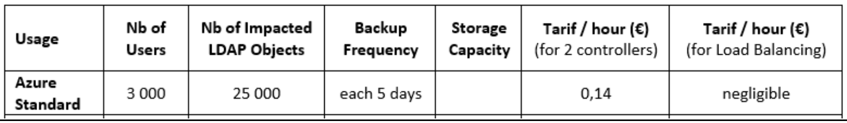
- Users have been gathered into User Groups
- Applicative Rights have been gathered into Roles
- Roles have been assigned to User Groups

A token exchange between DEQ Clients and the Domain Controller should be implemented when client machine starts, using Kerberos Protocol. This implementation is strongly secured but is expensive in terms of development. Firstly, for a POC solution, the Authentication could be managed with a simple check of the couple (Username, Password) through AD.

Any Requests entering the Cloud through the Gateway, embeds a username and an encrypted password. A dedicated Application (to be developed)

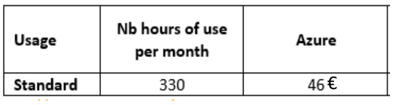
- authenticates the User by checking the validity of the couple (Username, Password)
- retrieves the LDAP Roles of the connected User
- checks if the User has the rights, according to its assigned Roles, to execute what he requests

Metrics:



Usage	Nb of Users	Nb of Impacted LDAP Objects	Backup Frequency	Storage Capacity	Tarif / hour (€) (for 2 controllers)	Tarif / hour (€) (for Load Balancing)
Azure Standard	3 000	25 000	each 5 days		0,14	negligible

Costs: 5/7 – 15/24 (from 5h to 20h)



Usage	Nb hours of use per month	Azure
Standard	330	46 €

Component: Microservices

Download Microservice
Upload Microservice
Data Restitution Microservice
Data Ingestion Microservice

Description: Microservice–based architecture will be highly adopted to implement different services.
All these microservices could be deployed on the same Talend VM to reduce costs.
This architecture including many open-source components will significantly reduce the **operating costs** and will provide the most flexible architecture and the most **reversible** solution. However, it can generate additional development costs. Note nevertheless that development costs would not be lower if Azure components (as Azure Logic Apps or App Service solution) were used.

Metrics:

Number of microservices

Development cost

Costs:
Microservices deployed on the same Talend VM to reduce costs.
Some microservices will be developed withing the scope of the CDMP. PostGreSql database will be used as database for the CDMP

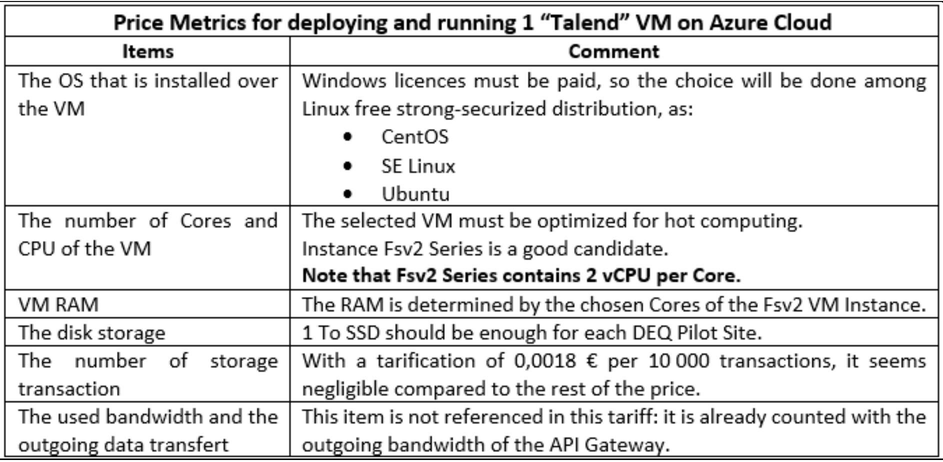
Component: Talend (ETL tool)

Description: Talend OS ESB (Open Studio Enterprise Server Bus) will be used for development. Talend runtime will be deployed over a dedicated Azure VM. Talend OS ESB:

- performs Extraction, Transformation, Loading of large data sets
- provides trigger connectors when REST API or HTTP Requests are consumed and any other connectors to connect any storages

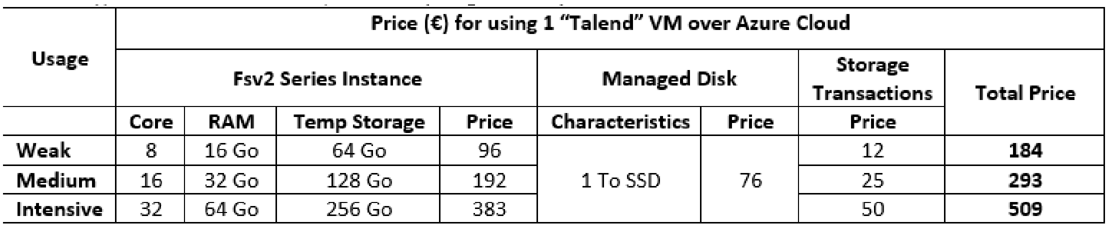
Among the tasks it performs, Talend must determine where the file must be dropped down or retrieved, based on the file nomenclature or its related metadata.

Metrics: The price of a VM over Azure Cloud is determined through these metrics:



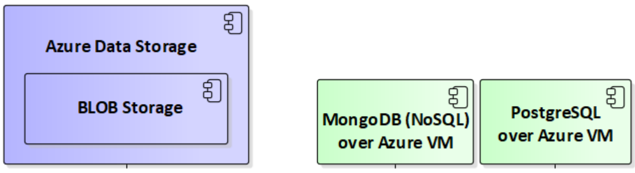
Price Metrics for deploying and running 1 "Talend" VM on Azure Cloud	
Items	**Comment**
The OS that is installed over the VM	Windows licences must be paid, so the choice will be done among Linux free strong-securized distribution, as: • CentOS • SE Linux • Ubuntu
The number of Cores and CPU of the VM	The selected VM must be optimized for hot computing. Instance Fsv2 Series is a good candidate. **Note that Fsv2 Series contains 2 vCPU per Core.**
VM RAM	The RAM is determined by the chosen Cores of the Fsv2 VM Instance.
The disk storage	1 To SSD should be enough for each DEQ Pilot Site.
The number of storage transaction	With a tarification of 0,0018 € per 10 000 transactions, it seems negligible compared to the rest of the price.
The used bandwidth and the outgoing data transfert	This item is not referenced in this tariff: it is already counted with the outgoing bandwidth of the API Gateway.

Costs: To minimize the cost, a 3-year reserved VM instance is chosen.



Usage	Price (€) for using 1 "Talend" VM over Azure Cloud							
	Fsv2 Series Instance				Managed Disk		Storage Transactions	Total Price
	Core	RAM	Temp Storage	Price	Characteristics	Price	Price	
Weak	8	16 Go	64 Go	96	1 To SSD	76	12	**184**
Medium	16	32 Go	128 Go	192			25	**293**
Intensive	32	64 Go	256 Go	383			50	**509**

Component: Azure Data Storage, Mongo DB, PostgreSQL (Storage tools)



Description: If some Pilot Sites need it, File Storage will be used to store files directly from a local File System (LAN) of Pilot Sites (via SMB Protocol).

Detailed specifications will determine if a NoSQL Database is necessary for the project. In that case, PostgreSQL and MongoDB might be hosted by the same VM.

Metrics:

Usage	Azure	Open-Source
Weak	· Storage 200 Go · 10^4 writes · 10^6 reads	1 D4s v3, 1 HDD S4, 1 year
Medium	· Storage 2 To · snapshot 100 Go · 10^6 writes · 10^7 reads	1 D4s v3, 1 SSD E6, 1 year
Intensive	· Storage 10 To · 10^6 writes · 10^7 reads	1 D8s v3, 2 SSD P10, 1 year

Costs:

Usage	Azure	
Weak	BLOB Storage on General purpose storage account v2	4,47 € / month
Medium		44,98 € / month
Intensive		186,50 € / month
Weak	Data Lake Storage Gen2	4,60 € / month
Medium		36,53 € / month
Intensive		178,05 € / month
Weak	File Storage	13,55 € / month
Medium		147,20 € / month
Intensive		698,28 € / month

Usage	Open-Source (on VM over Azure)	
Weak	PostGre SQL	97,83 € / month
Medium		101,19 € / month
Intensive		232,25 € / month
Weak	MongoDB *NoSQL*	97,83 € / month
Medium		101,19 € / month
Intensive		232,25 € / month

3.2.2 Interface with the IQS:

3.2.2.1 Centralized Data management Platform (CDMP)

The CDMP is a centralised platform to be developed by AKKA using open-source frameworks. It aims to collect and store data from the Pilot Sites, and allows IQS and quarry partners to browse, access and download data. The data will be associated with metadata -data description-, stored in a database, used to fetch, and retrieve data. Data itself will be stored in a data lake.

Uploading, browsing, accessing, and downloading data will be done using REST APIs provided by the CDMP.

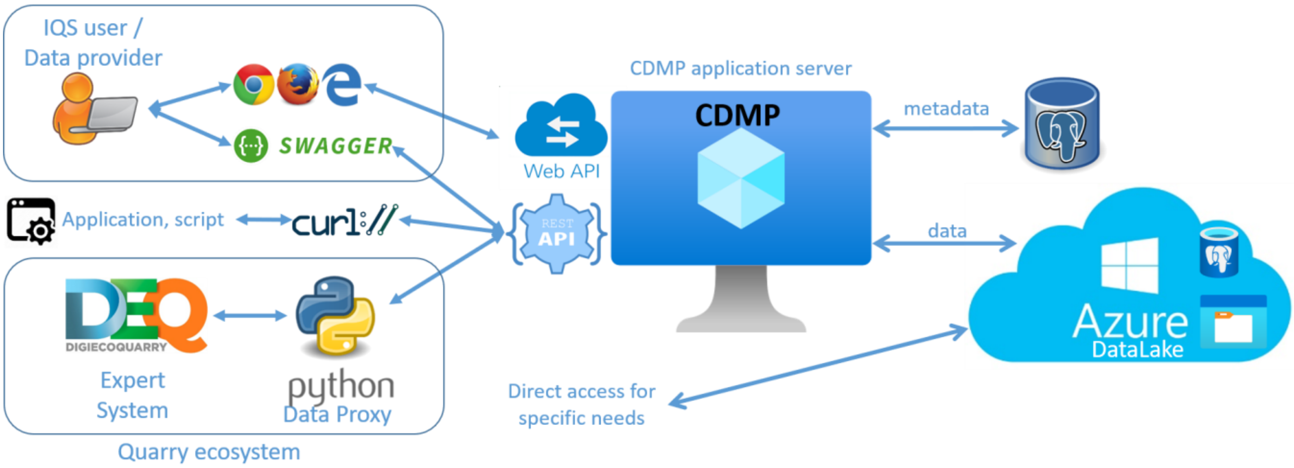


Figure 22: CDMP general architecture

Data will be uploaded to CDMP along with metadata, an accurate and complete description of the data, formalized in an XML description file. A common description model will be agreed between partners, based on Pilot Site, processes, etc.

This metadata, stored in a dedicated database, will be used to organize data in the data lake containers and databases, and later to browse and retrieve data. The data lake containers and databases will be created following data models defined by data providers, enabling cross access to data coming from different processes. Access to data lake containers will be granted following an authorization policy defined by PS and partners.

The development of CDMP architecture on top of the data lake was made because of the necessity to have an adequate data management system of metadata required for data upload, download and data sharing. Besides, the implementation of metadata using Azure tools appears to be complex, requiring a huge effort of development using expensive tools and services (LogicApp, Azure functions, data factory, data bases) and above all, we would like to avoid a lock-in situation. Vendor lock-in, also called proprietary lock-in or customer lock-in, is a technique used by some technology vendors to make their customers dependent on them for products and services by making it hard to switch to a competitor without substantial costs or difficulty.

The CDMP will provide a web interface enabling access to uploaded data, and data sharing between partners of PS according to authorizations defined by PS.

For specific needs, and usage not covered by CDMP mechanism, PS and partners can use directly low-level data lake API to take advantage of data lake features.

The data lake specifications will follow the benchmark recommendations (see Annex).

3.2.2.2 Data lake interface

Besides CDMP, there are several ways to upload data, navigate in the data lake, and download data from the data lake. In the DIGIECOQUARRY context, the main drawback of using Azure built-in or third-party interfaces is that there is no management of any associated metadata: data uploaded using these means won't be described in a metadata database, and to retrieve data, one must know what he is searching for. Therefore, these ways of accessing the data lake will be reserved for special needs, the CDMP being the recommended way for nominal or customized usages.

Interfaces provided by Azure data lake to store and share data

Users authenticated by Azure Active Directory, and granted to access data lake and containers, can explore the data lake, and manage data using their Internet browser.

Other users can upload data using either a copy tool "**AzCopy**"[1], or an SFTP connection. In this late case, data lake must have been configured to authorize SFTP. Both solutions require authentication strings or SSH keys.

Alternative tool for data lake exploration

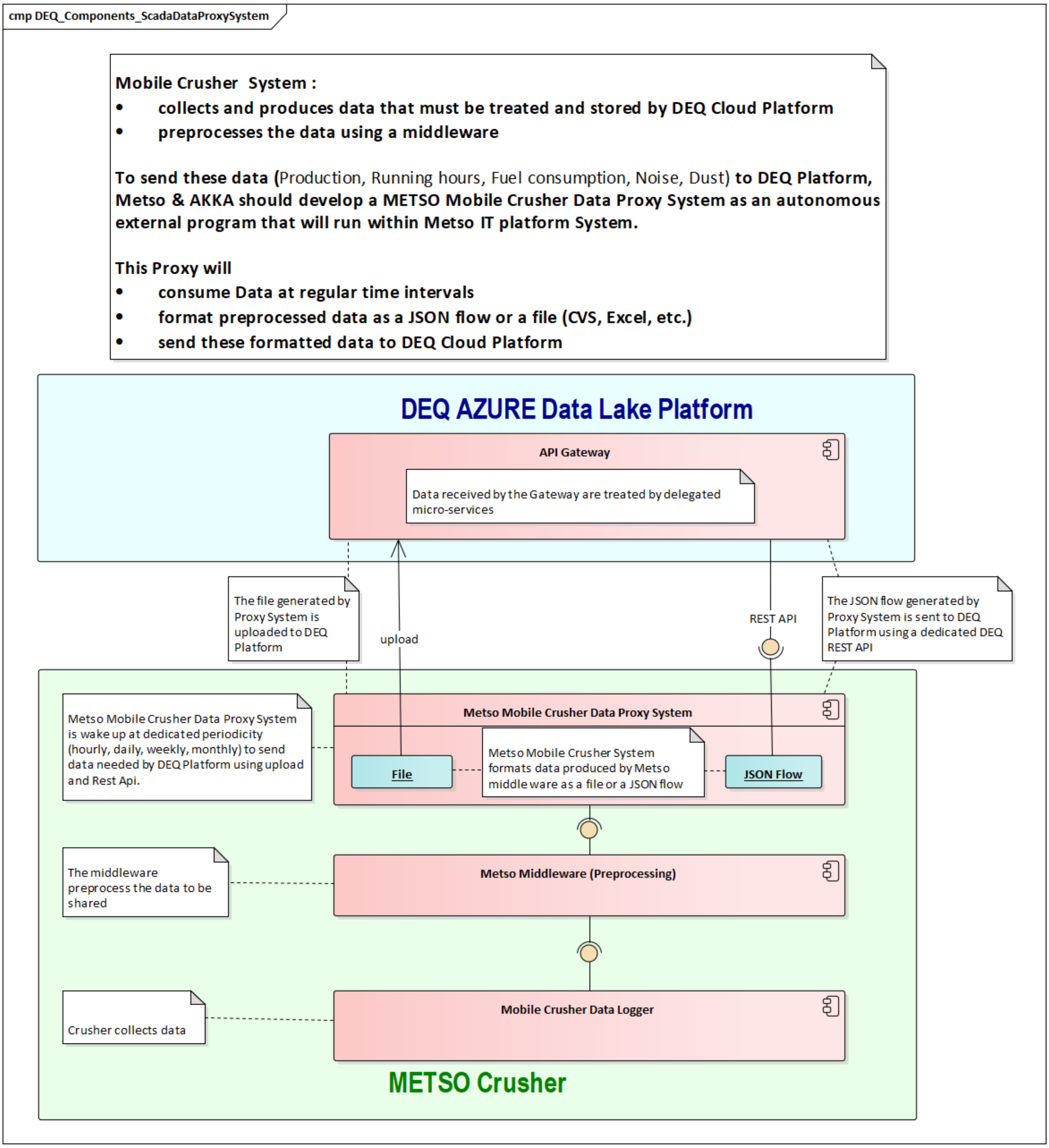
Microsoft Azure provides a standalone application, the "**Storage Explorer**"[2], that can be used to explore, upload, or download data from the data lake (an "Azure Storage account"). Using a Shared Access Signature, provided by AKKA, one can connect the Storage Explorer to specific containers in the data lake. Once connected, the Storage Explorer allows browsing and managing data as a simple file explorer.

3.2.2.3 Interface with Mobile crusher system developed by Metso [KTA 2.1]

The following picture depicts the mechanism that will be used to connect the IQS with the Metso mobile crusher system developed by Metso. A data pull process is defined to extract data at regular basis from Metso's middleware system and upload the data to the data lake. After the data transfer the data will be used by the business management tools.

[1] See https://docs.microsoft.com/en-us/azure/storage/common/storage-use-azcopy-v10

[2] See https://docs.microsoft.com/en-us/azure/vs-azure-tools-storage-manage-with-storage-explorer



cmp DEQ_Components_ScadaDataProxySystem
DEQ AZURE Data Lake Platform
API Gateway
upload
REST API
Metso Mobile Crusher Data Proxy System
File
JSON Flow
Metso Middleware (Preprocessing)
Mobile Crusher Data Logger
METSO Crusher

Mobile Crusher System :

- **collects and produces data that must be treated and stored by DEQ Cloud Platform**
- **preprocesses the data using a middleware**

To send these data (Production, Running hours, Fuel consumption, Noise, Dust) **to DEQ Platform, Metso & AKKA should develop a METSO Mobile Crusher Data Proxy System as an autonomous external program that will run within Metso IT platform System.**

This Proxy will

- **consume Data at regular time intervals**
- **format preprocessed data as a JSON flow or a file (CVS, Excel, etc.)**
- **send these formatted data to DEQ Cloud Platform**

Data received by the Gateway are treated by delegated micro-services

The file generated by Proxy System is uploaded to DEQ Platform

The JSON flow generated by Proxy System is sent to DEQ Platform using a dedicated DEQ REST API

Metso Mobile Crusher Data Proxy System is wake up at dedicated periodicity (hourly, daily, weekly, monthly) to send data needed by DEQ Platform using upload and Rest Api.

Metso Mobile Crusher System formats data produced by Metso middle ware as a file or a JSON flow

The middleware preprocess the data to be shared

Crusher collects data

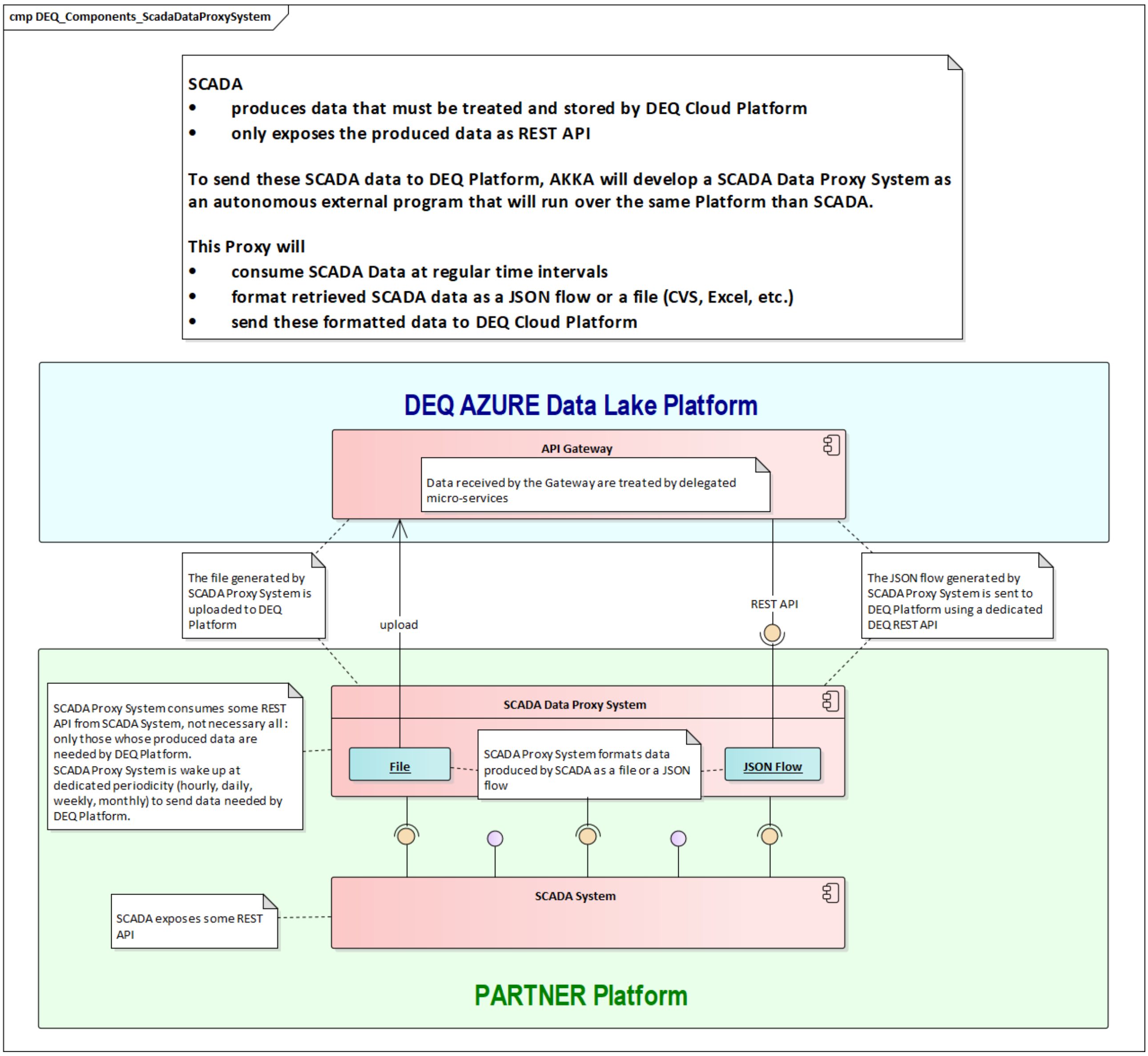
Figure 23: Mechanism used to connect the IQS with the Metso mobile crusher system

3.2.2.4 Interface with MINTEK’s simulation platform for crushing and screening optimization [KTA 2.2]

In HOLCIM pilot site, the IQS will provide the data management tools to enable data sharing with Mintek. The same concept of data sharing as the one described for HANSON will be used to share a file or a set of files related to the crushing and screening process. The data to be shared includes excel file, report results, crusher configuration file, etc. Data will be organized by Pilot Site, by process and by additional metadata to be defined to store the data.

3.2.2.5 Interface with automation & scada system developed by MAESTRO [KTA 3.1]

The following picture depicts the mechanism that will be used to connect the IQS with the SCADA system developed by MAESTRO and used to control the production. A data pull process is defined to extract data at regular basis from QProduction system and upload the data to the data lake. After the data transfer the data will be used by the business management tools.



SCADA
- produces data that must be treated and stored by DEQ Cloud Platform
- only exposes the produced data as REST API

To send these SCADA data to DEQ Platform, AKKA will develop a SCADA Data Proxy System as an autonomous external program that will run over the same Platform than SCADA.

This Proxy will
- consume SCADA Data at regular time intervals
- format retrieved SCADA data as a JSON flow or a file (CVS, Excel, etc.)
- send these formatted data to DEQ Cloud Platform

Data received by the Gateway are treated by delegated micro-services

The file generated by SCADA Proxy System is uploaded to DEQ Platform

The JSON flow generated by SCADA Proxy System is sent to DEQ Platform using a dedicated DEQ REST API

SCADA Proxy System consumes some REST API from SCADA System, not necessary all : only those whose produced data are needed by DEQ Platform. SCADA Proxy System is wake up at dedicated periodicity (hourly, daily, weekly, monthly) to send data needed by DEQ Platform.

SCADA Proxy System formats data produced by SCADA as a file or a JSON flow

SCADA exposes some REST API

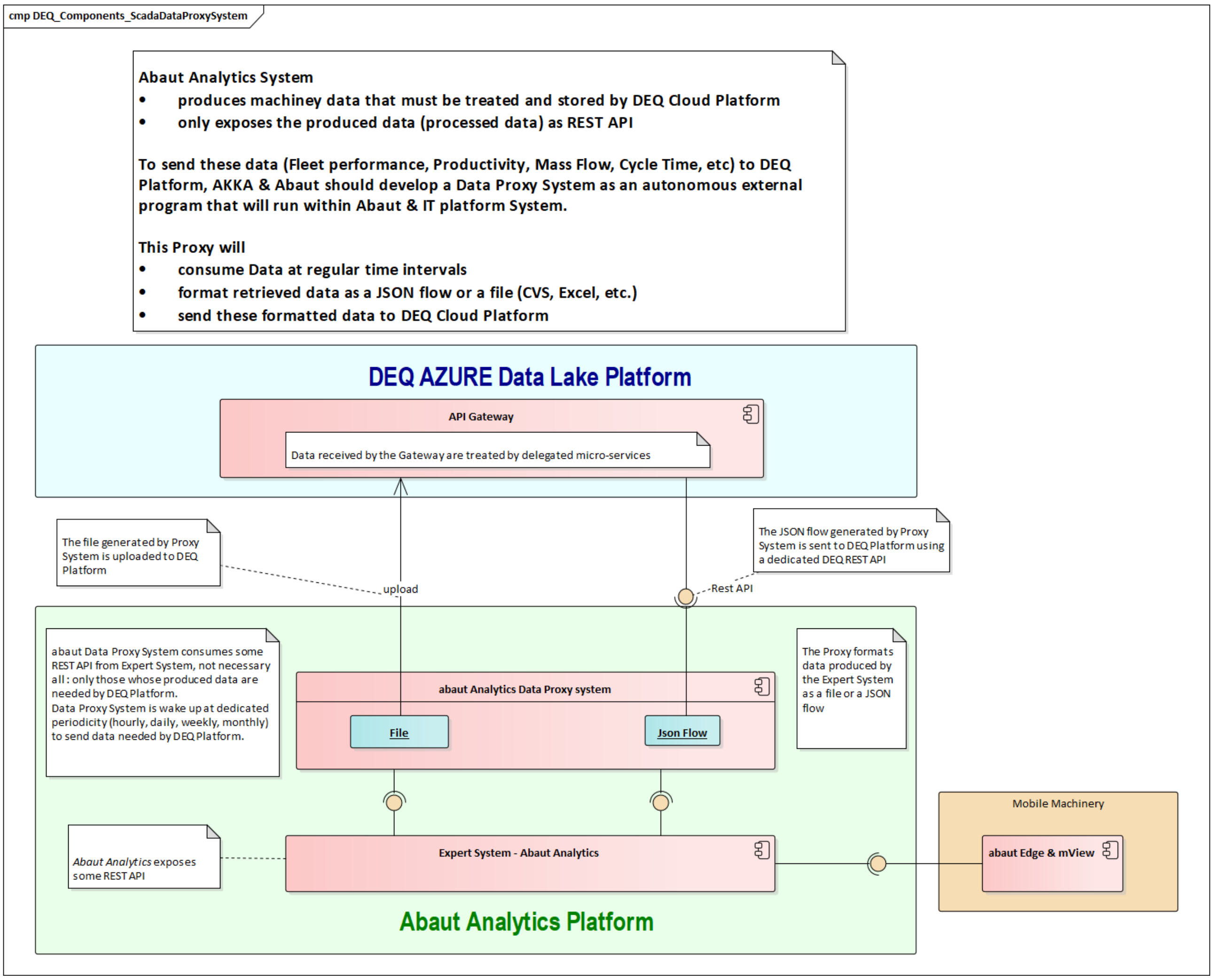
Figure 24: Mechanism used to connect the IQS with the SCADA system developed by MAESTRO

3.2.2.6 Interface with automation & SCADA system developed by ARCO [KTA 3.1]

The same data proxy pattern defined in 3.2.2.5 Interface with automation & scada system developed by MAESTRO [KTA 3.1] for Maestro will be used to collect data from ARCO's system and store data within the data lake.

3.2.2.7 Interface with Abaut analytics system [KTA 3.2]

The following picture depicts the mechanism that will be used to connect the IQS with the Abaut analytics system used to control the fleet performance. A data pull process is defined to extract data at regular basis from Abaut analytics system and upload the data to the data lake. After the data transfer the data will be used by the business management tools.


cmp DEQ_Components_ScadaDataProxySystem
DEQ AZURE Data Lake Platform
API Gateway
upload
Rest API
abaut Analytics Data Proxy system
File
Json Flow
Expert System - Abaut Analytics
Abaut Analytics Platform
Mobile Machinery
abaut Edge & mView

Abaut Analytics System

- produces machiney data that must be treated and stored by DEQ Cloud Platform
- only exposes the produced data (processed data) as REST API

To send these data (Fleet performance, Productivity, Mass Flow, Cycle Time, etc) to DEQ Platform, AKKA & Abaut should develop a Data Proxy System as an autonomous external program that will run within Abaut & IT platform System.

This Proxy will

- consume Data at regular time intervals
- format retrieved data as a JSON flow or a file (CVS, Excel, etc.)
- send these formatted data to DEQ Cloud Platform

Data received by the Gateway are treated by delegated micro-services

The file generated by Proxy System is uploaded to DEQ Platform

The JSON flow generated by Proxy System is sent to DEQ Platform using a dedicated DEQ REST API

abaut Data Proxy System consumes some REST API from Expert System, not necessary all : only those whose produced data are needed by DEQ Platform.
Data Proxy System is wake up at dedicated periodicity (hourly, daily, weekly, monthly) to send data needed by DEQ Platform.

The Proxy formats data produced by the Expert System as a file or a JSON flow

Abaut Analytics exposes some REST API

Figure 25: Mechanism used to connect the IQS with the Abaut analytics system

After the analysis of Transport Plant production data provided by Vicat (flat files) and Holcim (Scada) we created a first data model shown in the following diagram:

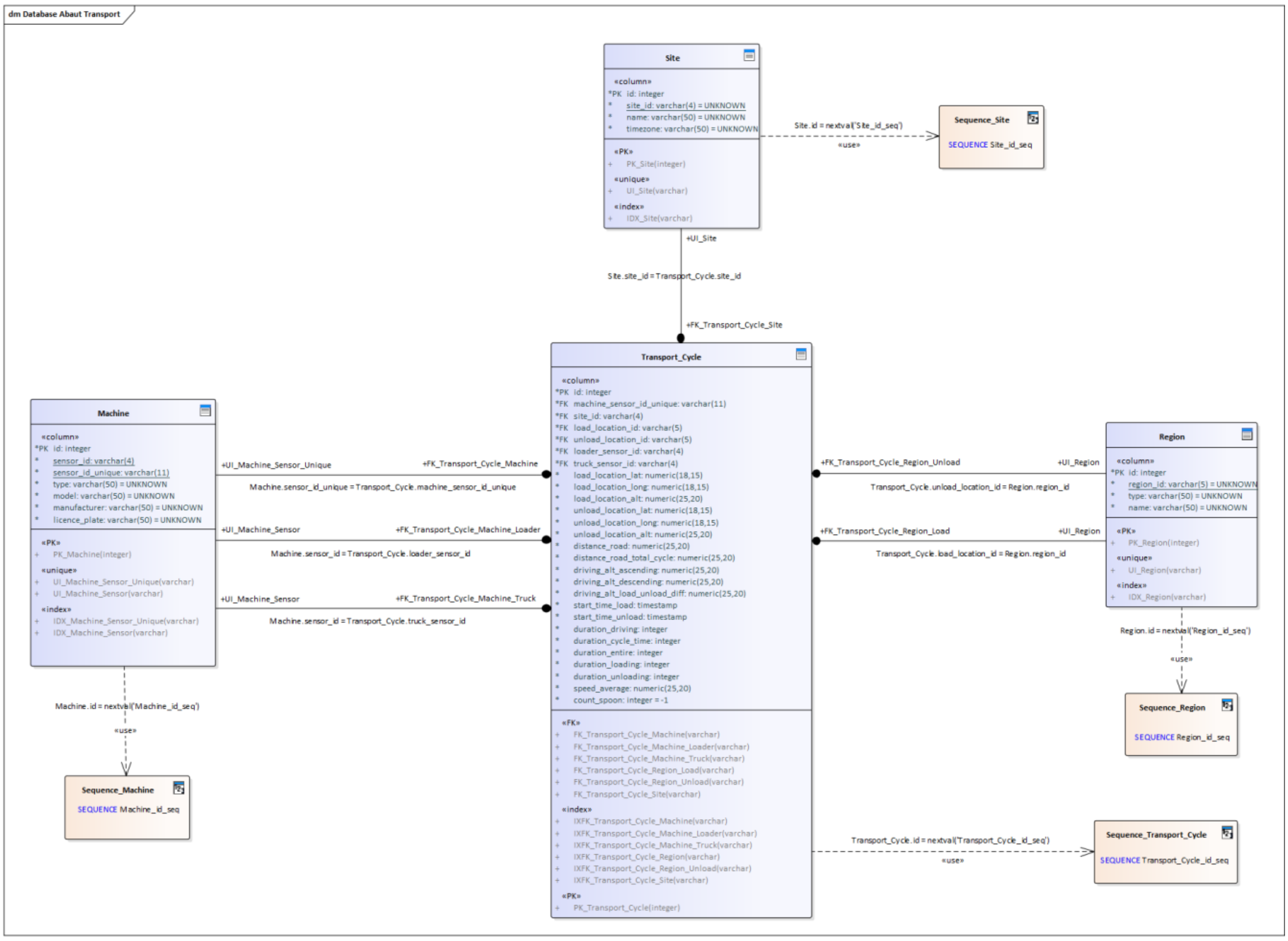
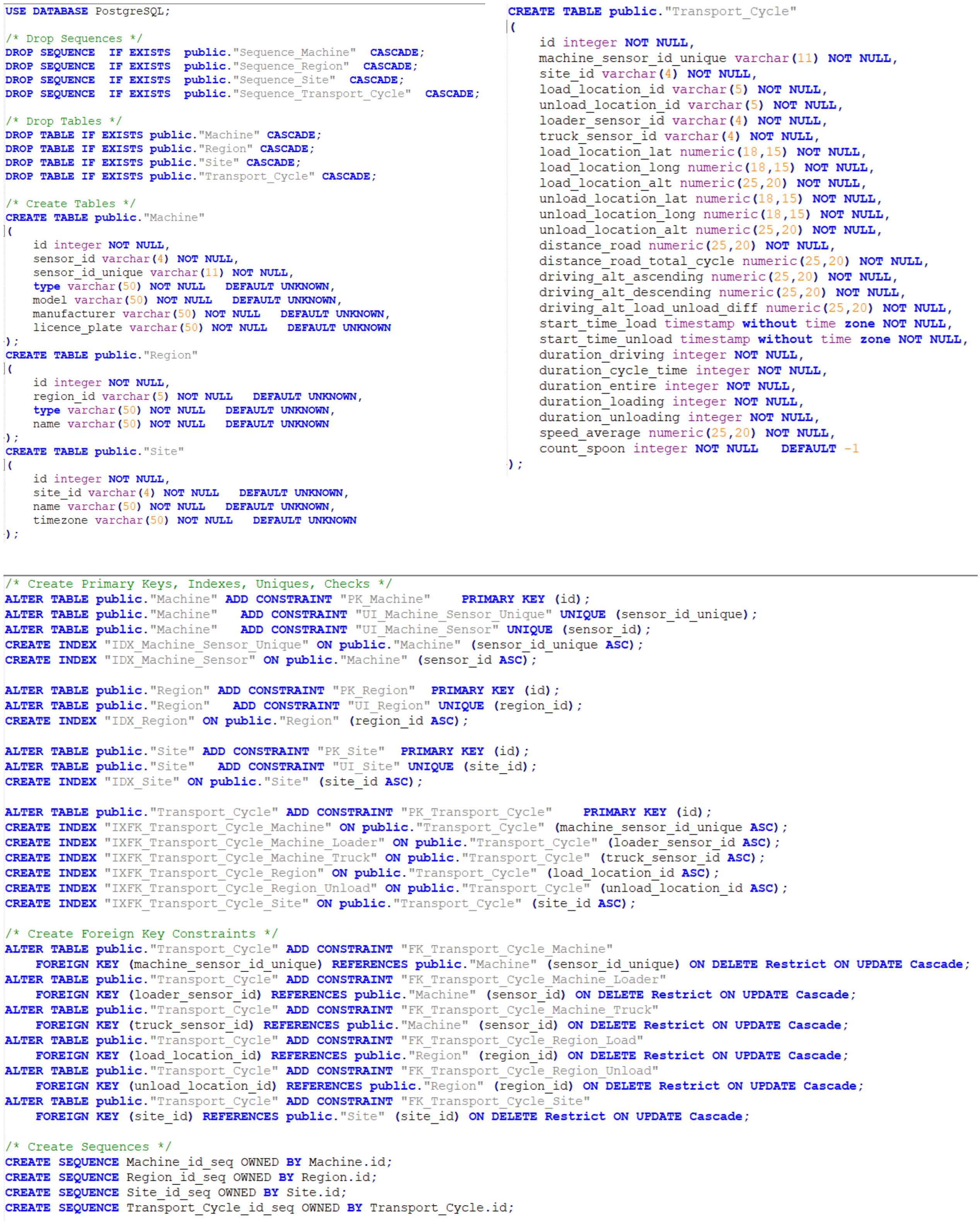

dm Database Abaut Transport
Site
«column»
*PK id: integer
* site_id: varchar(4) = UNKNOWN
* name: varchar(50) = UNKNOWN
* timezone: varchar(50) = UNKNOWN
«PK»
+ PK_Site(integer)
«unique»
+ UI_Site(varchar)
«index»
+ IDX_Site(varchar)
Site.id = nextval('Site_id_seq')
«use»
Sequence_Site
SEQUENCE Site_id_seq
+UI_Site
Site.site_id = Transport_Cycle.site_id
+FK_Transport_Cycle_Site
Transport_Cycle
«column»
*PK id: integer
*FK machine_sensor_id_unique: varchar(11)
*FK site_id: varchar(4)
*FK load_location_id: varchar(5)
*FK unload_location_id: varchar(5)
*FK loader_sensor_id: varchar(4)
*FK truck_sensor_id: varchar(4)
* load_location_lat: numeric(18,15)
* load_location_long: numeric(18,15)
* load_location_alt: numeric(25,20)
* unload_location_lat: numeric(18,15)
* unload_location_long: numeric(18,15)
* unload_location_alt: numeric(25,20)
* distance_road: numeric(25,20)
* distance_road_total_cycle: numeric(25,20)
* driving_alt_ascending: numeric(25,20)
* driving_alt_descending: numeric(25,20)
* driving_alt_load_unload_diff: numeric(25,20)
* start_time_load: timestamp
* start_time_unload: timestamp
* duration_driving: integer
* duration_cycle_time: integer
* duration_entire: integer
* duration_loading: integer
* duration_unloading: integer
* speed_average: numeric(25,20)
* count_spoon: integer = -1
«FK»
+ FK_Transport_Cycle_Machine(varchar)
+ FK_Transport_Cycle_Machine_Loader(varchar)
+ FK_Transport_Cycle_Machine_Truck(varchar)
+ FK_Transport_Cycle_Region_Load(varchar)
+ FK_Transport_Cycle_Region_Unload(varchar)
+ FK_Transport_Cycle_Site(varchar)
«index»
+ IXFK_Transport_Cycle_Machine(varchar)
+ IXFK_Transport_Cycle_Machine_Loader(varchar)
+ IXFK_Transport_Cycle_Machine_Truck(varchar)
+ IXFK_Transport_Cycle_Region(varchar)
+ IXFK_Transport_Cycle_Region_Unload(varchar)
+ IXFK_Transport_Cycle_Site(varchar)
«PK»
+ PK_Transport_Cycle(integer)
Machine
«column»
*PK id: integer
* sensor_id: varchar(4)
* sensor_id_unique: varchar(11)
* type: varchar(50) = UNKNOWN
* model: varchar(50) = UNKNOWN
* manufacturer: varchar(50) = UNKNOWN
* licence_plate: varchar(50) = UNKNOWN
«PK»
+ PK_Machine(integer)
«unique»
+ UI_Machine_Sensor_Unique(varchar)
+ UI_Machine_Sensor(varchar)
«index»
+ IDX_Machine_Sensor_Unique(varchar)
+ IDX_Machine_Sensor(varchar)
+UI_Machine_Sensor_Unique
+FK_Transport_Cycle_Machine
Machine.sensor_id_unique = Transport_Cycle.machine_sensor_id_unique
+UI_Machine_Sensor
+FK_Transport_Cycle_Machine_Loader
Machine.sensor_id = Transport_Cycle.loader_sensor_id
+UI_Machine_Sensor
+FK_Transport_Cycle_Machine_Truck
Machine.sensor_id = Transport_Cycle.truck_sensor_id
Machine.id = nextval('Machine_id_seq')
«use»
Sequence_Machine
SEQUENCE Machine_id_seq
Region
«column»
*PK id: integer
* region_id: varchar(5) = UNKNOWN
* type: varchar(50) = UNKNOWN
* name: varchar(50) = UNKNOWN
«PK»
+ PK_Region(integer)
«unique»
+ UI_Region(varchar)
«index»
+ IDX_Region(varchar)
+FK_Transport_Cycle_Region_Unload
+UI_Region
Transport_Cycle.unload_location_id = Region.region_id
+FK_Transport_Cycle_Region_Load
+UI_Region
Transport_Cycle.load_location_id = Region.region_id
Region.id = nextval('Region_id_seq')
«use»
Sequence_Region
SEQUENCE Region_id_seq
Transport_Cycle.id = nextval('Transport_Cycle_id_seq')
«use»
Sequence_Transport_Cycle
SEQUENCE Transport_Cycle_id_seq

Figure 26: "Transport" Data Model used in the Data Lake to be compliant with ABAUT Data Structures

The following figure shows the database creation scripts:



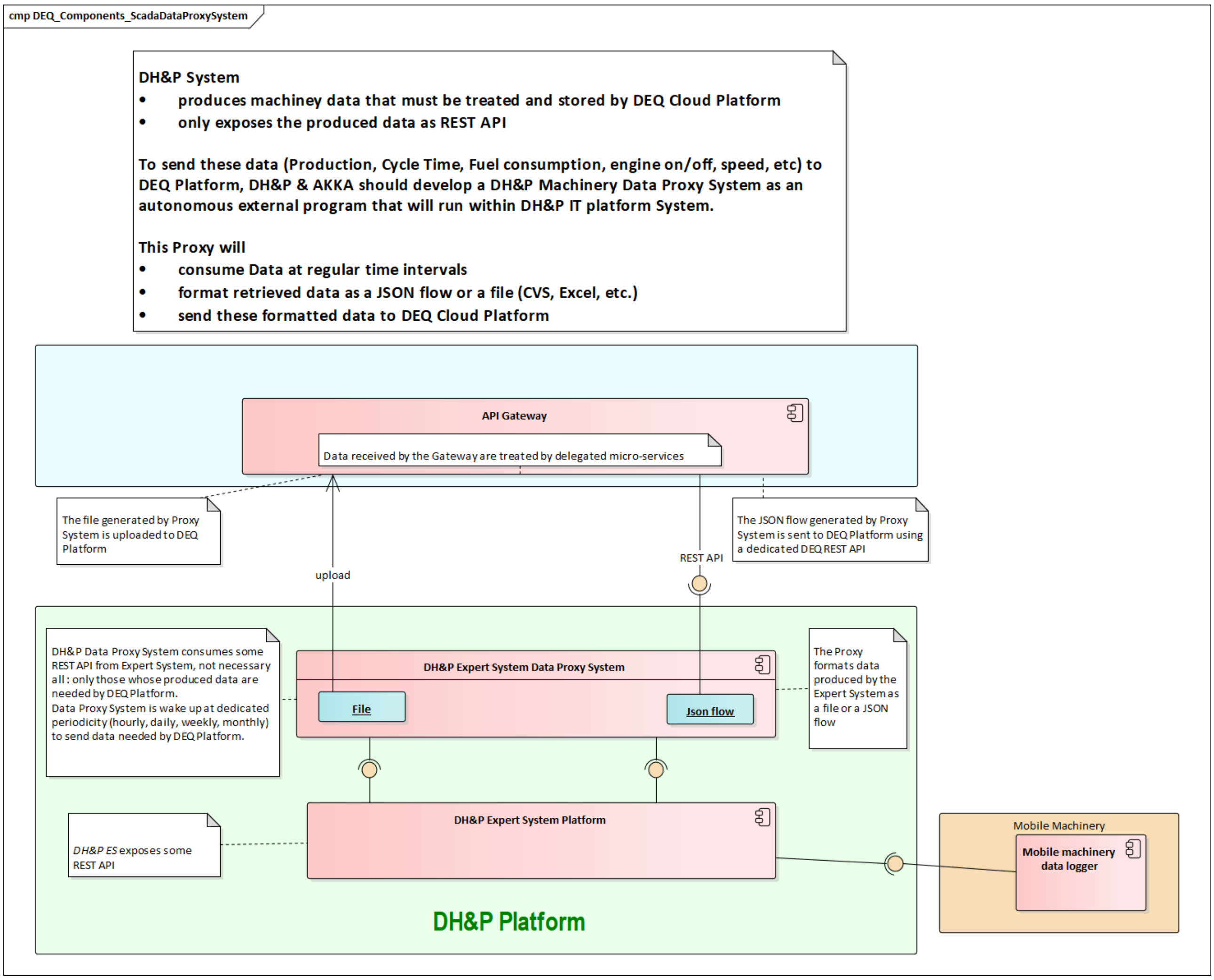
```
USE DATABASE PostgreSQL;

/* Drop Sequences */
DROP SEQUENCE  IF EXISTS  public."Sequence_Machine"  CASCADE;
DROP SEQUENCE  IF EXISTS  public."Sequence_Region"  CASCADE;
DROP SEQUENCE  IF EXISTS  public."Sequence_Site"  CASCADE;
DROP SEQUENCE  IF EXISTS  public."Sequence_Transport_Cycle"  CASCADE;

/* Drop Tables */
DROP TABLE IF EXISTS public."Machine" CASCADE;
DROP TABLE IF EXISTS public."Region" CASCADE;
DROP TABLE IF EXISTS public."Site" CASCADE;
DROP TABLE IF EXISTS public."Transport_Cycle" CASCADE;

/* Create Tables */
CREATE TABLE public."Machine"
(
    id integer NOT NULL,
    sensor_id varchar(4) NOT NULL,
    sensor_id_unique varchar(11) NOT NULL,
    type varchar(50) NOT NULL   DEFAULT UNKNOWN,
    model varchar(50) NOT NULL   DEFAULT UNKNOWN,
    manufacturer varchar(50) NOT NULL   DEFAULT UNKNOWN,
    licence_plate varchar(50) NOT NULL   DEFAULT UNKNOWN
);
CREATE TABLE public."Region"
(
    id integer NOT NULL,
    region_id varchar(5) NOT NULL   DEFAULT UNKNOWN,
    type varchar(50) NOT NULL   DEFAULT UNKNOWN,
    name varchar(50) NOT NULL   DEFAULT UNKNOWN
);
CREATE TABLE public."Site"
(
    id integer NOT NULL,
    site_id varchar(4) NOT NULL   DEFAULT UNKNOWN,
    name varchar(50) NOT NULL   DEFAULT UNKNOWN,
    timezone varchar(50) NOT NULL   DEFAULT UNKNOWN
);
```

```
CREATE TABLE public."Transport_Cycle"
(
    id integer NOT NULL,
    machine_sensor_id_unique varchar(11) NOT NULL,
    site_id varchar(4) NOT NULL,
    load_location_id varchar(5) NOT NULL,
    unload_location_id varchar(5) NOT NULL,
    loader_sensor_id varchar(4) NOT NULL,
    truck_sensor_id varchar(4) NOT NULL,
    load_location_lat numeric(18,15) NOT NULL,
    load_location_long numeric(18,15) NOT NULL,
    load_location_alt numeric(25,20) NOT NULL,
    unload_location_lat numeric(18,15) NOT NULL,
    unload_location_long numeric(18,15) NOT NULL,
    unload_location_alt numeric(25,20) NOT NULL,
    distance_road numeric(25,20) NOT NULL,
    distance_road_total_cycle numeric(25,20) NOT NULL,
    driving_alt_ascending numeric(25,20) NOT NULL,
    driving_alt_descending numeric(25,20) NOT NULL,
    driving_alt_load_unload_diff numeric(25,20) NOT NULL,
    start_time_load timestamp without time zone NOT NULL,
    start_time_unload timestamp without time zone NOT NULL,
    duration_driving integer NOT NULL,
    duration_cycle_time integer NOT NULL,
    duration_entire integer NOT NULL,
    duration_loading integer NOT NULL,
    duration_unloading integer NOT NULL,
    speed_average numeric(25,20) NOT NULL,
    count_spoon integer NOT NULL   DEFAULT -1
);
```

```
/* Create Primary Keys, Indexes, Uniques, Checks */
ALTER TABLE public."Machine" ADD CONSTRAINT "PK_Machine"   PRIMARY KEY (id);
ALTER TABLE public."Machine"   ADD CONSTRAINT "UI_Machine_Sensor_Unique" UNIQUE (sensor_id_unique);
ALTER TABLE public."Machine"   ADD CONSTRAINT "UI_Machine_Sensor" UNIQUE (sensor_id);
CREATE INDEX "IDX_Machine_Sensor_Unique" ON public."Machine" (sensor_id_unique ASC);
CREATE INDEX "IDX_Machine_Sensor" ON public."Machine" (sensor_id ASC);

ALTER TABLE public."Region" ADD CONSTRAINT "PK_Region"  PRIMARY KEY (id);
ALTER TABLE public."Region"   ADD CONSTRAINT "UI_Region" UNIQUE (region_id);
CREATE INDEX "IDX_Region" ON public."Region" (region_id ASC);

ALTER TABLE public."Site" ADD CONSTRAINT "PK_Site"  PRIMARY KEY (id);
ALTER TABLE public."Site"   ADD CONSTRAINT "UI_Site" UNIQUE (site_id);
CREATE INDEX "IDX_Site" ON public."Site" (site_id ASC);

ALTER TABLE public."Transport_Cycle" ADD CONSTRAINT "PK_Transport_Cycle"   PRIMARY KEY (id);
CREATE INDEX "IXFK_Transport_Cycle_Machine" ON public."Transport_Cycle" (machine_sensor_id_unique ASC);
CREATE INDEX "IXFK_Transport_Cycle_Machine_Loader" ON public."Transport_Cycle" (loader_sensor_id ASC);
CREATE INDEX "IXFK_Transport_Cycle_Machine_Truck" ON public."Transport_Cycle" (truck_sensor_id ASC);
CREATE INDEX "IXFK_Transport_Cycle_Region" ON public."Transport_Cycle" (load_location_id ASC);
CREATE INDEX "IXFK_Transport_Cycle_Region_Unload" ON public."Transport_Cycle" (unload_location_id ASC);
CREATE INDEX "IXFK_Transport_Cycle_Site" ON public."Transport_Cycle" (site_id ASC);

/* Create Foreign Key Constraints */
ALTER TABLE public."Transport_Cycle" ADD CONSTRAINT "FK_Transport_Cycle_Machine"
    FOREIGN KEY (machine_sensor_id_unique) REFERENCES public."Machine" (sensor_id_unique) ON DELETE Restrict ON UPDATE Cascade;
ALTER TABLE public."Transport_Cycle" ADD CONSTRAINT "FK_Transport_Cycle_Machine_Loader"
    FOREIGN KEY (loader_sensor_id) REFERENCES public."Machine" (sensor_id) ON DELETE Restrict ON UPDATE Cascade;
ALTER TABLE public."Transport_Cycle" ADD CONSTRAINT "FK_Transport_Cycle_Machine_Truck"
    FOREIGN KEY (truck_sensor_id) REFERENCES public."Machine" (sensor_id) ON DELETE Restrict ON UPDATE Cascade;
ALTER TABLE public."Transport_Cycle" ADD CONSTRAINT "FK_Transport_Cycle_Region_Load"
    FOREIGN KEY (load_location_id) REFERENCES public."Region" (region_id) ON DELETE Restrict ON UPDATE Cascade;
ALTER TABLE public."Transport_Cycle" ADD CONSTRAINT "FK_Transport_Cycle_Region_Unload"
    FOREIGN KEY (unload_location_id) REFERENCES public."Region" (region_id) ON DELETE Restrict ON UPDATE Cascade;
ALTER TABLE public."Transport_Cycle" ADD CONSTRAINT "FK_Transport_Cycle_Site"
    FOREIGN KEY (site_id) REFERENCES public."Site" (site_id) ON DELETE Restrict ON UPDATE Cascade;

/* Create Sequences */
CREATE SEQUENCE Machine_id_seq OWNED BY Machine.id;
CREATE SEQUENCE Region_id_seq OWNED BY Region.id;
CREATE SEQUENCE Site_id_seq OWNED BY Site.id;
CREATE SEQUENCE Transport_Cycle_id_seq OWNED BY Transport_Cycle.id;
```

3.2.2.8 Interface with SmartQuarry developed by DH&P [KTA 3.3]

The following picture depicts the mechanism that will be used to connect the IQS with the DH&P SmartQuarry system used to control the fleet performance. A data pull process is defined to extract data at regular basis from DH&P expert system and upload the data to the data lake. After the data transfer the data will be used by the business management tools.



DH&P System
- produces machiney data that must be treated and stored by DEQ Cloud Platform
- only exposes the produced data as REST API

To send these data (Production, Cycle Time, Fuel consumption, engine on/off, speed, etc) to DEQ Platform, DH&P & AKKA should develop a DH&P Machinery Data Proxy System as an autonomous external program that will run within DH&P IT platform System.

This Proxy will
- consume Data at regular time intervals
- format retrieved data as a JSON flow or a file (CVS, Excel, etc.)
- send these formatted data to DEQ Cloud Platform

Data received by the Gateway are treated by delegated micro-services

The file generated by Proxy System is uploaded to DEQ Platform

The JSON flow generated by Proxy System is sent to DEQ Platform using a dedicated DEQ REST API

DH&P Data Proxy System consumes some REST API from Expert System, not necessary all : only those whose produced data are needed by DEQ Platform.
Data Proxy System is wake up at dedicated periodicity (hourly, daily, weekly, monthly) to send data needed by DEQ Platform.

The Proxy formats data produced by the Expert System as a file or a JSON flow

DH&P ES exposes some REST API

cmp DEQ_Components_ScadaDataProxySystem
API Gateway
upload
REST API
DH&P Expert System Data Proxy System
File
Json flow
DH&P Expert System Platform
Mobile Machinery
Mobile machinery data logger
DH&P Platform

Figure 27: Mechanism used to connect the IQS with the DH&P SmartQuarry system

3.2.2.9 Interface BIM [KTA 4.1]

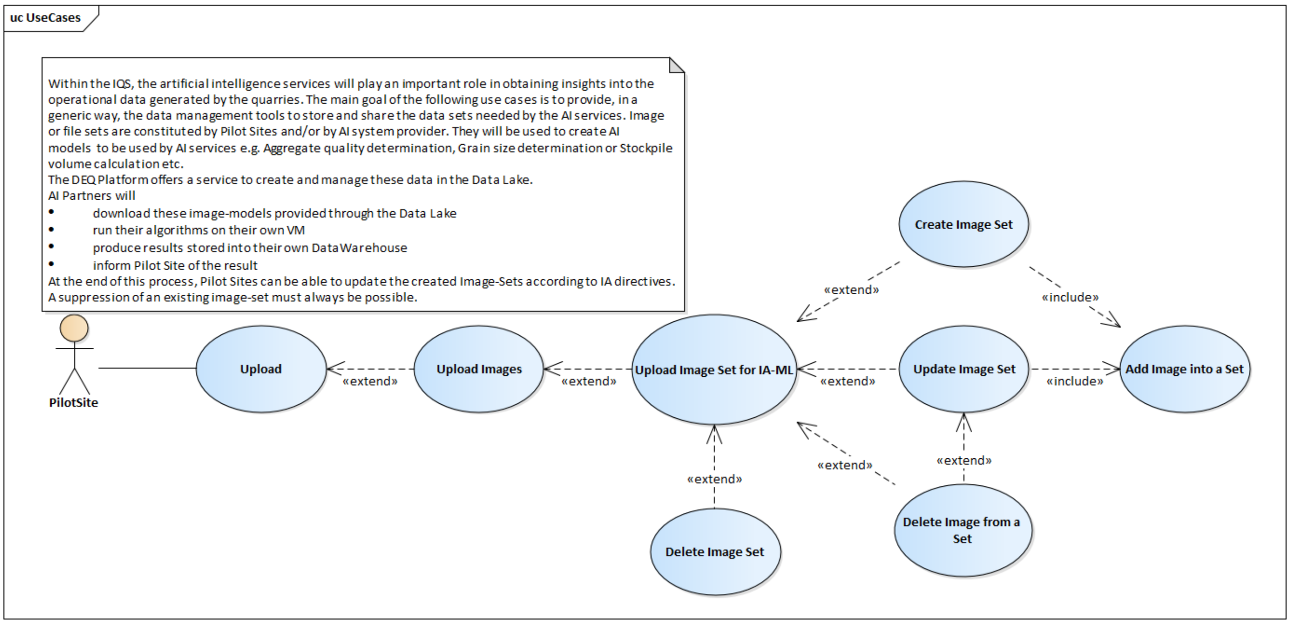
The integration of BIM system and the IoT architecture defined for the IQS will rely on an extensive usage of the interface provided by the chosen components of the benchmark: IoT devices, IoT Hub, Event Hub and Event Grid (see section 3.3 IoT platform)

Hence, the access to IoT data can be performed with pull mechanism for IoT Hub and for Event Hub or with a pub/sub protocol for Event Grid. The interface will be finalized during the development phases in task 4.2 and 4.4.

3.2.2.10 Interfaces with AI services [KTA 4.2]

The integration with Sigma and AI in particular with the IQS will be based on an extensive usage of a Centralized Data Management Platform.

The CDMP will enable the implementation of the uses cases depicted in the following picture:



Within the IQS, the artificial intelligence services will play an important role in obtaining insights into the operational data generated by the quarries. The main goal of the following use cases is to provide, in a generic way, the data management tools to store and share the data sets needed by the AI services. Image or file sets are constituted by Pilot Sites and/or by AI system provider. They will be used to create AI models to be used by AI services e.g. Aggregate quality determination, Grain size determination or Stockpile volume calculation etc.
The DEQ Platform offers a service to create and manage these data in the Data Lake.
AI Partners will

- download these image-models provided through the Data Lake
- run their algorithms on their own VM
- produce results stored into their own DataWarehouse
- inform Pilot Site of the result

At the end of this process, Pilot Sites can be able to update the created Image-Sets according to IA directives.
A suppression of an existing image-set must always be possible.

The following component view shows the interfaces that will be made available to provide access to any data needed by the AI components:

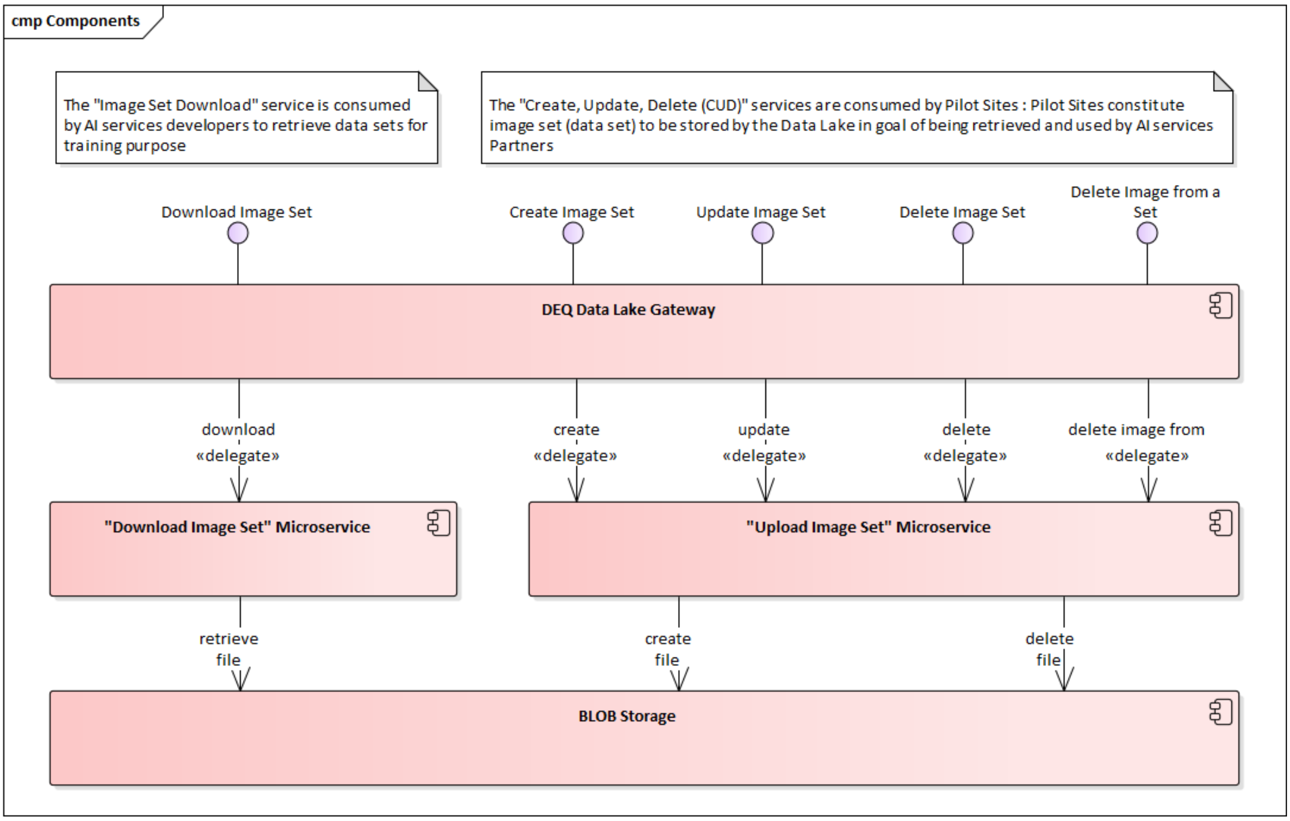

cmp Components
The "Image Set Download" service is consumed by AI services developers to retrieve data sets for training purpose
The "Create, Update, Delete (CUD)" services are consumed by Pilot Sites : Pilot Sites constitute image set (data set) to be stored by the Data Lake in goal of being retrieved and used by AI services Partners
Download Image Set
Create Image Set
Update Image Set
Delete Image Set
Delete Image from a Set
DEQ Data Lake Gateway
download «delegate»
create «delegate»
update «delegate»
delete «delegate»
delete image from «delegate»
"Download Image Set" Microservice
"Upload Image Set" Microservice
retrieve file
create file
delete file
BLOB Storage

Figure 28: Schema of the interfaces available to provide access to data needed by the AI components

3.3 IoT platform

3.3.1 Results of the Benchmark for the best IoT platform tools

The full benchmark's study results done by AKKA are available in Appendix 7.1. Here below is a synthesis of the main results related to the IoT platform but also to Business Intelligence components.

Here is a global view of the components that will be used to build IQS IoT platform and Business Intelligence solution:

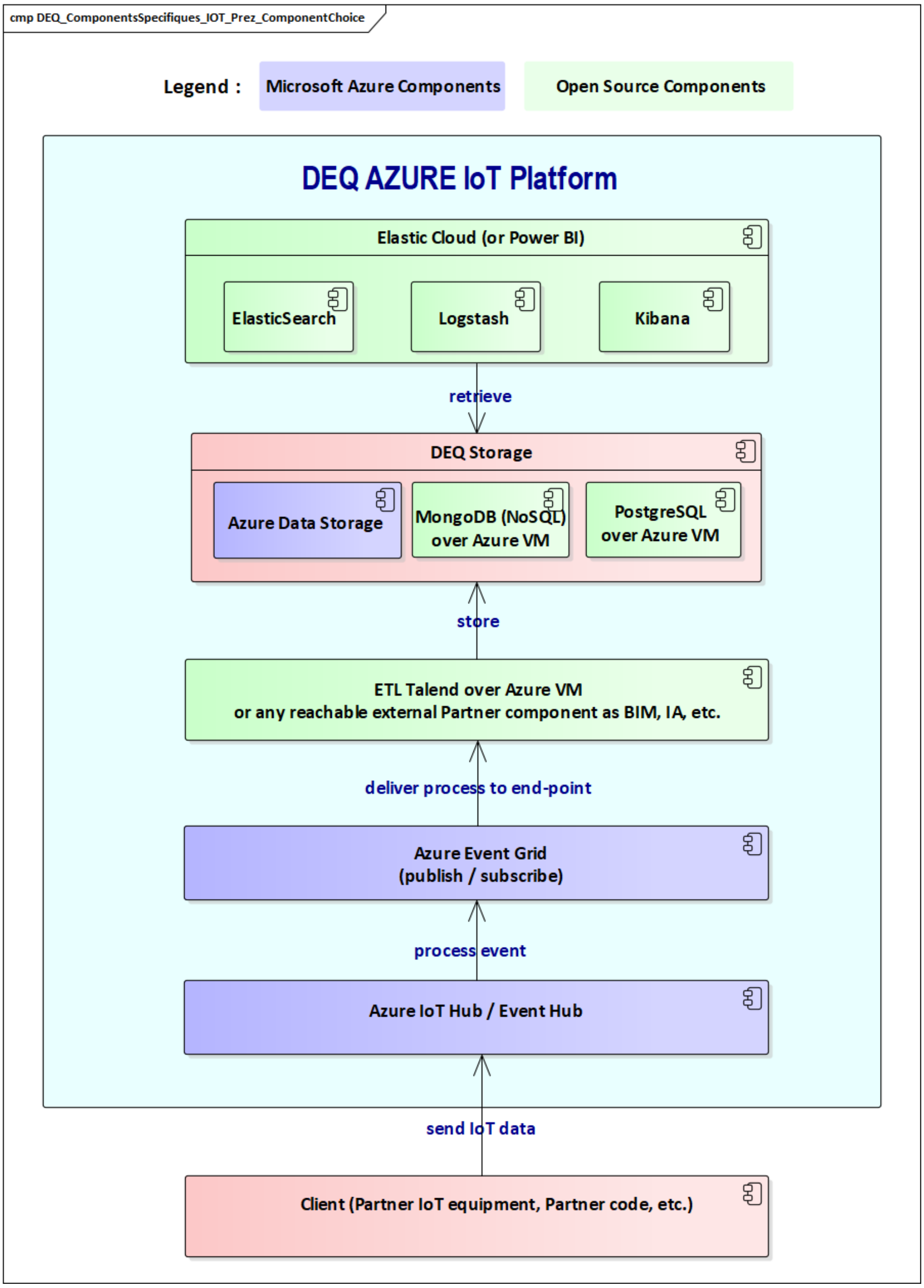

cmp DEQ_ComponentsSpecifiques_IOT_Prez_ComponentChoice
Legend :
Microsoft Azure Components
Open Source Components
DEQ AZURE IoT Platform
Elastic Cloud (or Power BI)
ElasticSearch
Logstash
Kibana
retrieve
DEQ Storage
Azure Data Storage
MongoDB (NoSQL) over Azure VM
PostgreSQL over Azure VM
store
ETL Talend over Azure VM or any reachable external Partner component as BIM, IA, etc.
deliver process to end-point
Azure Event Grid (publish / subscribe)
process event
Azure IoT Hub / Event Hub
send IoT data
Client (Partner IoT equipment, Partner code, etc.)

Figure 29: Diagram of the components selected for the IoT platform and Business Intelligence architecture

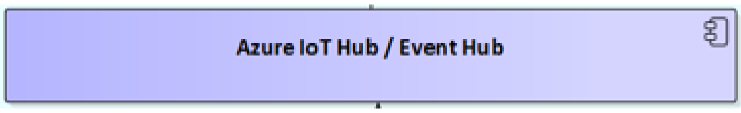
The global cost is estimated to less than 250€ per month per each pilot site:

IoT Frontal	Event Grid	Business Intelligence	TOTAL
IoT Hub: 55€	40€	Elastic Cloud: 125€	**< 250 € / month**
Event Hub: 15€		Power BI: 85€ (5 licenses) 170€ (10 licenses)	

Below, some details are given for each component:

- Description
- Metrics
- Costs

Components: IoT Hub and Event Hub



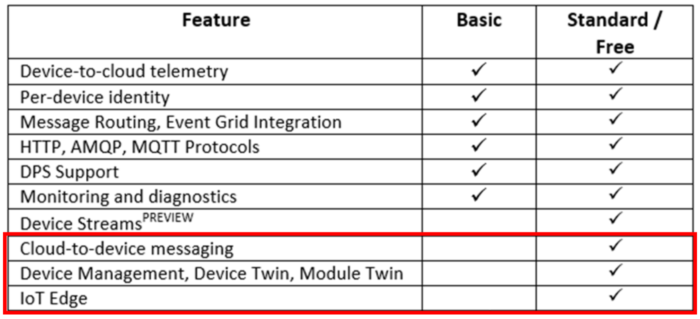
Description:

The two services are similar in that they both support data ingestion with low latency and high reliability, but they are designed for different purposes.

IoT Hub has been designed for connecting IoT devices to the Azure Cloud.

Event Hubs service is more used for streaming Big Data (mainly for hot computing).

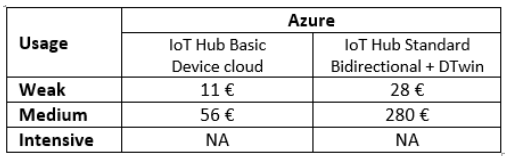
According to Pilot Sites' needs, one or the other should be used.



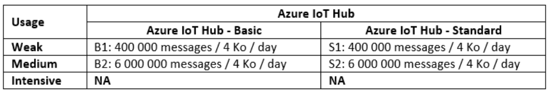
Feature	Basic	Standard / Free
Device-to-cloud telemetry	✓	✓
Per-device identity	✓	✓
Message Routing, Event Grid Integration	✓	✓
HTTP, AMQP, MQTT Protocols	✓	✓
DPS Support	✓	✓
Monitoring and diagnostics	✓	✓
Device Streams^PREVIEW		✓
Cloud-to-device messaging		✓
Device Management, Device Twin, Module Twin		✓
IoT Edge		✓

IoT Hub

Metrics:

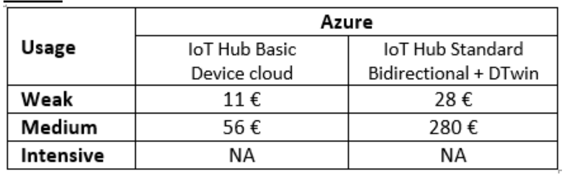


Usage	Azure	
	IoT Hub Basic Device cloud	IoT Hub Standard Bidirectional + DTwin
Weak	11 €	28 €
Medium	56 €	280 €
Intensive	NA	NA



Usage	Azure IoT Hub	
	Azure IoT Hub - Basic	**Azure IoT Hub - Standard**
Weak	B1: 400 000 messages / 4 Ko / day	S1: 400 000 messages / 4 Ko / day
Medium	B2: 6 000 000 messages / 4 Ko / day	S2: 6 000 000 messages / 4 Ko / day
Intensive	**NA**	**NA**

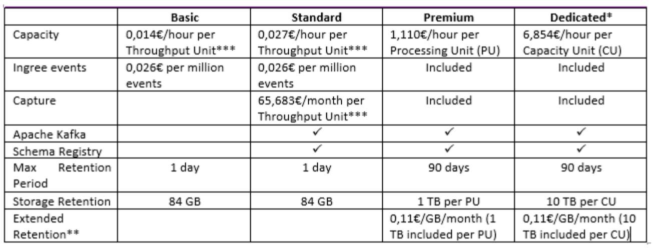
Costs:



Usage	Azure	
	IoT Hub Basic Device cloud	IoT Hub Standard Bidirectional + DTwin
Weak	11 €	28 €
Medium	56 €	280 €
Intensive	NA	NA

Event Hub

Metrics:

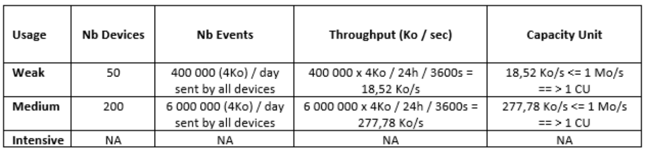


	Basic	Standard	Premium	Dedicated*
Capacity	0,014€/hour per Throughput Unit***	0,027€/hour per Throughput Unit***	1,110€/hour per Processing Unit (PU)	6,854€/hour per Capacity Unit (CU)
Ingree events	0,026€ per million events	0,026€ per million events	Included	Included
Capture		65,683€/month per Throughput Unit***	Included	Included
Apache Kafka		✓	✓	✓
Schema Registry		✓	✓	✓
Max Retention Period	1 day	1 day	90 days	90 days
Storage Retention	84 GB	84 GB	1 TB per PU	10 TB per CU
Extended Retention**			0,11€/GB/month (1 TB included per PU)	0,11€/GB/month (10 TB included per CU)

*Dedicated: Usage will be charged in one-hour increments with a minimum charge for four hours of usage

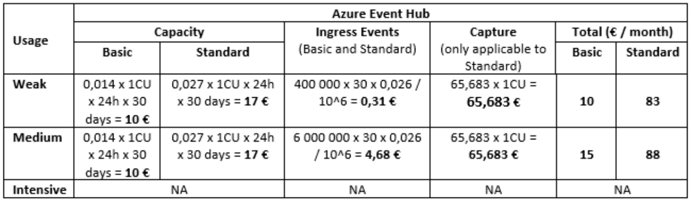
** Message retention above the included storage quotas will result in overage charges.

*** Throughput Unit provides 1 MB/s ingress and 2 MB/s egress.



Usage	Nb Devices	Nb Events	Throughput (Ko / sec)	Capacity Unit
Weak	50	400 000 (4Ko) / day sent by all devices	400 000 x 4Ko / 24h / 3600s = 18,52 Ko/s	18,52 Ko/s <= 1 Mo/s == > 1 CU
Medium	200	6 000 000 (4Ko) / day sent by all devices	6 000 000 x 4Ko / 24h / 3600s = 277,78 Ko/s	277,78 Ko/s <= 1 Mo/s == > 1 CU
Intensive	NA	NA	NA	NA

Costs:



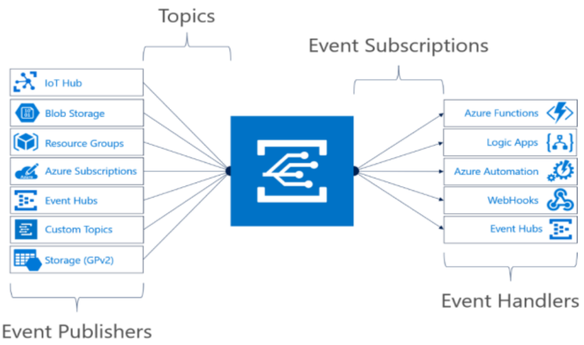
Usage	Azure Event Hub					
	Capacity		Ingress Events (Basic and Standard)	Capture (only applicable to Standard)	Total (€ / month)	
	Basic	Standard			Basic	Standard
Weak	0,014 x 1CU x 24h x 30 days = **10 €**	0,027 x 1CU x 24h x 30 days = **17 €**	400 000 x 30 x 0,026 / 10^6 = **0,31 €**	65,683 x 1CU = **65,683 €**	**10**	**83**
Medium	0,014 x 1CU x 24h x 30 days = **10 €**	0,027 x 1CU x 24h x 30 days = **17 €**	6 000 000 x 30 x 0,026 / 10^6 = **4,68 €**	65,683 x 1CU = **65,683 €**	**15**	**88**
Intensive	NA		NA	NA	NA	

Component: Event Grid

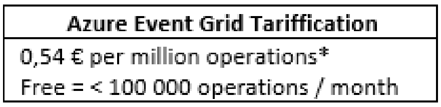
Description: Event Grid is the pub/sub solution for Azure Cloud. It is priced as pay-per-use based on operations performed.

Operations include:

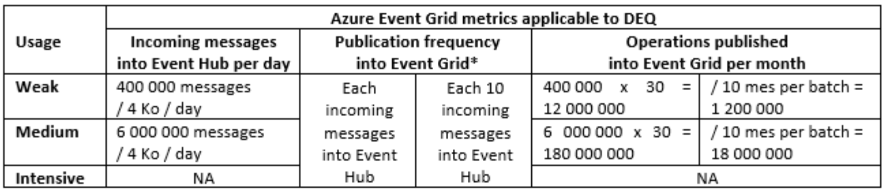
- ingress of events to Domains or Topics,
- advanced matches (using filtering to route to end-points),
- delivery attempt,
- management calls.



Metrics:

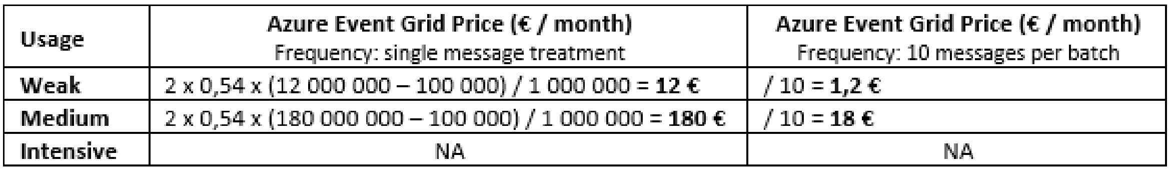


Azure Event Grid Tariffication
0,54 € per million operations* Free = < 100 000 operations / month



Usage	Azure Event Grid metrics applicable to DEQ				
	Incoming messages into Event Hub per day	Publication frequency into Event Grid*		Operations published into Event Grid per month	
Weak	400 000 messages / 4 Ko / day	Each incoming messages into Event Hub	Each 10 incoming messages into Event Hub	400 000 x 30 = 12 000 000	/ 10 mes per batch = 1 200 000
Medium	6 000 000 messages / 4 Ko / day			6 000 000 x 30 = 180 000 000	/ 10 mes per batch = 18 000 000
Intensive	NA			NA	

Costs:



Usage	Azure Event Grid Price (€ / month) Frequency: single message treatment	Azure Event Grid Price (€ / month) Frequency: 10 messages per batch
Weak	2 x 0,54 x (12 000 000 – 100 000) / 1 000 000 = **12 €**	/ 10 = **1,2 €**
Medium	2 x 0,54 x (180 000 000 – 100 000) / 1 000 000 = **180 €**	/ 10 = **18 €**
Intensive	NA	NA

3.4 Data warehouse and AI system architecture

3.4.1 Results of the Benchmark for the best data warehouse tools

This section describes the results obtained from a simple evaluation carried out to provide a data warehouse solution to the project having in mind all the needed use cases and KPI analysis. Our first step is to briefly describe the alternative solutions that have been considered. Then an analysis of the prices/costs for some of those solutions is presented. Finally, an evaluation of several features such as documentation for some of the proposed alternatives.

3.4.1.1 Alternatives being analyzed

The data warehouse solutions were organized into three groups depending on how the deployment of each solution is performed. The different alternatives are listed from the easiest deployment needs to the hardest ones. Every cloud solution that was analyzed includes tools to perform some data analytics tasks. However, these tools are unlikely to fit to the specific needs of the project.

3.4.1.1.1 Specific cloud solutions

A specific solution for the data warehouse requires a little administration effort but this advantage is reflected in a higher price than generic cloud solutions. We analyzed as a prominent commercial data warehouse, the solution provided by *Snowflake*.

- Snowflake

Snowflake for data warehouse is a solution that unifies the storage and analysis interface based on one or more generic cloud providers services. One can deploy its data warehouse, or its data analysis processes across the main cloud providers (*Amazon*, *Google*, or *Microsoft*). The main drawback of this solution is that price is higher than any of the cloud providers the solution is deployed on top of.

For our analysis we are using the information provided in the platform's pricing website and the advice provided by the technical staff from Snowflake we contacted with.

3.4.1.1.2 Generic cloud providers

A generic cloud solution like the analyzed in this benchmark provide an execution infrastructure but all provisioning of the services is up to the user. It provides much more flexibility than a specific solution. However, there include a higher platform administration effort.

The solutions provided by the three most important cloud providers were analyzed: Google, Amazon, and Microsoft. Contacts were stablished with sales staff from *Snowflake* and with the sales technical staff from *Microsoft*. Unfortunately, the later did not provide a lot of useful information.

- Google BigQuerry

Google Cloud provides a service called BigQuery that implements a distributed and scalable data warehouse solution. It is offered as a *serverless* solution, but resources can be *reserved* to reduce the cost when a specific service implementation has a high demand.

For the analysis presented below, the information about pricing and the documentation of the BigQuery service has been used.

- Amazon Redshift

Amazon Web Services offers a service called Redshift that provides a scalable cloud data warehouse solution. It is offered as a *provisioned* service, but AWS also offers a *serverless* option. However, the latter is still in *preview* so it may not be suitable for a production environment.

For the analysis presented below, the information about pricing and the documentation of the *Redshift* service has been used.

- Microsoft Azure Synapse Analytics

Microsoft Azure has a service called Synapse Analytics that provides big data analytics along with a data warehouse storage platform. Both the storage and the analytics services are mainly offered as a *provisioned* infrastructure but is it possible to use them as a *serverless* solution. *Synapse Analytics* seems to be more oriented to a data analysis platform so it seems more likely to offer services that in most of the cases will not be used by the DigiEcoQuarry solutions.

For the analysis presented below, the information about pricing and the documentation of the *Synapse Analytics* service has been used.

3.4.1.1.3 On-premises (Custom) Data Warehouse solutions

Implementing an *on-premises* data warehouse it the most flexible option. However, it comes with the drawback that building an acceptable solution requires a great effort. For example, designing a scalable infrastructure requires taking into account hardware provisioning, fault tolerance, data access security, encryption, etcetera. Considering that a huge amount of data won't be used for the DigiEcoQuarry infrastructure and services, it may not be worthwhile building an *on-premises* solution when considering the price of a generic cloud provider solution. Moreover, the cost of maintaining and operating such a platform at the long run will make this type of solution not viable for the Quarries type of business, where specific IT teams are not available on board. Therefore, an ad-hoc solution based on open-source software solutions is not considered.

3.4.1.2 Pricing

The price is one of the most important factors to consider when choosing a technical solution to implement operations on a daily basis of any given business. It can determine whether to discard or not a given platform solution. The cost of the exploitation of a platform was initially considered. A spreadsheet was designed to provide a fair comparison between the options considered in this analysis. The latter spreadsheet is attached at the Appendix section of this deliverable. <attachment: prices-spreadsheet>.

The cost of the data stored in the platform is more or less calculated the same way for all providers. However, the prices for the *Synapse Analytics* solution could not be found. The storage costs for this platform are nearly the same as the costs for similar solutions (that are, *BigQuery* and *Redshift*).

Calculating computing costs is trickier because there are two points of view. For *Synapse Analytics* and *BigQuery* the cost of data instantaneous computing depends on how much data from the data warehouse is retrieved to complete the computing task. On the other hand, *Redshift* and *Snowflake* calculate the cost of computing considering the time required to complete the computing task.

To calculate the costs of the computing platform in this benchmark, the amount of data used for a query is considered as the main factor, because the cost philosophy of the Microsoft solution was used as the main reference due to the fact that the data lake solution prototype will be most probably based on the latter. This made nearly impossible to estimate the computing costs for *Redshift and for Snowflake* from this point of view.

3.4.1.2.1 Example scenario

To provide an example of how the prices were calculated, the BigQuery solution is taken as the reference since it provides very detailed and clear pricing policies. It is considered that the platform will be ingesting an average of 2.5TB each month and the latter data won't be removed (historic data is important for AI data processing). Each month the dashboards would perform queries to the data warehouse and could retrieve about 2.2TB of data as an estimation. If a period of 4 years is considered, the total cost per year is summarized in the table below. Note that the scenario is set up to check the prices for all the three categories, the computation required for the queries, the amount of storage required for the data and the transfer rates applied to ingest and store the data.

Year	Compute	Storage	Transfer	Total
First	$87.24	$4,444.52	$181.92	$4,713.74
Second	$87.24	$12,654.84	$181.92	$12,924.06
Third	$87.24	$20,865.17	$181.92	$21,134.38
Fourth	$87.24	$29,075.49	$181.92	$29,344.71

Table 1: Prices per year in the example DWH scenario

Note that no data is removed, then at the end of a four-year period, 150TB will be stored!

3.4.1.2.2 Free-tier scenario

Let's consider now a less demanding scenario. One GB of data will be ingested each month and queries are being performed from the dashboards that are retrieving a maximum of 1TB each month. The total cost per year is summarized in the table 2. Note queries are about each inserted data 1000 times! In this scenario the costs included are only the costs of the storage because in the *BigQuery* service the free-tier quotas are not exceeded for computing and transfer,

Year	Compute	Storage	Transfer	Total
First	$0.00	$0.07	$0.00	$0.07
Second	$0.00	$2.33	$0.00	$2.33
Third	$0.00	$5.61	$0.00	$5.61
Fourth	$0.00	$8.89	$0.00	$8.89

Table 2: Prices per year in the free-tier DWH scenario

The prices included in table 1 and table 2 refer to the *BigQuery* service because is the only platform that comprises in its web offering the prices for *compute*, *storage*, and *transfer*. However, the prices of other similar services such as *Redshift* of *Synapse Analytics* seem to be fairly similar.

3.4.1.2.3 Conclusions from pricing

The Snowflake platform is the most expensive from all the analyzed options. Even only considering the costs of its storage, it shows to be more expensive that the total cost of any competitor. We believe that **it is not worth using the *Snowflake* platform** considering the facilities it provides from the point of view of system administration.

The cost of computing in the Redshift platform could not be calculated because the pricing model is different from the one we are using. However, the cost of a platform billed by time is very hard to estimate because it is not known how much time it will require to complete a process/calculation. It is also important to highlight the *serverless* service is still in *preview* so it seems not a good idea to use it in a production environment. For these reasons **the Redshift service is also discarded.**

Just considering the cost of the cloud solutions, the *BigQuery* service from *Google* and the *Synapse Analytics* platform service from *Microsoft,* have similar prices for similar features. We believe that from the price point of view **either *BigQuery* or *Synapse Analytics*** should be chosen to implement the data warehouse.

3.4.1.3 Development

As stated, price is the most important factor to be considered in order to choose a data warehouse solution that can be used in a production environment. There are also some other factors related to development activities to take into account to select which data warehouse solution should be chosen.

Considering the development point of view to take a decision makes sense because of the need to build software using the service. It is desirable to avoid problems that may arise because the chosen platform does not provide good documentation, or the interfaces are hard to use. The benchmark is based on a list of questions designed to evaluate each solution. The final list of questions is summarized in table 3. However, those questions have been selected from a more exhaustive benchmark and some of them have been reformulated to be answered just using a score. The full benchmark is included in the annex <reference: annex>.

The questions related to development that need to be answered for each platform can be grouped into three categories:

1. How good is the documentation provided for development?
2. How good are the interfaces provided by the service?
3. How easy is it to use the service from an implementation point of view?

Documentation	*How easy is it to find documentation about the resources?*
	How clear is the service documentation?
	How accurate is the service documentation?
	How useful are the examples included in the documentation?
	How active is the user's community?
Interface	*How easy is it to use the RESTful API interface?*
	How easy is it to use the Python library interface?
	How easy is it to load data into the storage?
	How easy is it to perform SQL queries?
	How stable are the service interfaces?
Usage	*How easy is it to build a mock-up environment?*
	How easy is it to create a test suite?
	How useful is error reporting?
	How good is the design of the interfaces?
	How easy is it to extend the libraries?

Table 3: Concrete questions about development

Up to this point, the decision concerns only to between *BigQuery* and *Synapse Analytics*. For each platform a score from 0 to 10 is assigned to each question. Table 4 summarizes the score for each question and the final average score for each platform.

	BigQuery	Synapse
How easy is to find documentation about the resources?	9.00	1.50
How clear is the service documentation?	8.00	3.00
How accurate is the service documentation?	7.00	4.00
How useful are the examples included in the documentation?	8.00	2.00
How active is the user's community?	9.60	0.40
How easy is to use the RESTful API interface?	7.00	8.00
How easy is to use the Python library interface?	7.00	0.00
How easy is to load data into the storage?	6.00	3.00
How easy is to perform SQL queries?	6.00	7.00
How stable are the service interfaces?		
How easy is to build a mock-up environment?	0.00	0.00

	BigQuery	Synapse
How easy is to create a test suite?	6.00	4.00
How useful is error reporting?		
How good is the design of the interfaces?	6.50	2.00
How easy is to extend the libraries?	6.00	7.00
TOTAL	**6.62**	**3.22**

Table 4: Development scores for the selected platforms

From the results on table 4, the service that got the highest score is ***Google BigQuery.*** There are some final remarks about both services that is worth sharing. From this evaluation it is also important to consider:

- The activity of the *user's community* from the number of questions answered for each platform on *Stack Overflow was calculated*.
- The stability of interfaces cannot be easily determined if the evolution of the library is not considered. However, that does not affect the final decision.
- It is hard to determine *how useful is the error reporting feature* without facing many errors when using the service. We consider that this is not important enough to bias the final decision.

- **BigQuery Remarks**

It was observed that the *BigQuery* platform has a **good documentation**. It also provides good implementation examples using the Python programming language. However, not good usage examples were found for the RESTful interface.

However, **there is not an easy way to test the library without using the real platform**. There does not exist an emulator for this service so building integration tests requires to configure a service for testing. This kind of tests may slightly increase the price of using the service.

- **Synapse Analytics Remarks**

The same way as the *BigQuery* service **there does not exist an emulator for *Synapse Analytics*** so building integration tests is not as handy as it would be desirable.

Apart from that, the **documentation is not well organized, and it is hard to find and understand.** Some described concepts required to effectively use the platform to be understood.

It was surprising to discover that the *Synapse Analytics* platform **does not offer a Python library**. However, any SQL driver can be used to access the service. Despite of that, a mock-up of a SQL client for our tests was not implemented because tests of the SQL statements used within the SQL client would have to be performed

3.4.2 General data warehouse architecture and interfaces

This section provides an overview of how the chosen data warehouse solution fits into the whole deployment. The main reason to include a data warehouse solution is to store the results of the different data-related operations and AI modules and make them easily available, and to provide analysis capabilities to facilitate querying these results.

As was detailed in deliverable D1.3, the following image shows how the data warehouse fits in the global IQS architecture:

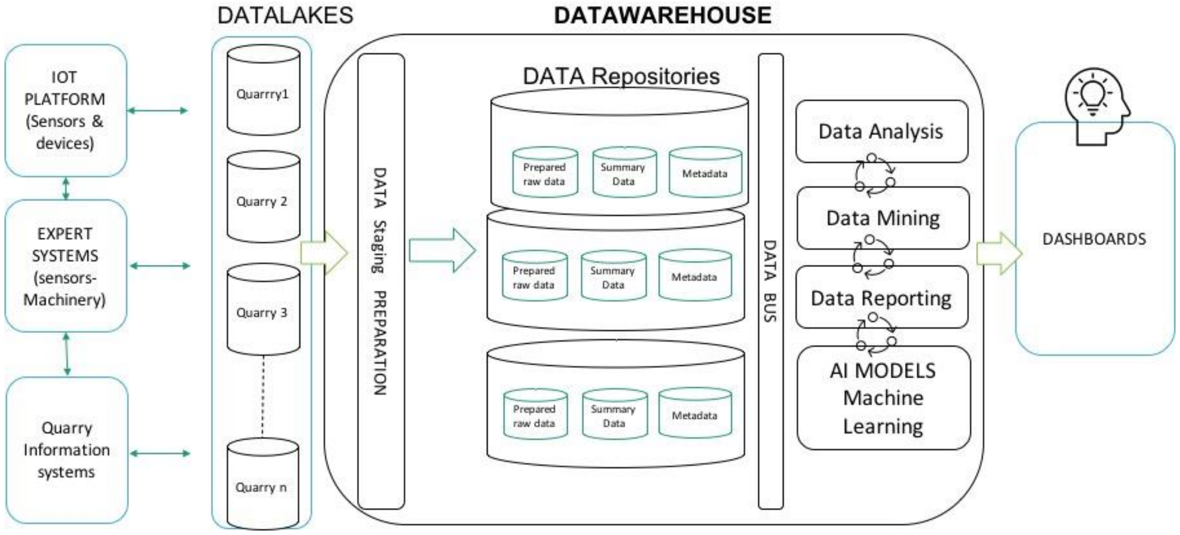


Figure 30: General IQS architecture

Individual storage space is given to each quarry in the data warehouse. This implies that, from the user's point of view, it seems that there is a single data warehouse for each quarry. However, this is not a rule of thumb: an operator that owns several quarries might use the same data warehouse for all its quarries in nearby locations. The most important points here are 1) data from different quarries is not shared and 2) data warehouse storage is located as close as possible to the quarry it belongs to.

Focusing only on the components that are relevant for the AI services, the image below illustrates the role of the data warehouse within the IQS:

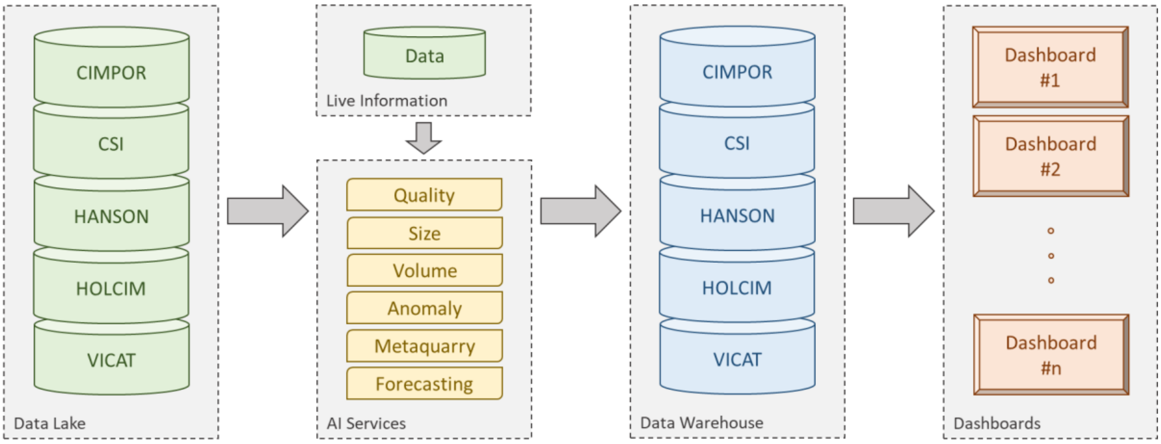


Figure 31: The role played by the data warehouse for the AI services

As shown in the figure, the services get the data required for their execution from the *data lake* and store their results in the *data warehouse.* Note that due to data privacy and security, each quarry has a logically separated storage space that will be physically located in the closest cloud region to each quarry. Services may also require accessing some publicly available data, such as satellite images, to be able to perform some training or inference tasks. The data warehouse provides a SQL interface to query the results provided by each service.

3.4.2.1 Query interface

As mentioned before, data stored in the Data Warehouse will be accessed through a SQL interface. SQL is supported in BigQuery by means of Google Standard SQL, an ANSI compliant Structured Query language that offers the following types of statements:

- **Query statements:** used to scan one or more tables or expressions. They are also known as Data Query Language (DQL) and are the main method to analyze data in BigQuery.
- **Procedural language:** procedural extensions to BigQuery SQL that allow to execute multiple SQL statements in one request.
- **Data definition language:** allows to create and modify database objects such as tables, views, functions, and row-level access policies.
- **Data Manipulation Language (DML):** allows to update, insert, and delete data from your BigQuery tables.
- **Data Control Language (DCL):** allows to control BigQuery system resources such as access and capacity.
- **Transaction Control Language (TCL):** allows to manage transactions for data modifications.
- **Other statements:** provide additional functionality, such as exporting data.

BigQuery offers two possibilities to run SQL queries: interactive and batch queries. Interactive queries, which are the default, are executed as soon as possible. In contrast, batch queries are queued automatically and are run as soon as idle resources are available in the BigQuery resource pool, which typically occurs within a few minutes. If a batch query has not been run within 24 hours, its priority will be changed to interactive.

Only authorized users will be able to submit queries to BigQuery. These permissions will be detailed at the end of this section.

3.4.2.2 Access from dashboard:

The dashboards that allow to visualize the data generated by the different AI services will be implemented in Power BI. Power BI natively supports BigQuery, so importing data from this warehouse is as simple as selecting BigQuery as the data source and logging in with a user that is authorized to access the required resources.

3.4.2.3 Access rights (IAM):

Google's BigQuery provides a system to control user access to the resources stored in the data warehouse, called IAM (Identity and Access Management). This system allows to specify which users (identities) have the appropriate access rights (roles) to check a certain resource. These resources can be SQL databases or other data sources, but also specific views or tables within an SQL database.

Thus, in IAM, permissions are not granted directly to end-users. Instead, permissions are grouped into roles, which can be then granted to authenticated users, or *principals*.

Lastly, an *IAM policy* (which can be either an *allow* or *deny* policy), defines and enforces what roles are granted to which principals. Policies are attached to resources, so when an authenticated principal attempts to access a resource, IAM checks the resource's policy to decide whether the action is allowed.

To sum up, the IAM model has three main parts:

Principal: user that wishes to access a certain resource. It can be an end-user or an application or compute workload. Its identity can be an email address associated with a user, service account or Google group, or a domain name associated with a Google Workspace account or a Cloud Identity domain.

Role: set of permissions that determine the operations that are allowed on a resource. By granting a role to a principal, all permissions contained in that role will be granted.

Policy: collection of role bindings that bind one or more principals to individual roles.

The following figure illustrates the IAM model:

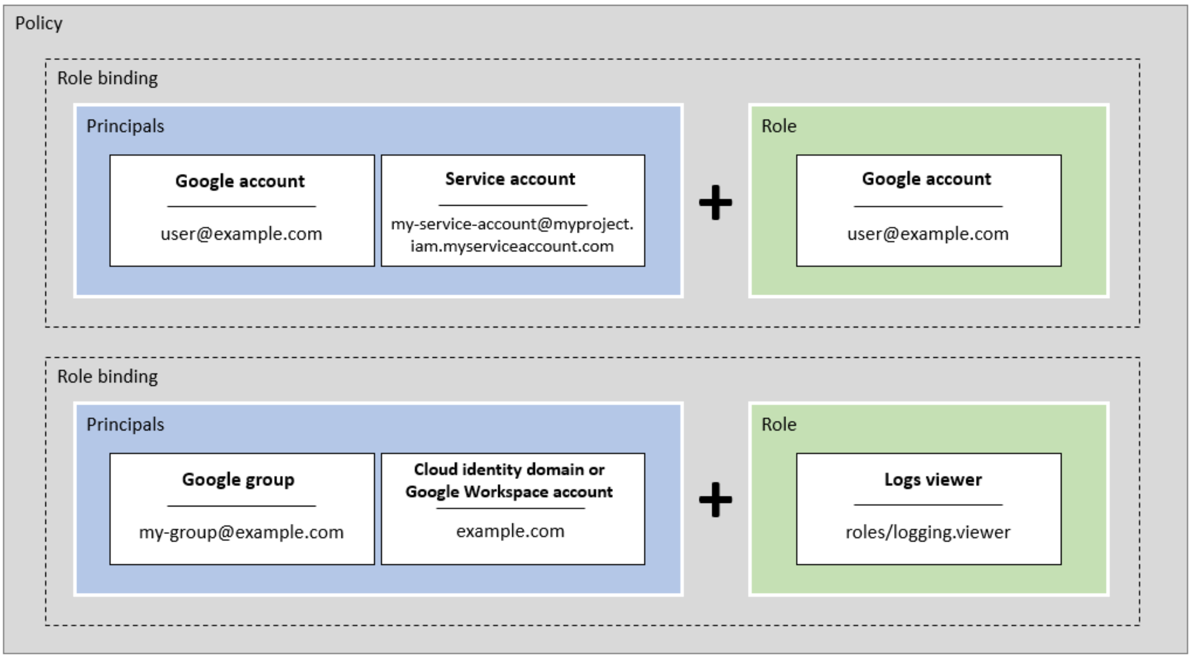


Figure 32: IAM's permission management

3.4.2.4 Permissions to run SQL queries:

As it was introduced in the SQL interface description section, only authorized users will be able to submit SQL queries and to retrieve data from the SQL database.

On the one hand, to run a query the **bigquery.jobs.create** permission is required. This permission is included in the following predefined roles:

- **roles/bigquery.admin**
- **roles/bigquery.jobUser**
- **roles/bigquery.user**

On the other hand, it is necessary for a user to have access to all tables and views that the query reference, which is granted by the **bigquery.tables.getData** permission. The following predefined roles include said permission:

- **roles/bigquery.admin**
- **roles/bigquery.dataOwner**
- **roles/bigquery.dataEditor**
- **roles/bigquery.dataViewer**

3.4.3 General AI system architecture and interfaces

The artificial intelligence services play a role in obtaining insights into the operational data generated by the quarries. The main goal is to extract value from raw data in order to provide quarry operators with information that helps them in driving their business.

A total of six services are planned to be implemented, namely:

Aggregate quality determination

Grain size determination

Stockpile volume calculation

Detection of mechanical failures

NLP information and document search engine (Metaquarry)

Consumptions & product forecasting

These services have been re-defined with respect to the plan presented in the Grant Agreement, after taking into consideration the context, real business necessities and infrastructure of the different pilots. However, they all keep a close link to the service areas that were defined in the proposal and address similar needs. The correspondence between the original and newly proposed services is shown in the following table:

Original Service Areas	New defined services
Metaquarry	*No Change*
Stock Forecast	1.-Stock Pile Volume calculation
	2.-Consumptions and Product Forecasting
Predictive Maintenance	Detection of Mechanical Failures
Hawkeye	1.-Aggregate Quality Determination
	2.-Grain Size Determination

Table 5: Relationship between original and new services

Each of these services will be described in detail below, though there are some common points that are worth mentioning.

In an ideal scenario, operation data comes only from the data lake. However, there are some circumstances that require to access data directly from the quarry, such as sensor values or images from cameras placed on-site. Whenever it is possible, data required by the AI services is retrieved from the data lake.

The following sections describe the architecture of each artificial intelligence solution. The base design principle is to make them as similar to each other as possible to ease the understanding of the service. However, there are some differences (mainly due to real-time data requirements):

For those services that require to process real-time information coming from cameras, microphones, and other kind of sensors, the service runtime environment is divided into a training that is executed on a *cloud infrastructure* and some estimation process that is executed in the *quarry facilities* to process the real-time information.

For the services that process real-time information, some sensors have to be installed in *quarry facilities* to monitor some machinery. For example, in the services that require visual inspection of the materials in conveyor belts, some cameras have to be placed to obtain images that can be processed.

The services that require some processing in the quarry may face some connectivity issues. For example, the Internet provider may face some infrastructure issues that make it impossible to send processed results to the data warehouse. These issues require a *buffering aggregator* that is placed to avoid missing some of the processed results.

The general schema for services is to train a system using some existing data and the, once the system is trained, deploy it on a cloud environment or to the quarry facilities. Table 5: shows the meaning given to the icons used in the architecture diagrams of the services described in the following sections.

Document	Metrics	Microphone
Camera	Statistics	Text
Query	Estimation	Item list
Image	Model	Sensor
Audio signal	Alarm	“ … ” Sentence
Data set	Signal	

Table 6: Icons used on architecture diagrams

3.4.3.1 Aggregate quality determination

This service aims at estimating the quality of aggregates (composition) on the line during production, based on visual data captured by cameras, along with other external data, such as weather information.

The system will be non-intrusive and will allow to maximize the run-of-quarry process by improving quarry planning and controlling the grinding process. It will also support automated notification and will keep a historical record to enable further analysis of the data.

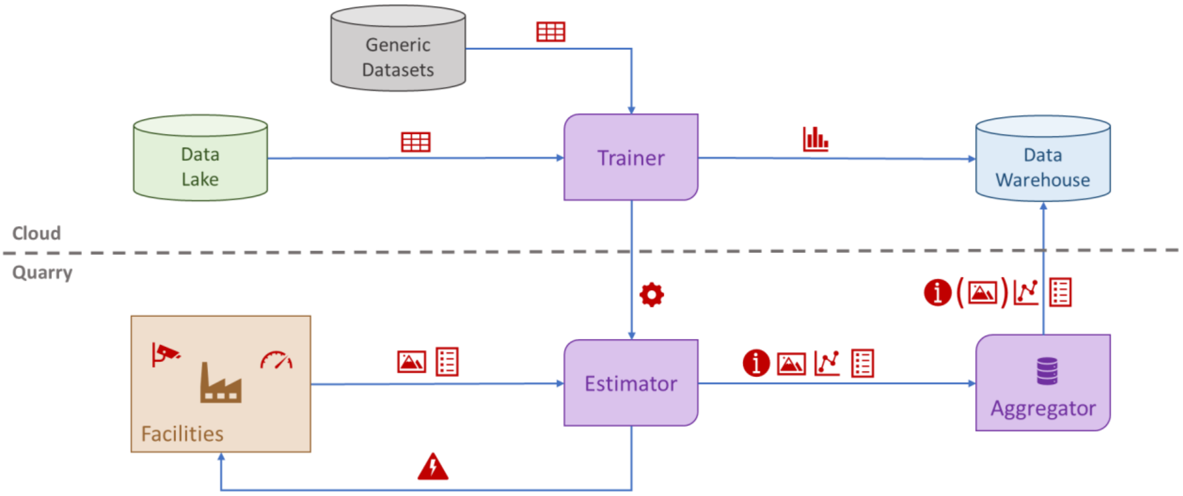


Figure 33: Base architecture of the aggregate quality determination service

Figure 33Figure 33 represents the architecture of the service. It has two distinct parts: the training components, deployed in the cloud and whose objective is to train the AI algorithms of the estimator, and the estimation components, which are deployed on-premises and analyze live data coming from the quarries to provide the quality of the extracted materials. Each component will be described in detail below.

The *Trainer* component is in charge of creating a quality model that can be used by the estimator. It takes as input historical data from the sites that have been stored in the data lake and rock images from generic datasets. Lastly, the trainer stores metrics related to the generation of the model in the data warehouse.

The *Estimator* receives the images of the on-line aggregates that have been captured by cameras located in the quarries, and applies the model generated by the trainer to provide an estimation of the aggregates' quality. It also utilizes sensor data to account for changes in the environment that can affect the captured images, such as ambient lighting.

Finally, the *Aggregator* has a two-fold purpose. On the one hand**,** it implements a buffer to keep the results of the estimator before storing them in the data warehouse, to avoid data loss in case of connectivity issues. On the other hand, it evaluates the relevance of the results to decide if they will be stored or discarded, based on the results themselves and the metrics received from the estimator.

Interfaces:

Cloud:

Input	Source	Output	Destination
Images of quarries	Data lake	Quality model	Estimator
Generic images	Generic data sets	Statistics	Data warehouse

Table 7: Interfaces for the aggregate quality determination service (cloud)

On-premises:

Input	Source	Output	Destination
Environment data	Sensors	Quality estimation	Data warehouse
Real-time quarry images	Cameras	Metrics	Data warehouse

Input	Source	Output	Destination
		Alarms	Data warehouse
		Corrective action	Quarry

Table 8: Interfaces for the aggregate quality determination service (on-premises)

3.4.3.2 Grain size determination

The Grain size determination service has the goal of analyzing visual inputs to estimate the size of the rock fragments extracted from the sites. The system will measure the grain size distribution as it goes through the quarry lines, being able to detect oversize material and evaluate grain uniformity, which, in turn, will allow to reduce damage in the crushing process and to maximize the efficiency of the run-of-quarry process.

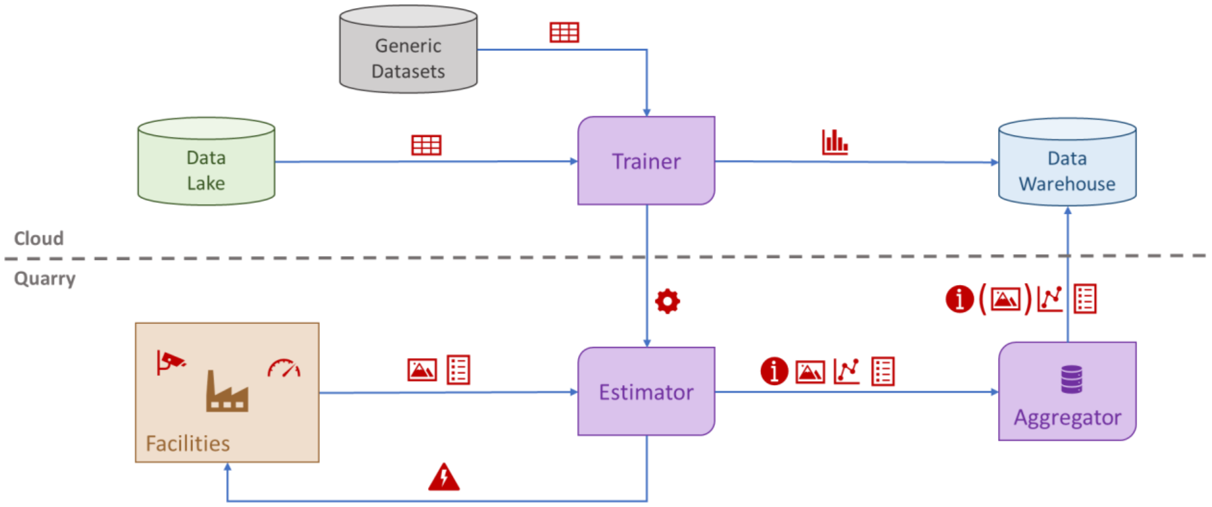


Figure 34: Base architecture of the grain size determination service

Figure 34 represents the architecture of the service. Similarly, to the aggregate quality determination service, the *Trainer* component aims at creating a model that can be used by the *Estimator*, taking as input rock images from the data lake and other generic data sets. These components are deployed in the cloud. The estimation components are deployed on premises and, using the model created by the trainer and images captured in the quarries along with other sensor data to correct according to the environmental conditions, provide an estimation of the grain sizes.

As was the case with the previous service, the *Aggregator* component buffers the results of the estimator to avoid data loss in case of connectivity issues and evaluates them before storing them in the data warehouse.

Interfaces:

Cloud:

Input	Source	Output	Destination
Images of quarries	Data lake	Grain size model	Estimator
Generic images	Generic data sets	Statistics	Data warehouse

Table 9: Interfaces for the grain size determination service (cloud)

On-premises:

Input	Source	Output	Destination
Environment data	Sensors	Grain size estimation	Data warehouse
Real time quarry images	Cameras	Metrics	Data warehouse
		Alarms	Data warehouse
		Corrective action	Quarry

Table 10: Interfaces for the grain size determination service (on-premises)

3.4.3.3 Stockpile volume calculation

This service aims at analyzing visual and data inputs to provide an estimation of the material volume in the different piles of the plants, allowing operators to keep track of the stock available across the quarry. This knowledge will also aid in optimizing production based on the stock level.

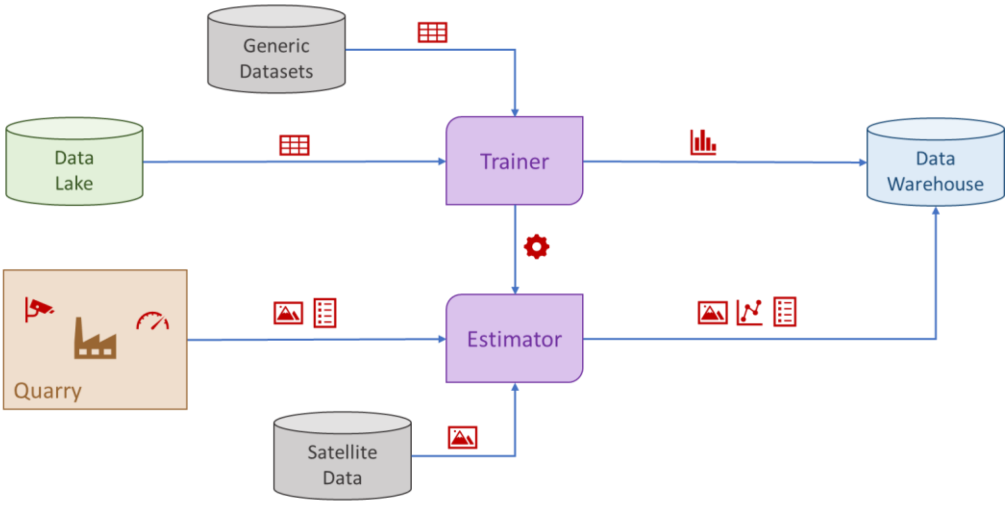


Figure 35: Base architecture of the stockpile volume calculation service

Figure 35 shows the architecture of the service. In contrast to the services described so far, all its components are deployed in the cloud, since there is no need for real-time data processing.

In this case, the goal of the *Trainer* is also to create a model that can be used by the *Estimator*, taking as input information from the data lake and other generic data sets, and storing metrics of this process in the data warehouse. The Estimator will make use of this model, along with images from the quarry (cameras and/or drone flights), environment information from sensors and, potentially, satellite data, etc., to generate an estimation of the volume of a certain stockpile. This information, together with the metrics of the process, will also be stored in the data warehouse.

Interfaces:

Input	Source	Output	Destination
Images of quarry stockpiles (historical)	Data lake	Stockpile volume estimation model	Estimator

Input	Source	Output	Destination
Generic stockpile images	Generic data sets	Statistics (training)	Data warehouse
Satellite Images	Public satellite imaging services	Stockpile volume estimation	Data warehouse
Images of quarry stockpiles	Cameras/drone flights	Metrics (estimation)	Data warehouse

Table 11: Interfaces for the stockpile volume calculation service

The aim of the stockpile volume is not just to get a static result but to track the current volume and to interpolate previous volumes when there is no further data (while awaiting updated inputs). This means the estimation is a single numeric value from an independent execution but the output from the service in time is a time series that shows the evolution of the stock.

3.4.3.4 Detection of mechanical failures

The anomaly detection service uses several kinds of devices (e.g., cameras, microphones, and sensors) to monitor the behavior of the production line in the quarry to detect and prevent the malfunction or the failure of machinery involved in the process. This service allows to implement a preventive maintenance system that, in turn, helps to lower production costs.

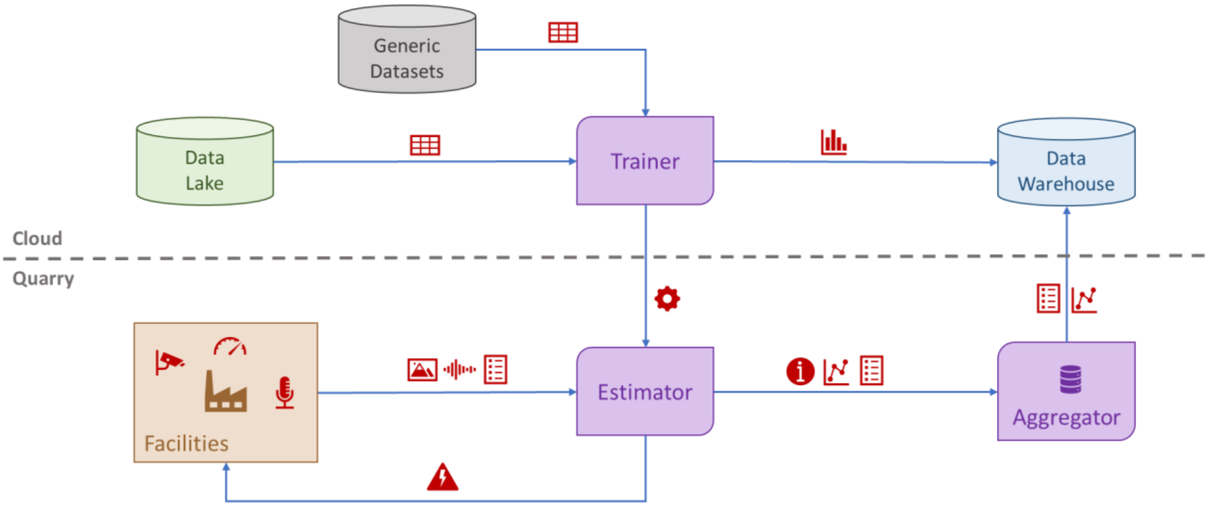


Figure 36: Base architecture of the anomaly detection of mechanical failures service

Figure 36 visually describes the architecture of the anomaly detection service. This service (as the aggregate quality and grain size determination ones) requires the *Estimator* to be executed in the quarry with real-time access to the information from monitoring devices. The estimator receives all sensor information and produces some results that are given to an *Aggregator* that schedules the insertion in the data warehouse. The estimator may eventually produce some alerts that may require sending signals to PLCs to adjust the operation of the machinery or messages to monitoring applications to inform about important issues related to machinery.

The execution of the *Estimator* is controlled by a *Trainer* that generates the models used to analyse the estimator's inputs. The trainer will use some external data sets to train an initial model that will be adjusted to the concrete quarry specificities using quarry data stored in the data lake. The training process will also produce some statistics that are used to evaluate the performance of the trained models

Interfaces:

Cloud:

Input	Source	Output	Destination
Historical data of anomalies	Data lake	Anomaly detection model	Estimator
Generic anomalies	Generic data sets	Statistics	Data warehouse

Table 12: Interfaces for the acoustic anomaly detection of mechanical failures service (cloud)

On-premises:

Input	Source	Output	Destination
Environment data	Sensors	Anomalies	Data warehouse
Machinery images	Cameras	Metrics	Data warehouse
Audio information	Microphones	Alarms	Data warehouse
		Corrective action	Quarry

Table 13: Interfaces for the acoustic anomaly detection of mechanical failures service (on-premises)

3.4.3.5 NLP information and document search engine (Metaquarry)

The NLP information and document search engine retrieves information from a knowledge base that contains documentation provided from each quarry. The goal is to find which documents are relevant to a query performed in natural language. When the query is formulated as a question the *Metaquarry* service will look for the response in the documents retrieved and return the answer in natural language.

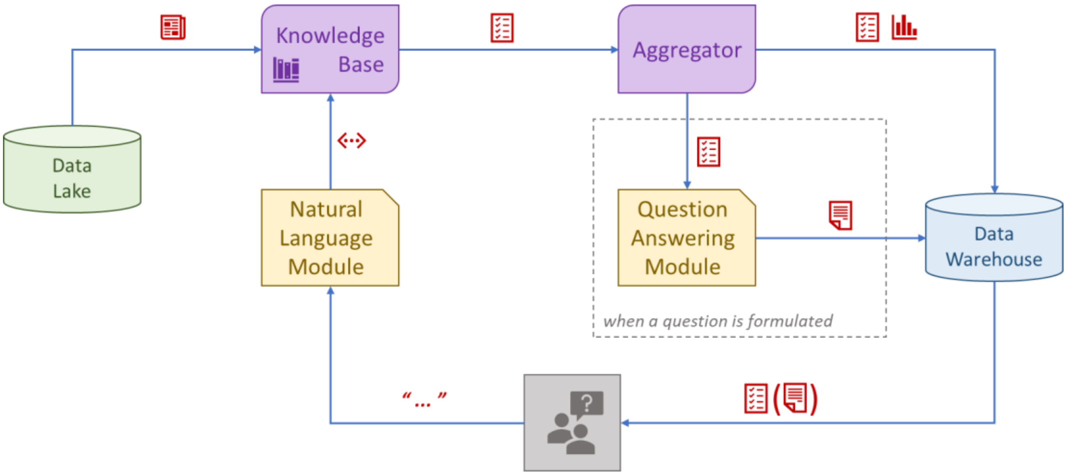


Figure 37: Base architecture of the NLP information and document search engine service

Figure 37 shows the base architecture of the NLP information and document search engine service. The user interacts with the system using an interface to be defined and gets a response for each natural language query he/she introduces.

There is a *Natural Language Module* that converts the provided sentence into a query to a *Knowledge Base*. The later produces a list of resulting documents for the query. Then an *Aggregator* component generates a response from the list of relevant documents using a *Question Answering Module* to make the final response available in the data warehouse.

The list of relevant documents returned by the *Knowledge Base* is calculated after indexing the documents in the data lake made available by each quarry. Note that in the other services the service naturally operates on each quarry in an independent way. The *Metaquarry* service requires a single instance per service.

Interfaces:

Input	Source	Output	Destination
Collection of documentation	Data lake	Document list	User/Data warehouse
Query	User	Textual response	User/Data warehouse
		Statistics	Data warehouse

Table 14: Interfaces for the NLP information and document search engine service

3.4.3.6 Consumptions and product forecasting

The consumption and product forecasting service analyzes operation information from the quarry to determine the real cost of production and the volume of material produced to get some insights about how to optimize the production in the sense of increase the profit of the quarry. The service also uses these estimations to produce some recommendations on how the operation can be adjusted to optimize the production.

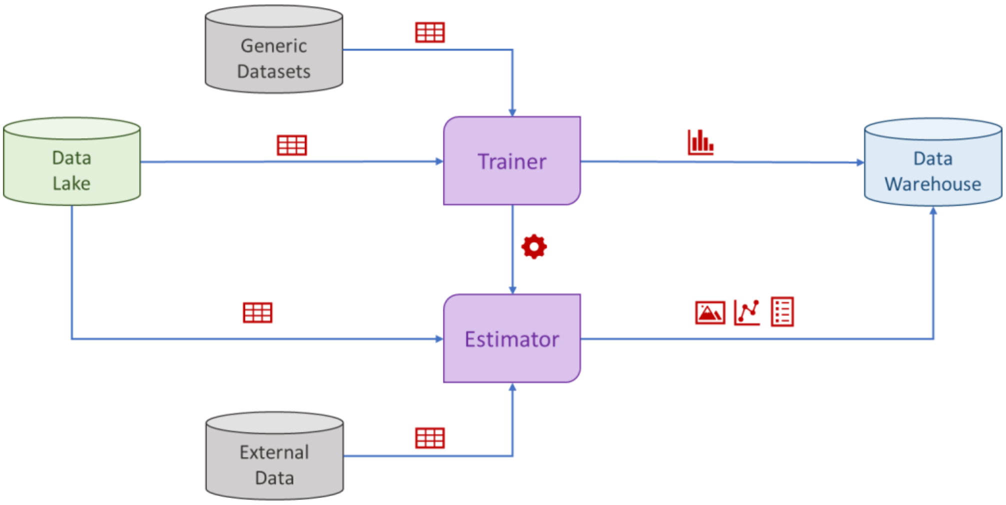


Figure 38: Base architecture of the consumptions and product forecasting service

Figure 38 illustrates the base architecture to deploy a service able to estimate the cost of production and provide some insights about actions to be taken to increase production efficiency. The architecture is divided into a training activity, that prepares an estimator for the concrete quarry and the estimation activity that analyses the operation data from the quarry. The *Trainer* component is responsible to retrieve historical data from the concrete quarry's data lake and some generic data that is publicly available, to train a model for the quarry. It will also store some training statistics in the data warehouse to evaluate its performance. The *Estimator* component is responsible of processing the latest unprocessed data from the quarry's data lake in combination with some external data, such as weather information, to produce a set of hourly estimations about production volume and costs that will be inserted into the data warehouse for reference.

During training, the system is fed with generic data such as the price of fuel, weather predictions, and some other publicly available data that may be relevant to train a model. After the first model was trained, the data used for training also includes specific data of the quarry the model is trained for. Training produces, as a result, the trained model and some statistics about the training itself, such as, training time, the error produced by the model, and some other information that is used to evaluate if the training was correct.

During estimation, the system is fed with the latest unseen data from the quarry. The estimator may also require data from external sources such as the weather forecast for several days ahead. The estimator then returns some metrics about the performance of the model, such as how much time it took to generate the forecast, or the deviation between the forecasted and the real scenarios. It also returns the forecast and some operation recommendations, that is, the estimator returns information of the production costs and volumes for the following hours and days, and some suggestion about when the best hours are to make machinery works in order to increase the quarry profit.

Interfaces:

Input	Source	Output	Destination
Historical production data	Data lake	Consumption model	Estimator
Generic production data	Generic data sets	Production and cost estimation	Data warehouse
External data (e.g., weather information)	Generic data sets	Statistics	Data warehouse

Table 15: Interfaces for the NLP information and document search engine service

3.4.4 Specificities by quarries

The selection of the services that will be implemented in each quarry has been carried out taking into account two considerations: the services should be useful and interesting for the quarries and all services should be implemented. To get an idea of the former, a survey was shared with all quarries' representatives, so that they could rate their interest in the different services. Once the answers to the surveys were received, the following service distribution was proposed and validated with quarry owners:

AI Service Area	HANSON	HOLCIM	VICAT	CIMPOR	CSI
Aggregate quality determination	x				
Grain size determination	x				
Stockpile volume calculation		x			
Detection of mechanical failures					x
NLP information and document search engine (Metaquarry)			x		
Consumptions & product forecasting				x	

Table 16: AI services distribution across quarries

It is important to note that this is a first approximation. As development progresses, the possibility of implementing services in quarries where they were not originally planned will be considered. Furthermore, results will be shared with all quarry owners to give them the chance to re-evaluate their interest.

3.5 BIM system

3.5.1 General BIM system architecture and interfaces

The intent of this section is to provide a general overview of how the BIM system integrates with the data lake and describe the overall BIM integration requirements.

BIM is planned to be used for 4D BIM planning and facility management. BIM for facility management provides visualization, access to the precise location and relationships of mining systems and equipment, and access to accurate existing condition attribute data. The main goal is to enhance project performance, produce better outcomes and produce an interactive BIM Common Data Environment (CDE) to enrich collaboration.

BIM service is divided into two different groups due to data required and integration differences.

- BIM Common Data Environment (CDE) for facility management.
- BIM Planning Environment

The planned scenario for BIM Common Data Environment (CDE) will directly integrate with the data lake using application programming interfaces or manually and store their results in Common Data Environment (CDE). BIM Common Data Environment (CDE) includes federated 3D BIM Models of each site and data to be received from Data Lake. All received data is stored facility data in 3D BIM element parameters and/or cloud-based databases of the BIM Expert system. The most important point here is the data will associate with the 3D BIM model elements in both cases.

The planned scenario for Planning Environment will manually fetch data from the data lake and store their results as a document locally. This service aims to optimize haulage process, improve planning quality, and risk mitigation due to visualization of the planned process.

Figure 39 illustrates the general architecture of the BIM system.

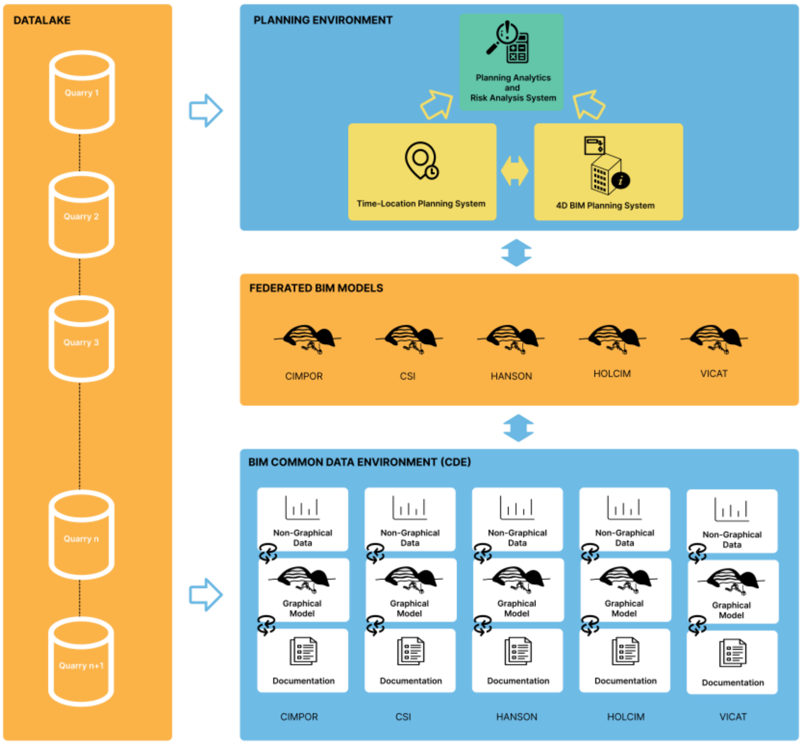


Figure 39: General architecture of BIM system

3.5.2 BIM Common Data Environment (CDE)

The BIM Common Data Environment service consumes mainly 3 different types of data: Graphical model, documentation and non-graphical data. The system connects BIM models and project data in one environment, which is deployed cloud. The project data and model will be associated with metadata stored in a database of Common Data Environment (CDE).

Graphical Model:

Graphical BIM models are developed and federated using design solutions such as Autodesk Revit, Bentley AECOsim, Graphisoft ArchiCAD, Tekla etc. A Federated BIM model means a set of 3D models related to specific disciplines (structural, MEP, machinery, etc.) that are integrated into a single view to create a single complete digital twin model of the building that is multidisciplinary and comprehensive. The purpose of generating a federated 3D BIM model of each site is sharing of information, coordination between disciplines and ease to use in the expert systems. Each quarry should have a divided 3D BIM model due to the different locations and different facilities and equipment types. These 3D BIM models will be stored and correlated separately in Planning Environment and BIM Common Data Environment (CDE). The important point here is that models can be shared in common file extensions supported by all expert system.

The table below shows the analysis of file extensions that the expert system supports.

Input	IFC 4	IFC 2x3	RVT	SKP	OBJ
4D BIM Planning System	x	x			
BIM Common Data Environment (CDE)	x	x	x	x	
Design Solutions	x	x	x	x	x

Table 17: Analysis of the 3D BIM model file extensions that the expert system supports

Graphical BIM Model is planned to be stored as IFC Schema that information can be shared in a format which enables and encourages interoperability.

Interfaces:

Input	Source	Output	Destination
Drone Flight Output	Pilot Sites	3D BIM models (IFC)	Design Solution
Design Documents	Pilot Sites	3D BIM models (IFC)	Design Solution

Table 18: Interfaces for 3D BIM models

Non-Graphical Model:

Non-graphical data consumes from the data lake using application programming interfaces (API) and manually. It can be analyzed in two ways: Static and dynamic data.

Static data refers to a fixed data set or, data that remains the same after it's collected. Static data includes data on facilities, machinery and assets such as machine model number, year of manufacture, crusher capacity, and area/volume data. These data will be stored directly in the model and integrated manually.

Dynamic data (IoT) refer to the data that continually changes after it's recorded in order to maintain its integrity such as energy consumption, frequency and operational mode. This type of data will be stored in the BIM Common Data Environment (CDE) database and integrated using application programming interfaces (API).

Figure 40 illustrates the Dynamic Data (IoT) General Architecture.

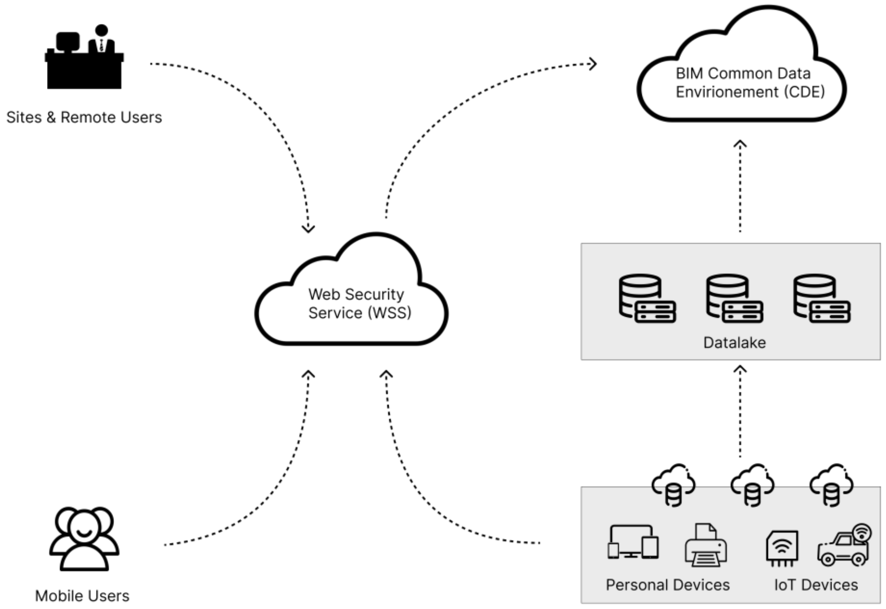


Figure 40: General Architecture of Dynamic Data (IoT)

Interfaces:

Input	Source	Output	Destination
Dynamic Data (IoT)	Data lake (Automated)	-	BIM Common Data Environment (CDE)
Static Data	Data lake (Manually)	-	3D BIM Model

Table 19: Interfaces for non-graphical data

Documentation:

Documents consume directly from partners or data lake according to availability. Although it is not possible to determine the document types at this stage of the project, to simplify the understanding the specification documents of the machines and maintenance documents of assets can be given as examples.

Interfaces:

Input	Source	Output	Destination
Documents	Data lake, Pilot Sites (Manually)	-	BIM Common Data Environment (CDE)

Table 20: Interfaces for documents

3.5.3 Planning Environment

Expert system in the planning environment requires windows-based computers. The planning environment is divided into 3 groups in itself.

- **4D BIM Planning** service is the only planning system in this group that will be integrated with the Federated 3D BIM model. It takes such as start time, finish time, quantity, quantity type, location data from the quarries that have been stored in the data lake and provide model-based scheduling and model-based estimating.
- **Time-Location Planning** service works together with the 4D BIM Planning system, and it is in charge of optimizing the haulage process according to the received data.
- **Planning Analytics and Risk Analysis** service take the data from 4D BIM Planning and Time-Location services manually and analyze the quality and risk of planning.

Interfaces:

Input	Source	Output	Destination
Mass Locations and Quantities	Data lake, Pilot Sites (Manually)	-	Time Location Planning, 4D BIM Planning System
Labour & Machinery Productivity Rate	Data lake, Pilot Sites (Manually)	-	Time Location Planning, 4D BIM Planning System
Hauling Distance	Data lake, Pilot Sites (Manually)	-	Time Location Planning, 4D BIM Planning System
Mass Type (Characterization of Earth)	Data lake, Pilot Sites (Manually)	-	Time Location Planning, 4D BIM Planning System
Labour & Machine Capacity	Data lake, Pilot Sites (Manually)	-	Time Location Planning, 4D BIM Planning System

Table 21: Interfaces for planning environment

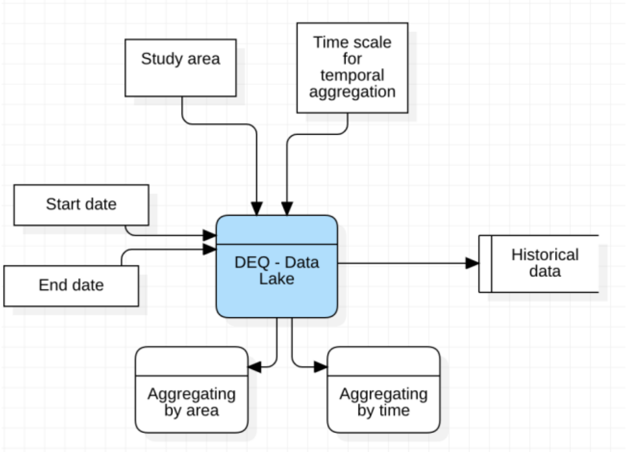
3.6 Reporting and Management tools

The management tool is the central part of the interaction with the end-user of the IQS and the DEQ. Because of the quarries, or rather of the workforce working in them. The management tools will not consider the typical tasks of IT departments (updating or changing the data lake architecture, computer infrastructures, assess potential threats, etc.).

The main function is to be a visualisation and reporting tool for the main KPI generated by the expert systems, combined with information available in the data lake to meaningful and easily digestible output. This avoids the user having to navigate between the different solutions of each partner. It is essential that all information is available in the data lake and can interconnect the information from the different systems. It is not convenient for the end-users to navigate between and use different tools for accessing information which could be provided conveniently in one single system. If the quarry data are compartmentalised in silos, it makes it impossible to create indicators (KPI) that involve data from different expert systems and company functions.

The main functions are:

- To allow the user to make modifications to the data stored in the data lake. Mainly because of errors in data collection or because of a later interpretation of the real circumstances of the quarry/operation.
- As manual input of quarry information or data. E.g., by emailing with attachments to the data lake.
- Being able to filter and aggregate data from all expert systems according to the user's needs. For example, display data by start and end date, aggregating by weeks, months, or days, by quarry, by country or area, etc.



- Enable the generation of performance, usage, or target evaluation reports (planned vs. actual).
- Enable planning and reporting to all sections of the quarry, both to operators (downstream) and to different departments of the organization (upstream).
- Export data to other data formats (e.g., csv or xlsx)
- Generate basic statistics according to needs (Exploratory Data Analysis), averages, maximums, minimums, standard deviation, etc.
- Generate graphical reports.
- Run Python or R scripts.

3.6.1 Results of the Benchmark for the best reporting software tool

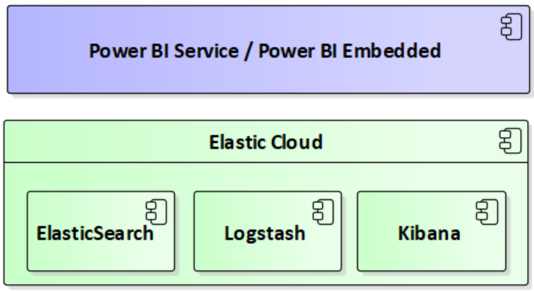
The full benchmark's study results done by AKKA are available in Appendix 7.1. Here below is a synthesis of the main results related to the Business Intelligence components.

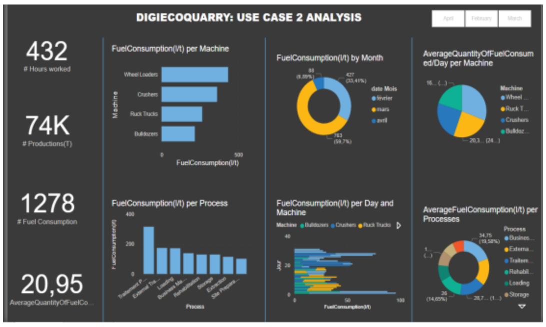
The global view of the components that will be used to build IQS Business Intelligence solution and the global cost are presented in section 3.3.1 Results of the Benchmark for the best IoT platform tools.

Below, some details are given for the business intelligence components:

- Description
- Metrics
- Costs

Components: Power BI and Elastic Cloud

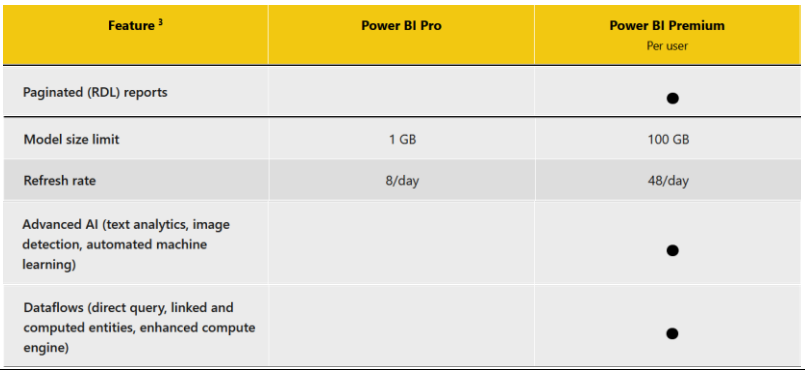




Description:

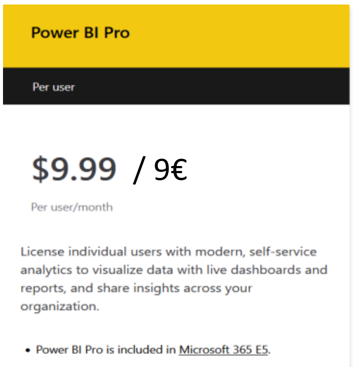
- Power BI is a Microsoft tool specifically dedicated to data exploration, analysis and visualization.
- Power BI offers the possibility to create dynamic and interactive dashboards.
- ELK Suite is used as BI Component over the Cloud.

Metrics:



Feature [3]	Power BI Pro	Power BI Premium Per user
Paginated (RDL) reports		●
Model size limit	1 GB	100 GB
Refresh rate	8/day	48/day
Advanced AI (text analytics, image detection, automated machine learning)		●
Dataflows (direct query, linked and computed entities, enhanced compute engine)		●

Costs:



Power BI Pro

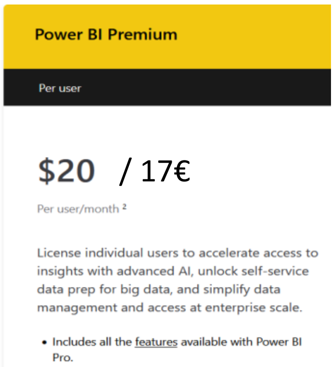
Per user

$9.99 / 9€

Per user/month

License individual users with modern, self-service analytics to visualize data with live dashboards and reports, and share insights across your organization.

- Power BI Pro is included in Microsoft 365 E5.



Power BI Premium

Per user

$20 / 17€

Per user/month [2]

License individual users to accelerate access to insights with advanced AI, unlock self-service data prep for big data, and simplify data management and access at enterprise scale.

- Includes all the features available with Power BI Pro.

3.6.2 Power BI ecosystem

The following figure shows the main components of Power BI solution and how different users contribute to the design of the dashboards and to the management of the environment, to make available the dashboards to the quarry end users.

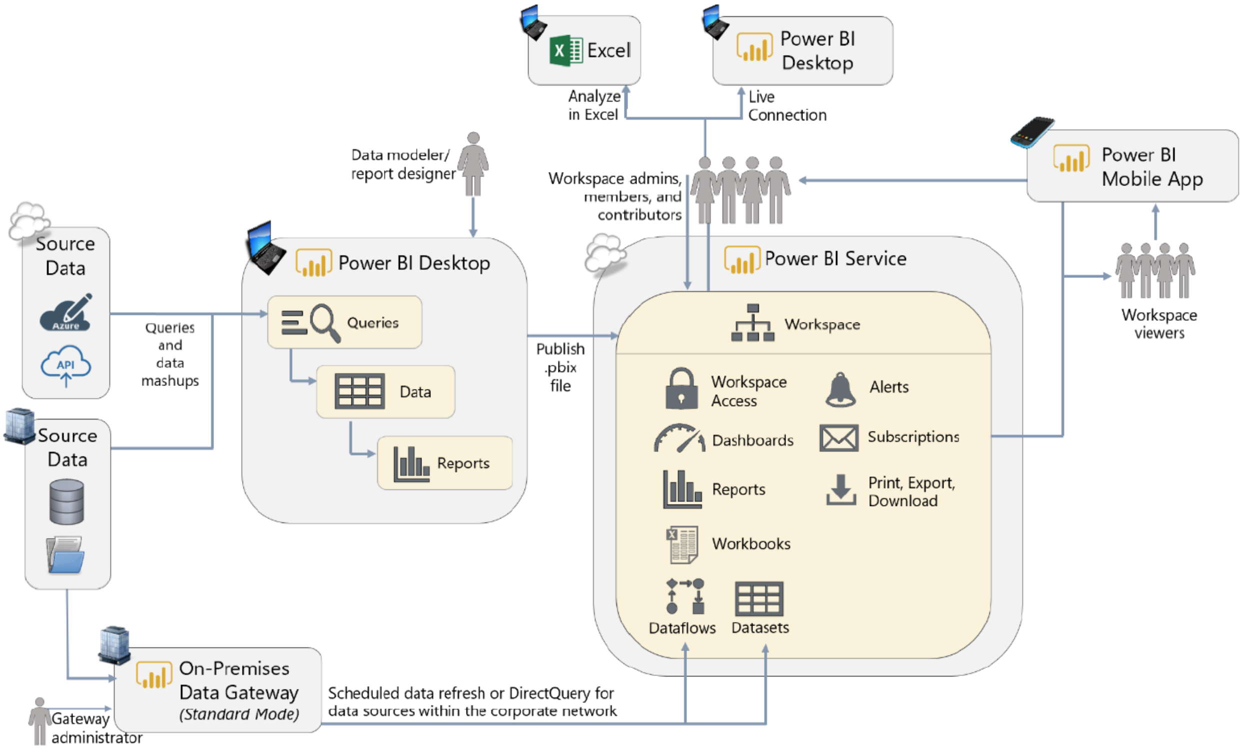


Figure 41: Power BI ecosystem and component's end users

Power BI Desktop is a desktop tools built for the analyst and used to:

- Create queries, datasets, import data from a wide variety of data sources
- Create relationships and enrich our data model with new measures and data formats
- Create, upload, publish and refresh publish reports

Power BI service is a cloud service where Power BI users can:

- Discover and access data, reports, dashboards and other business intelligence-related content which has been shared with them.
- Publish data, reports, dashboards and other business intelligence-related content that they have created.
- Connect to on-premises and cloud data sources seamlessly, with scheduled refresh.
- Share and distribute this content with authorized users, both inside and outside of the organization.

When a dataset author or report designer has finished developing and testing content created in Power BI Desktop, the .pbix file is published to a workspace in the Power BI service. There are two types of workspaces in the Power BI service:

- MyWorkspace: Every Power BI user has a private area called "My Workspace" which is intended purely for personal use.
- Workspace: Workspaces are shared workspaces where multiple users can collaborate. Workspaces are typically created for a specific purpose and a specific audience.

There are two modes[1] of interaction with reports in the Power BI service: Edit Mode and Read Mode. If you are a business user, you are more likely to use Read mode to consume reports created by other users. Edit mode is used by report designers, who create reports and share them with you. Read mode allows you to explore and interact with reports created by colleagues.

A user with a Power BI pro license can interact with a dashboard in either read or write mode depending on the permissions granted. A user with a free license can interact with the dashboards knowing that they are in a workspace with a Premium capacity.

[1] https://docs.microsoft.com/fr-fr/power-bi/consumer/end-user-reading-view

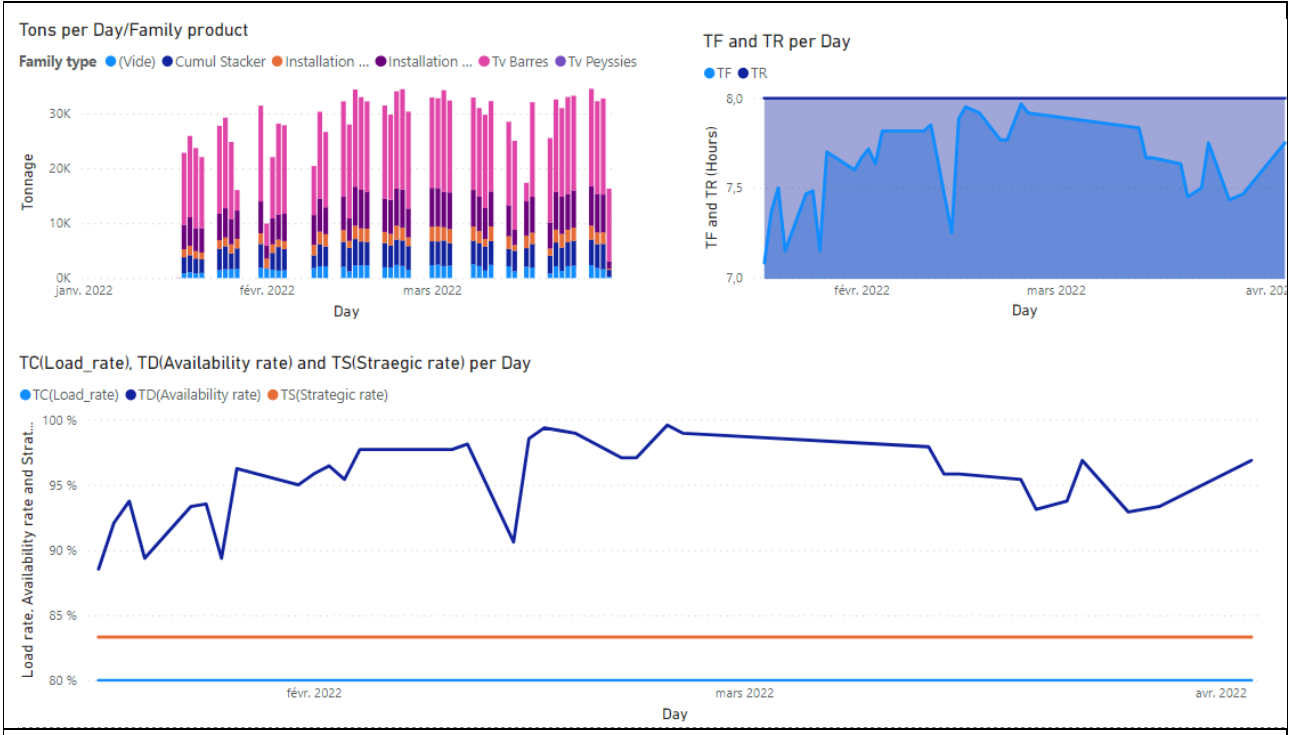
Multiple data sources (flat files, csv, excel) will be used. Postgres SQL Database will be used as the main database when the RDBMS is adapted. Concerning flat files (csv, excel...) and thanks to power query, retrieved files can be transformed before creating the reports, or loaded directly into Power BI to create the report. The integration of Power BI solution with the IQS is described in section 4.3 : IQS Integration: Focus on Power BI.

On-Premises Personal Data Gateway is used to refresh at regular intervals datasets uploaded to Power BI Services. An Enterprise gateway may be used to securely refresh corporate datasets in Power BI Service. A schedule refresh plan can be defined to schedule when the data model and the dashboards must be refreshed. Thereby, Power BI Pro license allows up to 8 refresh per day, while a Power BI Premium license allows to schedule up to 48 refresh per day.

3.6.3 Business management tools: First prototypes

This section will show the first prototypes created with Power BI using first datasets examples provided by DigiEcoQuarry partners and covering different quarry processes. The goal is to initiate these business management tools activities in an agile basis by enabling the creation of robust, common, and useful dashboard that convers project's needs.

Treatment plant production dashboard in Vicat Fenouillet pilot site:

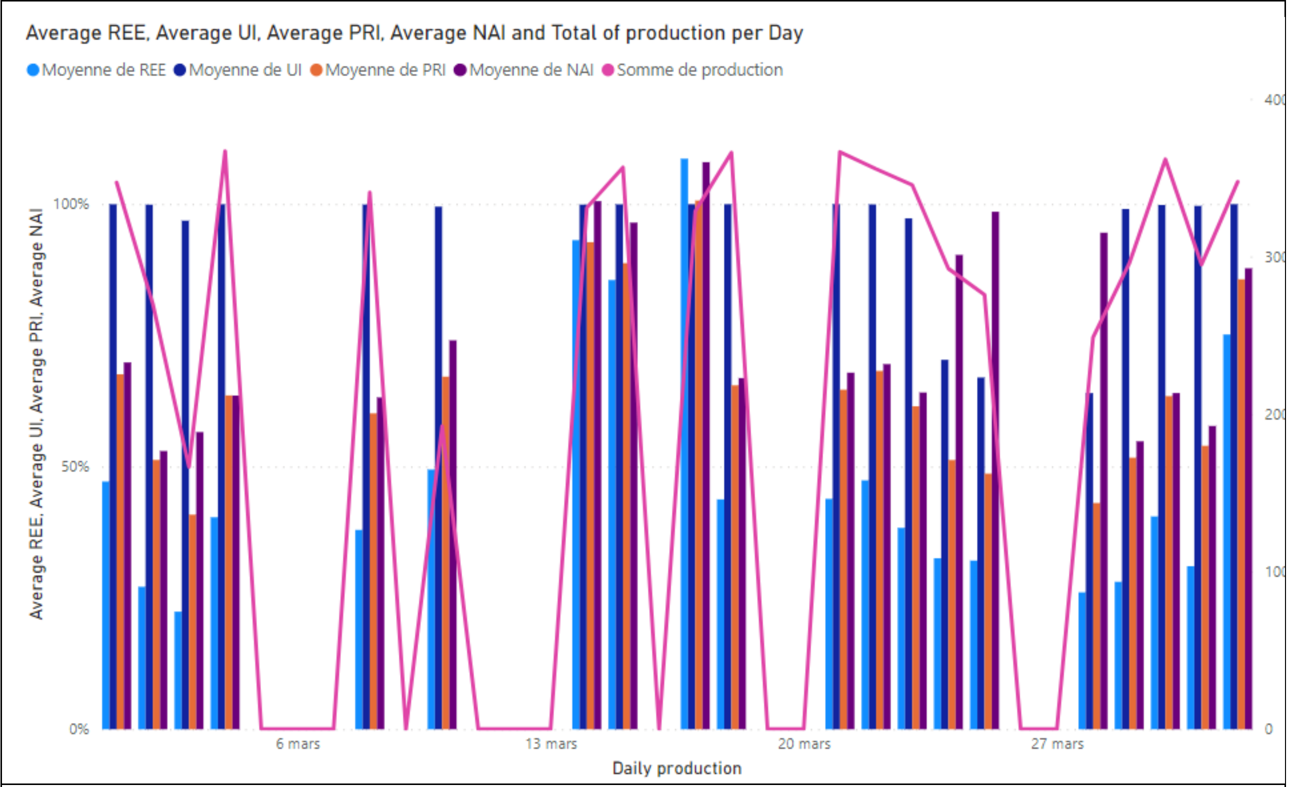

Tons per Day/Family product
Family type ● (Vide) ● Cumul Stacker ● Installation ... ● Installation ... ● Tv Barres ● Tv Peyssies
Tonnage
30K
20K
10K
0K
janv. 2022
févr. 2022
mars 2022
Day
TF and TR per Day
● TF ● TR
TF and TR (Hours)
8,0
7,5
7,0
févr. 2022
mars 2022
avr. 202
Day
TC(Load_rate), TD(Availability rate) and TS(Straegic rate) per Day
● TC(Load_rate) ● TD(Availability rate) ● TS(Strategic rate)
Load rate, Availability rate and Strat...
100 %
95 %
90 %
85 %
80 %
févr. 2022
mars 2022
avr. 2022
Day

To build this dashboard, we have relied on the Excel data provided by Vicat containing information on operating time, water consumption according to productions.

The first histogram shows the daily production (in tons) per product Type. Data can be filtered and displayed per day, week, etc.

The second diagram shows TF (operating time) and TR (required time) that are calculated according to the opening hours, maintenance hours.
The third diagram shows TC (load rate), TD (availability rate), TS (strategic rate) that are calculated according to different times (operating hours, required hours, opening hours). These indicators are expressed as a percentage and defined by period, in our case daily.

Treatment plant production dashboard in Holcim using a direct connection to Maestro/scada system:

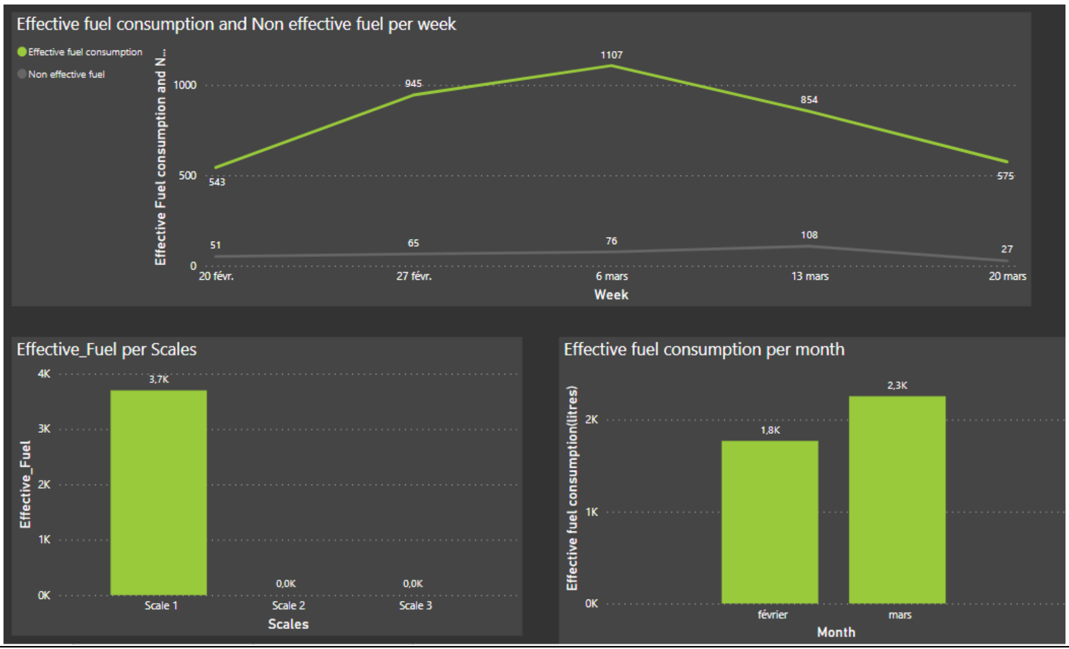


This dashboard represents the Running equipment effectiveness (REE), Net Availability Index –Aggregates (NAI)(%), Utilization Index –Aggregates (UI)(%), Production Rate Index–Aggregates (PRI)(%)) and the production per day,

To calculate those KPI we used $RunningTime, LackofFeedTime, ActualOperationTime$ present on a JSON file retrieved by calling the Rest API provided by QProduction cloud platform developed by Maestro for Holcim plant.

Since Holcim and Maestro are able to provide the data through an API REST in JSON format, we have created a python script to collect this production data in a daily basis then control and insert this data on a Postgres SQL Database. Power BI has the ability to connect with a Postgres SQL base, compute the KPI and generate the reports.

Fuel consumption dashboard using Metso's data related to mobile crusher:



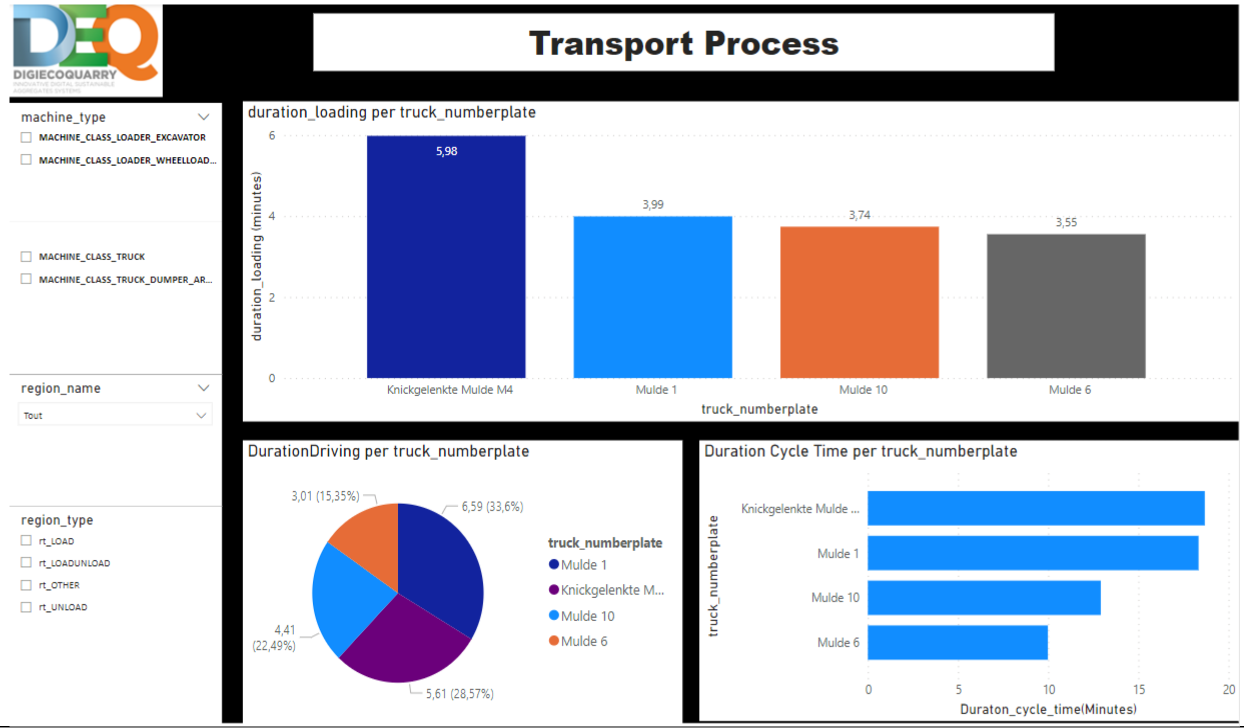
To build this dashboard, we have relied on the Excel data provided by Metso containing information on fuel consumption of mobile crusher to be deployed in Vicat.

The first one describes the proportion between the effective fuel consumption and the non-effective consumption of fuel per week. It can be seen that in the first two weeks there is a considerable increase in the consumption of effective fuel followed by a decrease in the next two weeks.

The second is the consumption of effective fuel per scale. It can be seen that only the scale 1 has been filled for the moment.

The third display shows an analysis of the Effective fuel consumption per month. This will allow you to see the actual amount of fuel consumed per month and take the right decisions based on this.

Transport process Dashboard using Abaut data related to mobile machinery:



To build this dashboard, we relied on the CSV data provided by Abaut containing information related to transportation and mobile machinery over a period of time.

The first describes the duration loading per truck identified by number plate.

The second is the duration of a cycle per truck which includes loading time, driving time and unloading time.

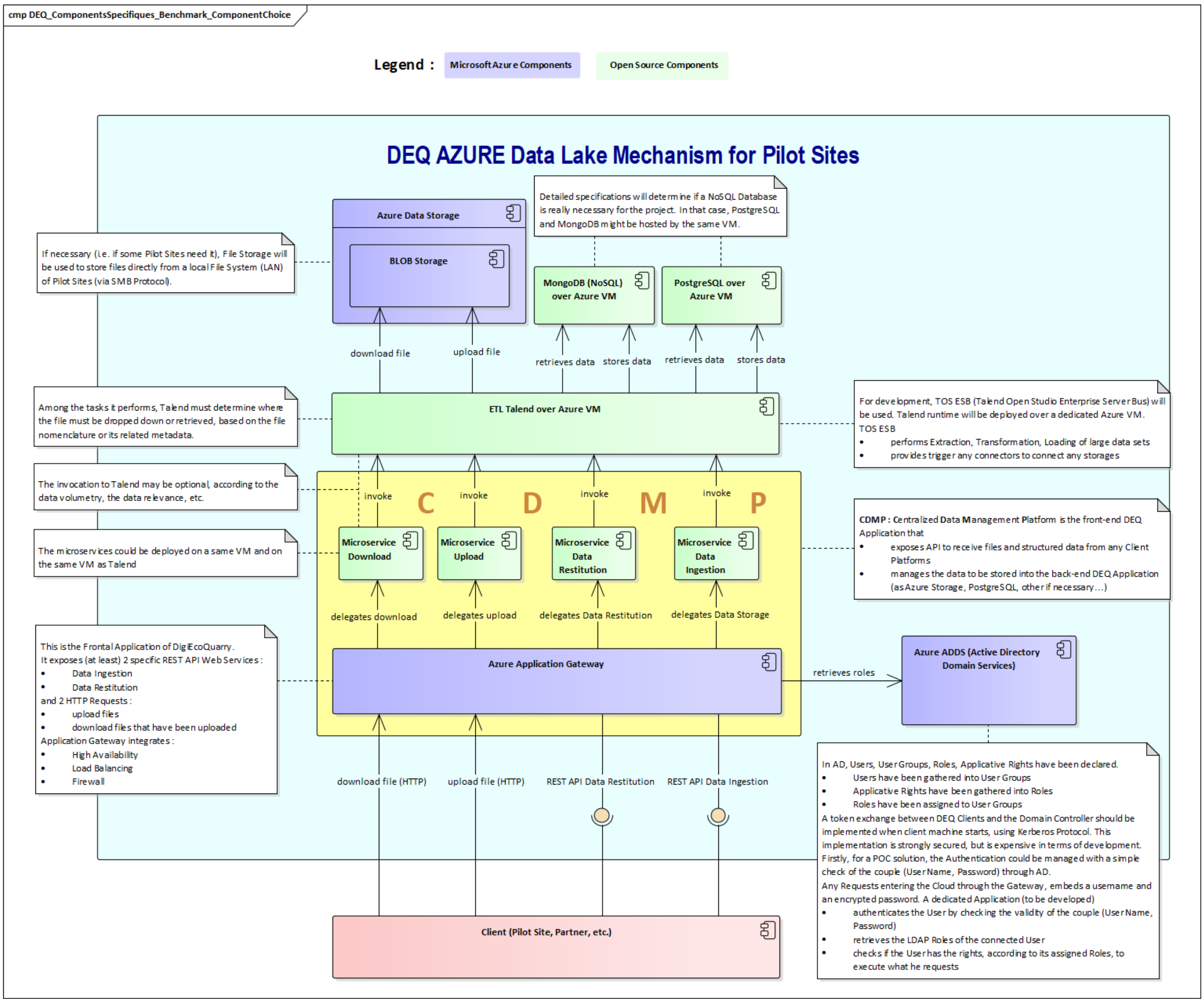
The third is a pie chart showing the duration of driving in minutes from the load location to the unload location per truck.

On the left, the data can also be filtered by machine type, region name, region type.

4 Global IQS integration

This section shows and describes all the components to be deployed or developed and how they will be integrated. For each subsection, an UML components diagram depicts in several layers the components, their roles and how they communicate.

4.1 IQS Integration: Focus on Data Lake



Detailed specifications will determine if a NoSQL Database is really necessary for the project. In that case, PostgreSQL and MongoDB might be hosted by the same VM.

If necessary (i.e. if some Pilot Sites need it), File Storage will be used to store files directly from a local File System (LAN) of Pilot Sites (via SMB Protocol).

Among the tasks it performs, Talend must determine where the file must be dropped down or retrieved, based on the file nomenclature or its related metadata.

For development, TOS ESB (Talend Open Studio Enterprise Server Bus) will be used. Talend runtime will be deployed over a dedicated Azure VM.
TOS ESB

- performs Extraction, Transformation, Loading of large data sets
- provides trigger any connectors to connect any storages

The invocation to Talend may be optional, according to the data volumetry, the data relevance, etc.

The microservices could be deployed on a same VM and on the same VM as Talend

CDMP : Centralized Data Management Platform is the front-end DEQ Application that

- exposes API to receive files and structured data from any Client Platforms
- manages the data to be stored into the back-end DEQ Application (as Azure Storage, PostgreSQL, other if necessary…)

This is the Frontal Application of DigiEcoQuarry.
It exposes (at least) 2 specific REST API Web Services :

- Data Ingestion
- Data Restitution

and 2 HTTP Requests :

- upload files
- download files that have been uploaded

Application Gateway integrates :

- High Availability
- Load Balancing
- Firewall

In AD, Users, User Groups, Roles, Applicative Rights have been declared.

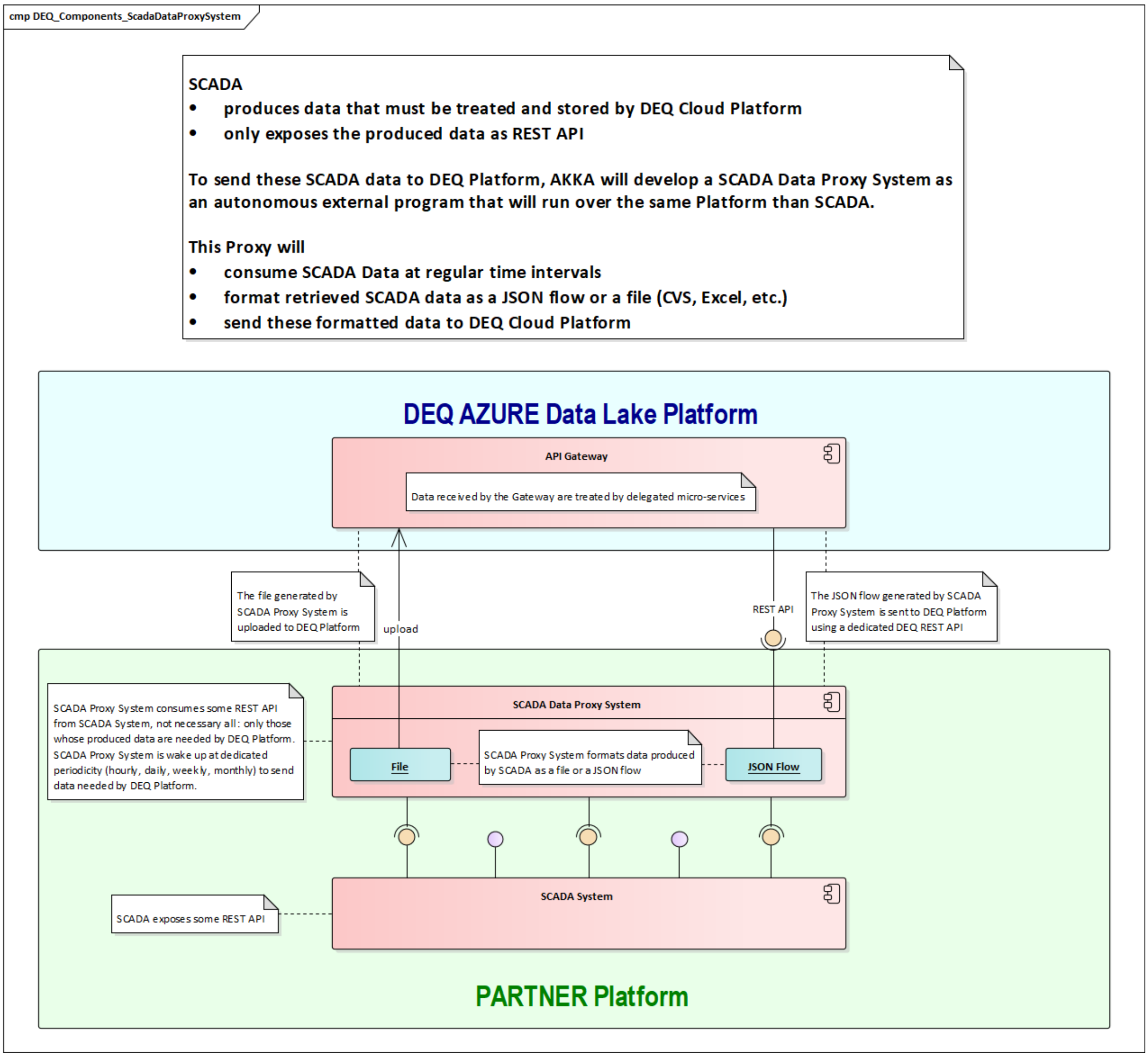
- Users have been gathered into User Groups
- Applicative Rights have been gathered into Roles
- Roles have been assigned to User Groups

A token exchange between DEQ Clients and the Domain Controller should be implemented when client machine starts, using Kerberos Protocol. This implementation is strongly secured, but is expensive in terms of development. Firstly, for a POC solution, the Authentication could be managed with a simple check of the couple (User Name, Password) through AD.
Any Requests entering the Cloud through the Gateway, embeds a username and an encrypted password. A dedicated Application (to be developed)

- authenticates the User by checking the validity of the couple (User Name, Password)
- retrieves the LDAP Roles of the connected User
- checks if the User has the rights, according to its assigned Roles, to execute what he requests

Figure 42: IQS Integration - Focus on Data Lake Platform

4.2 IQS Integration: Focus on Generic Data Provider Proxy



SCADA

- produces data that must be treated and stored by DEQ Cloud Platform
- only exposes the produced data as REST API

To send these SCADA data to DEQ Platform, AKKA will develop a SCADA Data Proxy System as an autonomous external program that will run over the same Platform than SCADA.

This Proxy will

- consume SCADA Data at regular time intervals
- format retrieved SCADA data as a JSON flow or a file (CVS, Excel, etc.)
- send these formatted data to DEQ Cloud Platform

Data received by the Gateway are treated by delegated micro-services

The file generated by SCADA Proxy System is uploaded to DEQ Platform

The JSON flow generated by SCADA Proxy System is sent to DEQ Platform using a dedicated DEQ REST API

SCADA Proxy System consumes some REST API from SCADA System, not necessary all: only those whose produced data are needed by DEQ Platform. SCADA Proxy System is wake up at dedicated periodicity (hourly, daily, weekly, monthly) to send data needed by DEQ Platform.

SCADA Proxy System formats data produced by SCADA as a file or a JSON flow

SCADA exposes some REST API

Figure 43: IQS Integration - Focus on Generic Data Provider Proxy

4.3 IQS Integration: Focus on Power BI

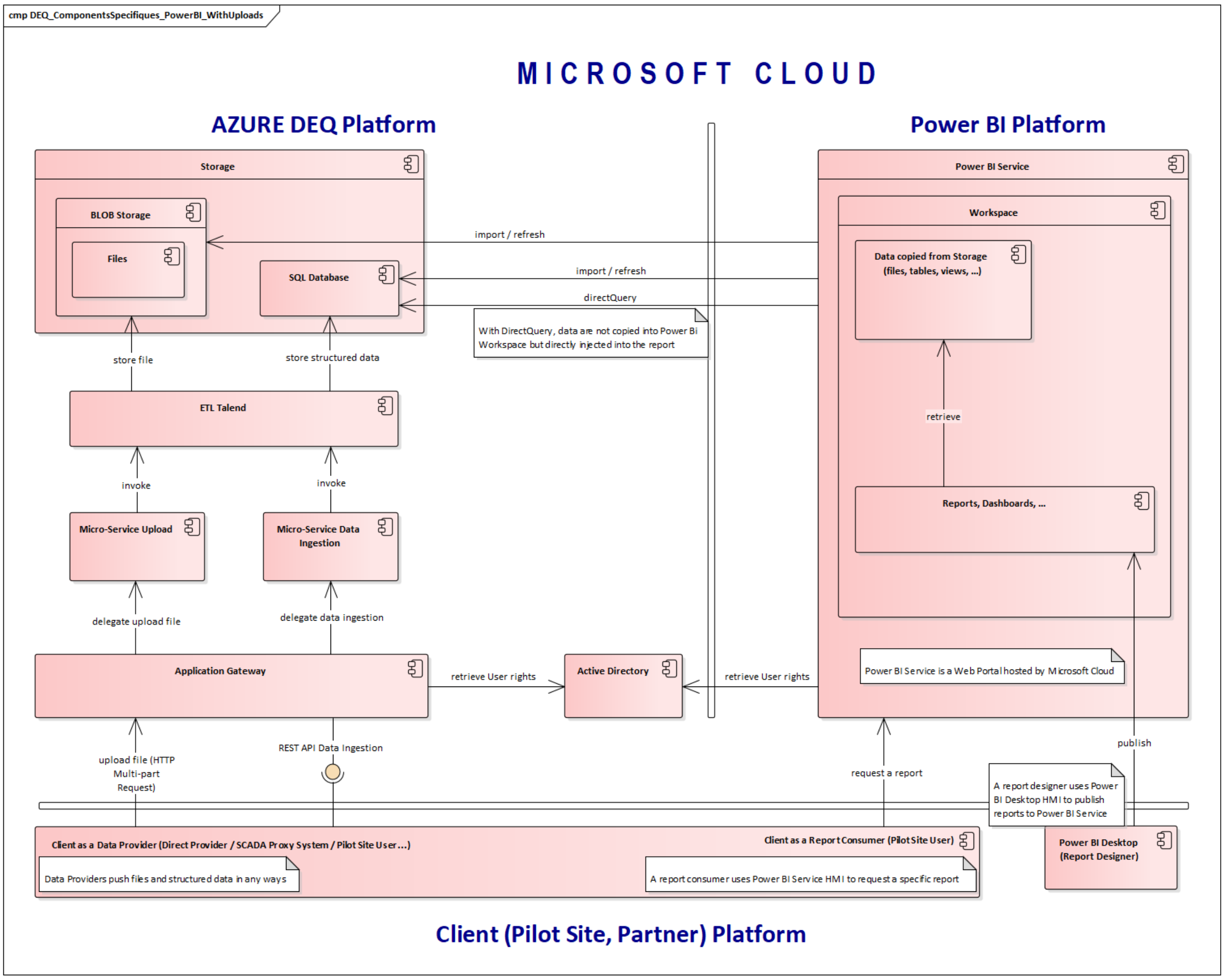
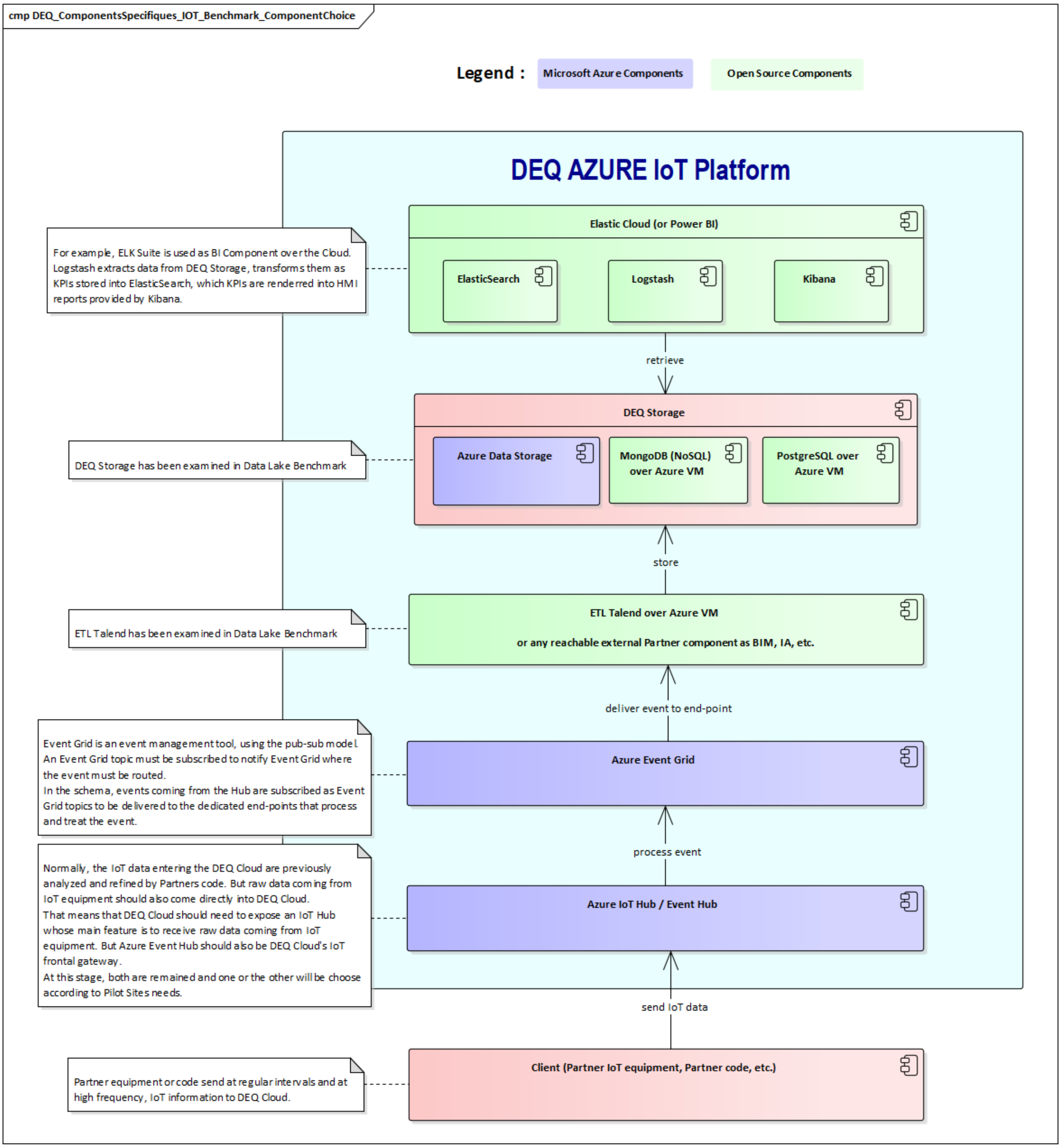

cmp DEQ_ComponentsSpecifiques_PowerBI_WithUploads
MICROSOFT CLOUD
AZURE DEQ Platform
Power BI Platform
Storage
BLOB Storage
Files
SQL Database
import / refresh
import / refresh
directQuery
With DirectQuery, data are not copied into Power BI Workspace but directly injected into the report
store file
store structured data
ETL Talend
invoke
invoke
Micro-Service Upload
Micro-Service Data Ingestion
delegate upload file
delegate data ingestion
Application Gateway
retrieve User rights
Active Directory
retrieve User rights
Power BI Service
Workspace
Data copied from Storage (files, tables, views, ...)
retrieve
Reports, Dashboards, ...
Power BI Service is a Web Portal hosted by Microsoft Cloud
upload file (HTTP Multi-part Request)
REST API Data Ingestion
request a report
publish
A report designer uses Power BI Desktop HMI to publish reports to Power BI Service
Client as a Data Provider (Direct Provider / SCADA Proxy System / Pilot Site User ...)
Data Providers push files and structured data in any ways
Client as a Report Consumer (Pilot Site User)
A report consumer uses Power BI Service HMI to request a specific report
Power BI Desktop (Report Designer)
Client (Pilot Site, Partner) Platform

Figure 44: IQS Integration - Focus on Power BI

4.4 IQS Integration: Focus on IoT Platform


cmp DEQ_ComponentsSpecifiques_IOT_Benchmark_ComponentChoice
Legend :
Microsoft Azure Components
Open Source Components
DEQ AZURE IoT Platform
Elastic Cloud (or Power BI)
ElasticSearch
Logstash
Kibana
retrieve
DEQ Storage
Azure Data Storage
MongoDB (NoSQL) over Azure VM
PostgreSQL over Azure VM
store
ETL Talend over Azure VM
or any reachable external Partner component as BIM, IA, etc.
deliver event to end-point
Azure Event Grid
process event
Azure IoT Hub / Event Hub
send IoT data
Client (Partner IoT equipment, Partner code, etc.)

For example, ELK Suite is used as BI Component over the Cloud. Logstash extracts data from DEQ Storage, transforms them as KPIs stored into ElasticSearch, which KPIs are renderred into HMI reports provided by Kibana.

DEQ Storage has been examined in Data Lake Benchmark

ETL Talend has been examined in Data Lake Benchmark

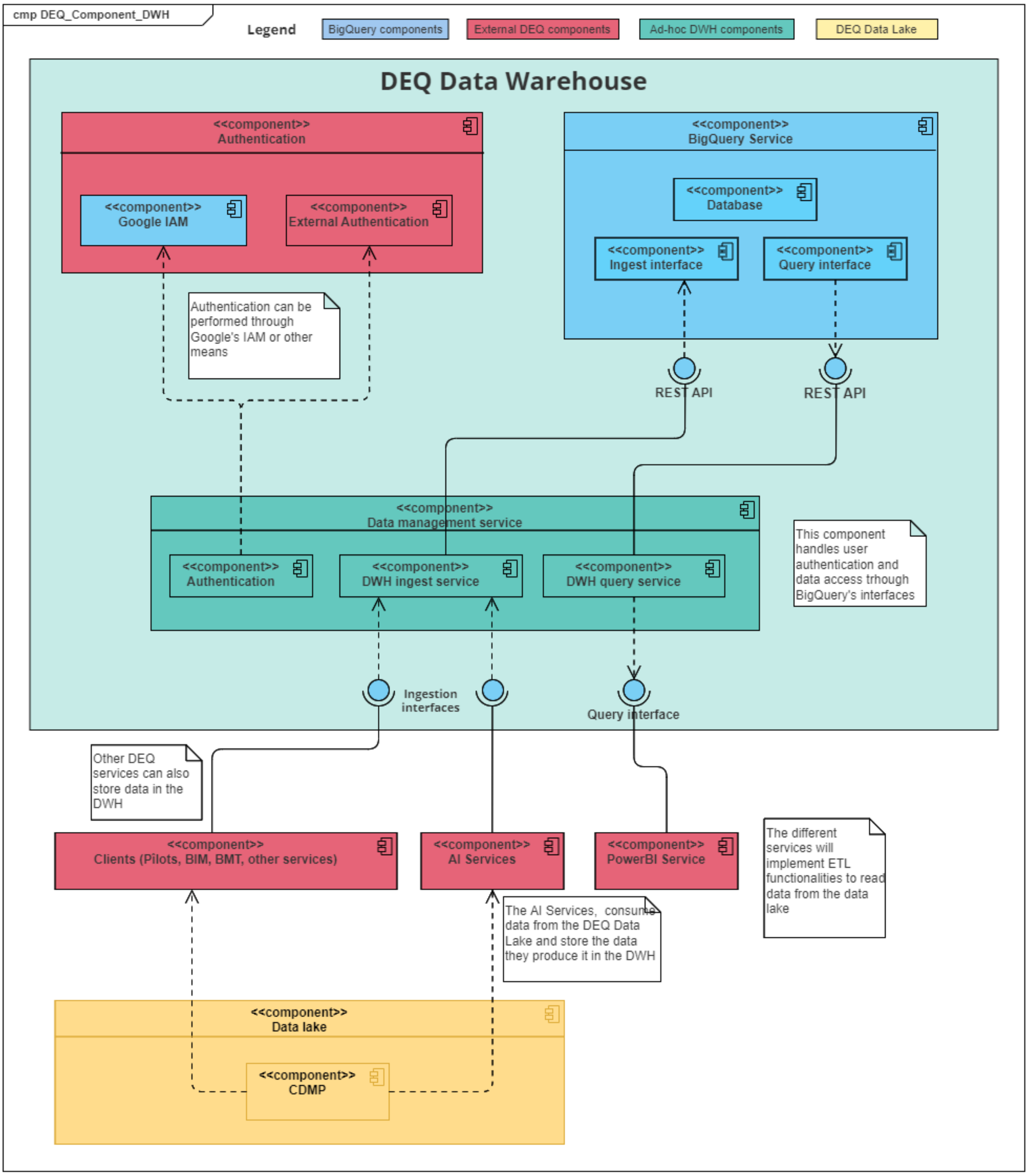
Event Grid is an event management tool, using the pub-sub model. An Event Grid topic must be subscribed to notify Event Grid where the event must be routed.
In the schema, events coming from the Hub are subscribed as Event Grid topics to be delivered to the dedicated end-points that process and treat the event.

Normally, the IoT data entering the DEQ Cloud are previously analyzed and refined by Partners code. But raw data coming from IoT equipment should also come directly into DEQ Cloud.
That means that DEQ Cloud should need to expose an IoT Hub whose main feature is to receive raw data coming from IoT equipment. But Azure Event Hub should also be DEQ Cloud's IoT frontal gateway.
At this stage, both are remained and one or the other will be choose according to Pilot Sites needs.

Partner equipment or code send at regular intervals and at high frequency, IoT information to DEQ Cloud.

Figure 45: IQS Integration - Focus on IoT Platform

4.5 IQS Integration: Focus on Data Warehouse


cmp DEQ_Component_DWH
Legend
BigQuery components
External DEQ components
Ad-hoc DWH components
DEQ Data Lake
DEQ Data Warehouse
<<component>> Authentication
<<component>> Google IAM
<<component>> External Authentication
<<component>> BigQuery Service
<<component>> Database
<<component>> Ingest interface
<<component>> Query interface
REST API
REST API
<<component>> Data management service
<<component>> Authentication
<<component>> DWH ingest service
<<component>> DWH query service
Ingestion interfaces
Query interface
<<component>> Clients (Pilots, BIM, BMT, other services)
<<component>> AI Services
<<component>> PowerBI Service
<<component>> Data lake
<<component>> CDMP

Authentication can be performed through Google's IAM or other means

This component handles user authentication and data access trhough BigQuery's interfaces

Other DEQ services can also store data in the DWH

The AI Services, consume data from the DEQ Data Lake and store the data they produce it in the DWH

The different services will implement ETL functionalities to read data from the data lake

Figure 46: IQS Integration - Focus on Data Warehouse

5 Conclusions

The ICT requirements analysis and assets inventory done in the frame of the WP4 and reported in this public deliverable D4.1 allowed to go more deeply in the descriptions of what could be the best digitalisation tools to implement on quarries to reach the DigiEcoQuarry project's main challenges: health & safety, security, efficiency, selectivity & profitability, environmental impact and social acceptance.

The five pilot sites will contribute, at different processes levels, to experiment, for the aggregates industry, the different solutions and digitalisation tools.

After the identification of the assets, all data flows between the partners or systems within this pilot site were identified. Data flows served as a starting point to the definition of the data sharing interfaces and data models to be used by the business management tools.

For all these pilot sites, the collaboration between the different involved partners will be facilitated by the implementation of the IQS which will allow the sharing of all the relevant data. Based on a deep ICT requirements analysis and on prototyping activities, AKKA performed a detailed benchmark study (refer to section 7.1) on the data lake components, IoT platform components and business intelligence tools. The best solution proposed for the digitalisation of the aggregates industry is composed by Microsoft Azure cloud components (Azure gateway, Active directory…) mixed with open-sources tools (microservices, ETL Talend, MongoDB…) for the DEQ data lakes (refer to section 3.2.1). For the DEQ IoT platforms (refer to section 3.3.1) and for the reporting and business management tools (refer to section 3.6.1**¡Error! No se encuentra el origen de la referencia.**), Microsoft Azure components (Azure IoT hub, event grid, event hub, Power BI…) mixed with open-sources tools (Talend, ELK suite) have also been selected. On their side, SIGMA and UPM-AI also performed a deep benchmark analysis on data warehouse components (refer to section 3.4.1) This benchmark concludes to the selection of the BigQuery application (Google component) as the best data warehouse solution for the quarries. This data warehouse solution will allow the storage of the results coming from the six AI services that will be delivered on the different pilot sites, as proposed by Sigma and UPM-AI. All these solutions have been costed.

On top of the data lake, the IQS will contain a CDMP, a centralised and structured platform to be developed by AKKA, to collect and store the pilot sites' data shared, and to allow IQS users to browse, access and download data thanks to REST APIs and web interface. The CDMP is the recommended way for nominal or customized data accesses nevertheless Azure data lake also offer native access to azure features, but these features are more recommended for special needs.

A harmonized approach for data collection and data sharing between the IQS and the main partners' expert systems (Maestro's SCADA, DH&P and Abaut ES) is described. This data push system "Data Proxy System" will consume the ES data at regular basis, format this data as a json flow or specific format files and then upload the formatted data to the data lake.

The IoT platform of the IQS will enable data sharing of IoT data. IoT components will be used to integrate the data necessary for the building of digital twins of the quarries. APP will provide such BIM service based on BIM Common Data Environment (CDE), the Planning Environment and the data available at PS.

All the data collected can be used by the business management tools (Power BI, ELK suite) to create dynamic dashboards for any business case. These dashboards can be then shared and distributed with authorized users, both inside and outside of the organization.

Data integration has been the cornerstone of the digital transformation, enabling the sharing and processing of data across the enterprise to enable data-driven decision making. Within the IQS, the data integration will make and extensive usage of the cutting-edge technologies, data processing patterns and reference architecture to build the IQS (Rest API, Talend, Microservice, Push principle, CDMP, data lake) while focusing on scalability, performance, and ease of development.

Finally, the global IQS integration and the interfaces will be detailed and implemented in task 4.2 (ICT platform design and implementation led by AKKA), task 4.3 (Data warehouse-AI led by UPM-AI and SIGMA) and task 4.4 (BIM integration led by APP) in close collaboration with the different KTAs' leaders and in line with the deployment coordinated by WP6 (Pilot scenarios for quarrying operations monitoring & assessment) led by Holcim.

6 References

Document Resource ID	Document Resource name and reference
DR1	EU Grant Agreement n°101003750
DR2	D1.1 Requirements for Improved extraction, rock mass characterisation and control report
DR3	D1.2 Requirements for Innovative Treatment processes
DR4	D1.3 Requirements for Quarry full digitalisation (for Smart Sensors, Automation &Process Control, and for ICT solutions, BIM and AI report
DR5	D1.4 Requirements for H&S improvement, Environmental impact minimization and energy and resources efficiency report
DR6	D3.1 List and characterisation of key data inputs

7 Appendix

7.1Benchmark for the best digitalisation tools (data lake, IoT platform elements and Business Intelligence)

7.1.1 Introduction

In the scope of the task 4.1, “ICT requirements analysis and assets inventory”, one of the activities was to perform a benchmark to select the best digitalisation tools (data lake, IoT platform elements and data warehouse) by considering the state of the art, defining evaluation criteria, and identifying potential solutions.

AKKA team contributed mainly to the benchmark study related to data lake and IoT platform elements while SIGMA and UPM/AI teams worked mainly on data warehouse elements. Both teams shared their results to produce this document.

Our approach was first to study the state of the art related to cloud solutions. This study, combined to our deep analysis of the requirements and first high-level architecture described in the Deliverable 1.3 (DR2), allows us to choose the components to study in more details in this benchmark. Then, we defined the relevant hypothesis and evaluation criteria to be analyzed in the frame of the DigiEcoQuarry project i.e., according to the estimated use cases to be implemented in the quarries. We identified mainly three levels of use: weak, medium, and intensive. According to these levels of use, we were able to propose a costing of the different possible solutions.

Compared to what has been presented in the Deliverable 1.3 (DR2), we finally decided to set out of scope the Google cloud provider as it doesn’t bring additional added value compared to the other two big cloud providers for a same level of price. AKKA estimated more relevant to minimize the costs while offering a sustainable solution, to study Open-Source tools, such as Talend (ETL), ElasticSearch or PostgreSQL.

Note that some AWS and Azure specific components have been removed from this study as considered as finally not fully necessary to be implemented for quarries usages; this will be detailed within this document.

The **dissemination level** of this deliverable is **public**.

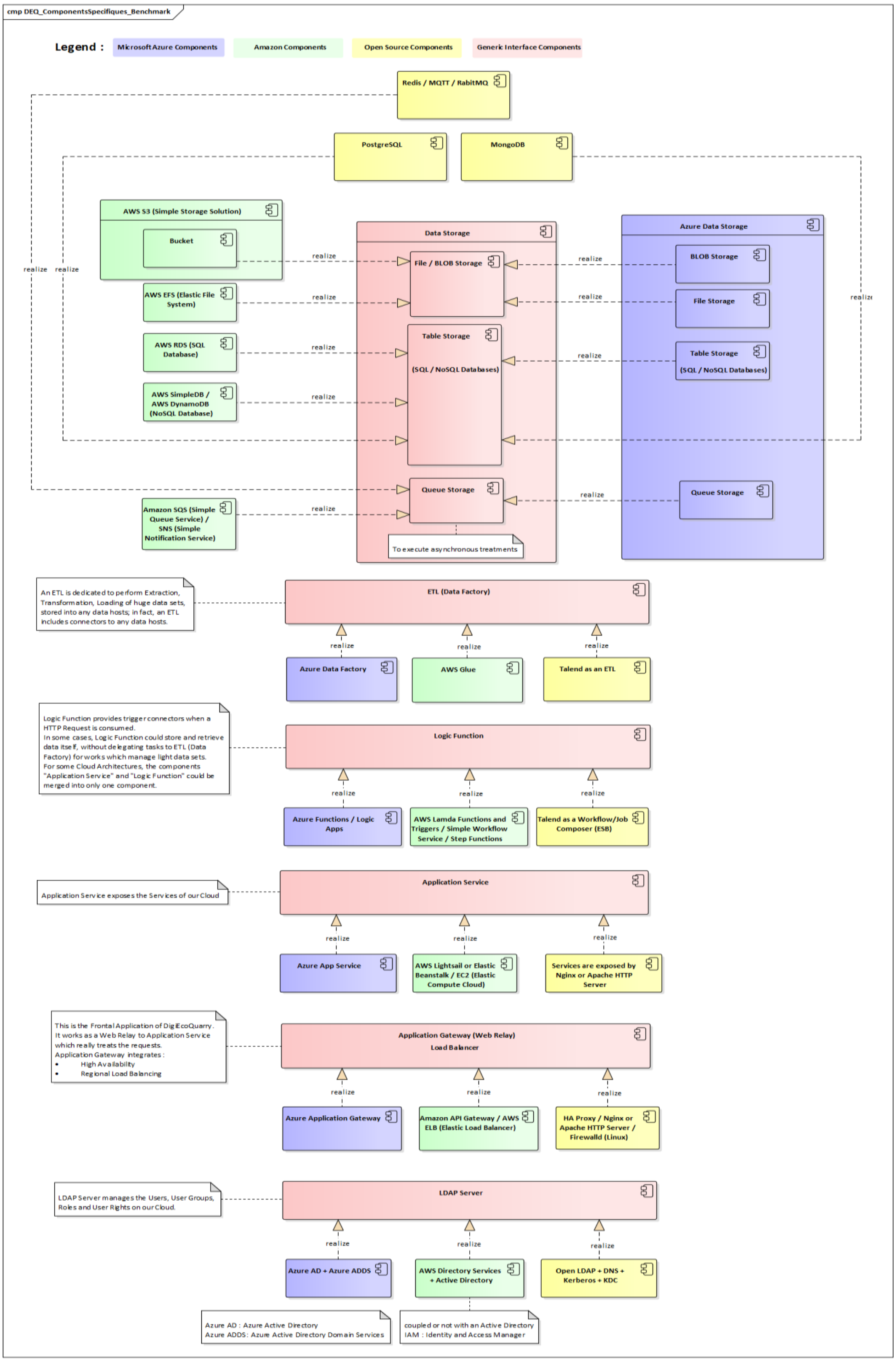
7.1.2 Data Lake components comparison

Obviously, not all datacenters charge the same prices… To simplify the reading, a “generic” datacenter has been chosen (France Central for Azure / Europe Paris for AWS), which more or less respects not only the average price of datacenters, but also the range of deployable components.

Anyway, except for small outlying data centers, the price and the features do not vary that much from one European data center to another (i.e., two datacenters hosted in the same zone)

For each Pilot site, a dedicated Data Lake will be implemented. The selection of a specific zone where is deployed the Data Lake, is very important and should take into consideration the availability of other services/components envisioned in the overall solution. The co-location of services deployed in the same zone will significantly reduce the costs of inter-zone data transfers. The closest datacenter from the Pilot sites’ location will be selected if they provide all the guarantees of proper functioning. For example, “Spain Central/Madrid” and “Italy North/Milan” should come soon as regions for Azure.

7.1.2.1 Overview: components presentation to be studied


cmp DEQ_ComponentsSpecifiques_Benchmark
Legend :
Microsoft Azure Components
Amazon Components
Open Source Components
Generic Interface Components
Redis / MQTT / RabitMQ
PostgreSQL
MongoDB
AWS S3 (Simple Storage Solution)
Bucket
AWS EFS (Elastic File System)
AWS RDS (SQL Database)
AWS SimpleDB / AWS DynamoDB (NoSQL Database)
Amazon SQS (Simple Queue Service) / SNS (Simple Notification Service)
Data Storage
File / BLOB Storage
Table Storage
(SQL / NoSQL Databases)
Queue Storage
Azure Data Storage
BLOB Storage
File Storage
Table Storage
(SQL / NoSQL Databases)
Queue Storage
realize
ETL (Data Factory)
Azure Data Factory
AWS Glue
Talend as an ETL
Logic Function
Azure Functions / Logic Apps
AWS Lamda Functions and Triggers / Simple Workflow Service / Step Functions
Talend as a Workflow/Job Composer (ESB)
Application Service
Azure App Service
AWS Lightsail or Elastic Beanstalk / EC2 (Elastic Compute Cloud)
Services are exposed by Nginx or Apache HTTP Server
Application Gateway (Web Relay)
Load Balancer
Azure Application Gateway
Amazon API Gateway / AWS ELB (Elastic Load Balancer)
HA Proxy / Nginx or Apache HTTP Server / Firewalld (Linux)
LDAP Server
Azure AD + Azure ADDS
AWS Directory Services + Active Directory
Open LDAP + DNS + Kerberos + KDC

To execute asynchronous treatments

An ETL is dedicated to perform Extraction, Transformation, Loading of huge data sets, stored into any data hosts; in fact, an ETL includes connectors to any data hosts.

Logic Function provides trigger connectors when a HTTP Request is consumed.
In some cases, Logic Function could store and retrieve data itself, without delegating tasks to ETL (Data Factory) for works which manage light data sets.
For some Cloud Architectures, the components "Application Service" and "Logic Function" could be merged into only one component.

Application Service exposes the Services of our Cloud

This is the Frontal Application of DigiEcoQuarry.
It works as a Web Relay to Application Service which really treats the requests.
Application Gateway integrates :
- High Availabilty
- Regional Load Balancing

LDAP Server manages the Users, User Groups, Roles and User Rights on our Cloud.

Azure AD : Azure Active Directory
Azure ADDS: Azure Active Directory Domain Services

coupled or not with an Active Directory
IAM : Identity and Access Manager

Figure 47: Data Lake components to be benchmarked

7.1.2.2 Application Gateway

7.1.2.2.1 Metrics Meaning

7.1.2.2.1.1 Azure Metrics

Azure Application Gateway price is based upon

- the amount of time that the gateway is provisioned and available
- and the amount of data processed by the application gateway

Usage	Traffic Volume Ranges	Traffic Price (€)
Weak	=< 10 To / month	0,0072 / Go
Medium	10 To / month < x =< 40 To / month	0,0063 / Go
Intensive	> 40 To / month	0,0032 / Go

Table 22: Azure Application Gateway Metrics - Price for a Traffic Volume Range

Usage	Outgoing Traffic Price (€)
Same Availability Zone	Free
Between Availability Zones	0,009 / Go
Between European Regions	0,018 / Go

Table 23: Azure Application Gateway Metrics - Price for the outgoing Traffic

A priori, given that 1 Data Lake will be implemented per Pilot site, and the Pilot sites have no valid reason to communicate with each other, the outgoing traffic should stay into the same availability zone, and should remain free.

Nevertheless, if data had to flow between availability zones or regions, it should be negligible compared to current main data. Maybe 5% of the total traffic each.

7.1.2.2.1.2 Amazon Metrics

AWS API Gateway is based upon

- the number of requests treated by the Gateway
- the amount of outgoing data from the Gateway

AWS does not charge the same price for HTTP requests and REST API requests.

Note that for AWS API Gateway, the amount of time to execute a given number of HTTP or REST API requests is not a criterion for computing the price per month. It only happens when computing Web Socket prices.

That means, except for Web Socket, AWS API Gateway generates "serverless" costs: we pay for use and not according to a rate of hours in the month.

7.1.2.2.1.2.1 HTTP Traffic

Usage	HTTP Traffic Ranges	Traffic Price (€)
Weak	=< 1 million requests / month	Free for 1st year then 1,03 / millions requests per outgoing 512 Ko*
Medium	=< 300 millions requests / month	1,03 / millions requests per outgoing 512 Ko*
Intensive	> 300 millions requests / month	0,924 / millions requests per outgoing 512 Ko*

Table 24: AWS API Gateway Metrics - Price for HTTP Traffic Range

***Important remark:** Note that only the outgoing HTTP traffic (the data which are downloaded to the emitter through the HTTP Response) is charged per range of 512 Ko.

7.1.2.2.1.2.2 API Rest Traffic

Usage	REST API Traffic Ranges	Traffic Price (€)
Weak	=< 1 million requests / month	Free for 1st year then 3,08 / million requests
Medium -	=< 333 million requests / month	3,08 / million requests
Medium +	333 million requests / month < x <= 1 milliard requests / month	2,9304 / million requests
Intensive -	1 milliard requests / month < x <= 20 milliard requests / month	2,4904 / million requests
Intensive +	> 20 milliard requests / month	1,584 / million requests

Table 25: AWS API Gateway Metrics - Price for REST API Traffic Range

7.1.2.2.1.2.3 API Web Socket

It exists another way to upload data to AWS API Gateway: the API Web Socket.

It is based upon

- the total number of messages (sent and received) per month
- and the total number of connection time per month

Important remark: Unfortunately, it is limited to 126 Ko ingoing data, and consequently, it does not fit with DEQ requirements.

7.1.2.2.2 Gateway Traffic Evaluation

7.1.2.2.2.1 Traffic Usage Range

The hypothesis for DEQ Traffic has been fixed as follow:

Usage	Traffic Hypothesis
Weak	5 To / month
Medium	25 To / month
Intensive	50 To / month

Table 26: Gateway Traffic Hypothesis

7.1.2.2.2.2 Azure Prices

A Z U R E						
Usage	Traffic Hypothesis	Traffic Volume Price	Outgoing Data			TOTAL (€)
			Same Zone (90%)	Different Availability Zones (5%)	Different European Regions (5%)	
Weak	5 To / month	36	-	250 Go x 0,009 = 2,25	250 Go x 0,018 = 4,5	**42,75**
Medium	25 To / month	180	-	11,25	22,5	**213,75**
Intensive	50 To / month	360	-	22,5	45	**427,5**

Table 27: Azure Gateway Traffic Price

7.1.2.2.2.3 AWS Prices

7.1.2.2.2.3.1 AWS Traffic Quantification

This is an investigation to reconcile Azure metrics with AWS metrics.

First, let’s estimate the constitution of the traffic according to HTTP and API REST.

A lot of files will be uploaded into – and downloaded from – Data Lake buckets: Excel and CSV files, video files, model templates, and possibly all kinds of files of all types. These upload and download process will pass through the frontal Gateway as HTTP Requests. Therefore, it can be estimated that they will be more numerous than API REST Requests.

With the hypothesis that the median volume of incoming requests is 1 Mo (i.e., 50% requests less than 1Mo, 50% more), the incoming volume in Mo equals the number of incoming requests.

As every data coming and stored into the Data Lake must be able to be restituted, it seems logical to share the global volume between 50% incoming and 50% outgoing.

Moreover, because most of the downloaded files or the outgoing flow should be heavy volumes, it can be estimated that 80% of the outgoing HTTP traffic

- exceeds 512 Ko
- and has a 5 Mo average volumetry

For the rest 20% of the outgoing HTTP traffic, the average could be 250 Ko per request. In consequence, if n Mo is transferred, an average of 4n requests might be expected.

For REST API requests, the hypothesis is also a 250 Ko average volumetry. In consequence, if n Mo is transferred, an average of 4n requests might be expected.

The table below summaries these described metrics:

	HTTP Requests			**API Rest Requests**
Traffic Estimation	65% of the global traffic			35% of the global traffic
	Incoming	**Outgoing**		
	50%	50%		
		< 512 Ko	**>= 512 Ko**	
		20%	80%	
Average Volumetry		250 Ko	5 Mo	250 Ko
Traffic in Mo **N = Global Traffic (Mo)**	N x 0,65 / 2	N x 0,65 / 2 / 5	(N x 0,65 / 2) x 4/5	N
Average Nb Requests	N x 0,65 / 2	4 x (N x 0,65 / 2 / 5)	((N x 0,65 / 2) x 4/5) / 5	4 x N

Table 28: Traffic estimation for AWS Gateway

	HTTP Requests			**API Rest Requests**
	Incoming	**Outgoing**		
		< 512 Ko	**>= 512 Ko**	
Pricing	N x 0,65 / 2 x 1,03 / 1 000 000	4 x (N x 0,65 / 2 / 5) x 1,03 / 1 000 000	((N x 0,65 / 2) x 4/5) / 5 x 1,03 / 1 000 000 x (5 / 0,5*) *512 Ko = 0,5 Mo	4 x N x 3,08 / 1 000 000

Table 29: Generic pricing according to a traffic estimation for AWS Gateway

7.1.2.2.2.3.2 AWS Traffic Prices

7.1.2.2.2.3.2.1 Weak Case: 5 To per month

Application of the grid for N = 5 To = 5 242 880 Mo

	HTTP Requests			**API Rest Requests**
Traffic Estimation	65% of the global traffic			35% of the global traffic
	Incoming	**Outgoing**		
	50%	50%		
		< 512 Ko	**>= 512 Ko**	
		20%	80%	
Average Volumetry		250 Ko	5 Mo	250 Ko
Traffic in Mo	1 703 936	340 787	1 363 149	1 835 008
Average Nb Requests	1 703 936	1 363 148	272 630	7 340 032
Price	**1,75**	**1,4**	**2,8**	**22,6**
TOTAL	**29**			

Table 30: Price for a traffic estimation of 5 To for AWS Gateway

7.1.2.2.2.3.2.2 Medium Case: 25 To per month

Application of the grid for N = 25 To = 26 214 400 Mo

	HTTP Requests			**API Rest Requests**
Traffic Estimation	65% of the global traffic			35% of the global traffic
	Incoming	**Outgoing**		
	50%	50%		
		< 512 Ko	**>= 512 Ko**	
		20%	80%	
Average Volumetry		250 Ko	5 Mo	250 Ko
Traffic in Mo	8 519 680	1 703 936	6 815 744	9 175 040
Average Nb Requests	8 519 680	6 815 744	1 363 149	36 700 160
Price	**8,75**	**7**	**14**	**113**
TOTAL	**143**			

Table 31: Price for a traffic estimation of 25 To for AWS Gateway

7.1.2.2.2.3.2.3 Intensive Case: 50 To per month

Application of the grid for N = 50 To = 52 428 800 Mo

	HTTP Requests			**API Rest Requests**
Traffic Estimation	65% of the global traffic			35% of the global traffic
	Incoming	**Outgoing**		
	50%	50%		
		< 512 Ko	**>= 512 Ko**	
		20%	80%	
Average Volumetry		250 Ko	5 Mo	250 Ko
Traffic in Mo	17 039 360	3 407 872	13 631 488 Volumétrie moyenne : 5 Mo	18 350 080
Average Nb Requests	17 039 360	13 631 488	2 726 298	73 400 320
Price	**17,55**	**14**	**28**	**226**
TOTAL	**286**			

Table 32: Price for a traffic estimation of 50 To for AWS Gateway

7.1.2.2.2.3.2.4 AWS Traffic Prices Summary

Some other fees exist concerning the calculation of the data price; AWS (as Azure) charges specific tariffs for the outgoing data:

- for data going out of the current region
- for "intra-region" data going from the current availability zone to another one

The prices are the same for Azure and AWS.

A W S						
Usage	Traffic Hypothesis	Traffic Volume Price	Outgoing Data			TOTAL (€)
			Same Zone (90%)	Different Availability Zones (5% = 250 Go)	Different European Regions (5% = 250 Go)	
Weak	5 To / month	**29**	-	2,25	4,5	**35,75**
Medium	25 To / month	**143**	-	11,25	22,5	**176,75**
Intensive	50 To / month	**286**	-	22,5	45	**353,5**

Table 33: Traffic (Data Treatment) Prices Summary for AWS Gateway

7.1.2.2.3 AZURE: Gateway Availability and Costs Aggregation

The pricing is given per month per 1 instance. The currency is in euros.

The price is based upon the amount of time during the gateway is available.

The results are intersected with the data treatment billing, based upon the data volumetry (see Gateway Traffic Evaluation – Azure Prices)

7.1.2.2.3.1 7/7 – 24/24

Usage	A Z U R E				
	Application Gateway / Load Balancer	WAF	**Gateway Availability**	**Data Treatment**	**TOTAL (€)**
Weak (5 To/month)	20	N/A	**20**	**43**	**63**
Medium (25 To/month)	56	101	**157**	**214**	**371**
Intensive (50 To/month)	258	361	**619**	**428**	**1047**

Table 34: Price for Azure Gateway 7/7 - 24/24

7.1.2.2.3.2 5/7 – 24/24

Usage	AZURE				
	Application Gateway / Load Balancer	WAF	**Gateway Availability**	**Data Treatment**	TOTAL (€)
Weak (5 To/month)	16	N/A	**16**	**43**	59
Medium (25 To/month)	45	81	**126**	**214**	340
Intensive (50 To/month)	206	289	**495**	**428**	923

Table 35: Price for Azure Gateway 5/7 - 24/24

7.1.2.2.3.3 5/7 – 15/24 (from 5h00 to 20h00)

Usage	AZURE				
	Application Gateway / Load Balancer	WAF	**Gateway Availability**	**Data Treatment**	TOTAL (€)
Weak (5 To/month)	10	N/A	**10**	**43**	53
Medium (25 To/month)	28	51	**79**	**214**	293
Intensive (50 To/month)	129	181	**310**	**428**	738

Table 36: Price for Azure Gateway 5/7 - 15/24 (from 5h00 to 20h00)

7.1.2.2.4 AWS: Firewall and Load Balancing Settings / Costs Aggregation

7.1.2.2.4.1 AWS WAF Metrics

AWS WAF Metrics per month
0,54 € per million of requests
0,9 € per rule

Table 37: AWS WAF Metrics

It might be necessary to write 10 rules max over the Firewall: 9€ for 10 rules.

7.1.2.2.4.2 AWS Load Balancer Metrics

To be entirely compliant with Azure Solution, an AWS Load Balancer must be costed. The load balancer will be charged according to an amount of Go per hour. The metric is given for:

- 1 Load Balancer with 10 rules per request
- routing to AWS EC2 (Elastic Compute Cloud) components, that gives cheaper prices than other AWS components (EC2 components are computing platforms – Vitual Machines – on which can be installed Open-Source components as Talend, PostgreSQL database, etc.)

Formula to be applied: (Nb Go / hour) x 0,0084 USD x 0,9 € x (Nb hours used per month)

Usage	**AWS** For routing to EC2 components
7/7 24/24	(Nb Go / hour) x 0,0084 x 0,9 € x 730 hours
5/7 24/24	(Nb Go / hour) x 0,0084 x 0,9 € x 530 hours
5/7 15/24 (from 5h00 to 20h00)	(Nb Go / hour) x 0,0084 x 0,9 € x 330 hours

Table 38: AWS Load Balancer Metrics

On average, it exists 30,416… days per months, so 1 hour is 0,00136986 month.

Usage	**Volumetry (To/month)**	**Volumetry (Go/hour)**	**Volumetry (Go/hour)**
Weak	5	5 x 1024 x 0,00136986 = 7,0136832	5 x 1024 x 0,00136986 = 7,0136832
Medium	25	25 x 1024 x 0,00136986 = 35,068416	25 x 1024 x 0,00136986 = 35,068416
Intensive	50	50 x 1024 x 0,00136986 = 70,136832	50 x 1024 x 0,00136986 = 70,136832

Table 39: AWS Traffic To/month - Go/hour

7.1.2.2.4.3 AWS Costs Summary: 7/7 – 24/24

Usage	**Volumetry (To/month)**	**Volumetry (Go/hour)**	**Nb Requests (millions/month)**	**AWS Price (€)**		
				AWS API Gateway	Load Balancer	WAF
Weak	5	7,0136832	11	36	38,7	6 + 9 = 15
Medium	25	35,068416	54	177	193,5	29 + 9 = 38
Intensive	50	70,136832	107	354	387	58 + 9 = 67

Table 40: Price of each component of AWS Gateway "7/7 – 24/24" availability

Usage	AWS Price (€)					
	AWS API Gateway	Load Balancer	WAF	**Total** **API Gateway + WAF**	**Total** **Load Balancer + WAF**	**Total** **API Gateway + Load Balancer + WAF**
Weak	36	38,7	15	**51**	**54**	**90**
Medium	177	193,5	38	**215**	**232**	**409**
Intensive	354	387	67	**421**	**454**	**808**

Table 41: AWS Gateway "7/7 – 24/24" availability total price

7.1.2.2.4.4 AWS Costs Summary: 5/7 – 24/24

Usage	Volumetry (To/month)	Volumetry (Go/hour)	Nb Requests (millions/month)	AWS Price (€)		
				AWS API Gateway	Load Balancer	WAF
Weak	5	7,0136832	11	36	28	6 + 9 = 15
Medium	25	35,068416	54	177	140	29 + 9 = 38
Intensive	50	70,136832	107	354	281	58 + 9 = 67

Table 42: Price of each component of AWS Gateway "5/7 – 24/24" availability

Usage	AWS Price (€)					
	AWS API Gateway	Load Balancer	WAF	**Total** **API Gateway + WAF**	**Total** **Load Balancer + WAF**	**Total** **API Gateway + Load Balancer + WAF**
Weak	36	28	15	**51**	**43**	**79**
Medium	177	140	38	**215**	**178**	**355**
Intensive	354	281	67	**421**	**348**	**702**

Table 43: AWS Gateway "5/7 – 24/24" availability total price

7.1.2.2.4.5 AWS Costs Summary: 5/7 – 15/24 (from 5h00 to 20h00)

Usage	Volumetry (To/month)	Volumetry (Go/hour)	Nb Requests (millions/month)	AWS Price (€)		
				AWS API Gateway	Load Balancer	WAF
Weak	5	7,0136832	11	36	17,5	6 + 9 = 15
Medium	25	35,068416	54	177	87,5	29 + 9 = 38
Intensive	50	70,136832	107	354	175	58 + 9 = 67

Table 44: Price of each component of AWS Gateway "5/7 – 15/24" availability

Usage	AWS Price (€)					
	AWS API Gateway	Load Balancer	WAF	**Total** **API Gateway + WAF**	**Total** **Load Balancer + WAF**	**Total** **API Gateway + Load Balancer + WAF**
Weak	36	17,5	15	**51**	**33**	**69**
Medium	177	87,5	38	**215**	**136**	**313**
Intensive	354	175	67	**421**	**242**	**596**

Table 45: AWS Gateway "5/7 – 15/24" availability total price

7.1.2.2.5 Open Source

Note: With a Linux Distribution installed over the VM that hosts the frontal Gateway, Firewalld can be used, which is easy to use and efficient. Example of an appropriate Linux Distribution: CentOS 7 and >.

7.1.2.2.5.1 VM Metrics

7.1.2.2.5.1.1 Generic VM Metrics

Metrics are based upon:

- CPU (Core)
- Processor
- RAM
- Bandwidth
- Storage capacity (managed disk and temporary storage)
- Number of hours of use per month
- Some other criteria to be defined…

Habitually, for frontal Gateways, VM are chosen for a generic usage or an optimized computing usage.

The billing is performed in different ways:

- pay as you go
- reserved VM instances for some years (and "save money" …)

7.1.2.2.5.1.2 Azure VM Metrics

AZURE – Ddsv5 Series (for a Gateway general usage)						
Usage	**CPU**	**Processor**	**RAM (Go)**	**Temp Storage Capacity (Go)**	**3-year reserved Price (€)**	**Disk Capacity (To)**
Weak	8	Intel® Xeon® Platinum 8370C Until 3,5 GHz	32	300	0,1655 / hour	1 = 76 €
Medium	16		64	600	0,331 / hour	1 = 76 €
Intensive	32		128	1 200	0,662 / hour	1 = 76 €

Table 46: Azure VM Metrics for an Open-Source Gateway

And these criteria must be coupled (intersected) with the number of hours of use par month.

Usage	Hours per month
7/7 24/24	730
5/7 24/24	530
5/7 15/24 (from 5h00 to 20h00)	330

Table 47: Usage hours / month for an Open-Source Gateway on Azure

Notes:

- Azure also charges for storage transactions, but for a VM hosting a Gateway, it does not make sense: the Gateway will delegate treatments to components dedicated to computation, transformation, and storage (as ETL, for example), and will not perform by itself any storage tasks.
- It is assumed that data will be transferred inside the same availability zone; for possible additional fees due to data exchange out of the current availability zone, refer to "Outgoing data prices" in paragraph "Azure Prices" of "Gateway Traffic Evaluation".

7.1.2.2.5.1.3 AWS VM Metrics

AWS EC2 VM will come with an SSD Disk. For AWS, it is charged as an Elactic Block Storage (ESB).

EBS Price: 0,1044 € / Go ==> 1 To = 1024 Go = 107 €

A W S – m6g instance type (for a Gateway general usage)						
Usage	**CPU**	**Processor**	**RAM (Go)**	**Network Bandwidth**	**3-year reserved Price (€)**	**Disk Capacity (To)**
Weak	8	Custom-built AWS Graviton2 processor with 64-bit Arm Neoverse cores	32	<= 10 Gbit/s	0,1593 / hour	1 = 107 €
Medium	16		64	<= 10 Gbit/s	0,3186 / hour	1 = 107 €
Intensive	32		128	10 Gbit/s	0,6381 / hour	1 = 107 €

Table 48: AWS VM Metrics for an Open-Source Gateway

And these criteria must be coupled (intersected) with the number of hours of use par month.

Usage	Hours per month
7/7 24/24	730
5/7 24/24	530
5/7 15/24 (from 5h00 to 20h00)	330

Table 49: VM Usage - Hours / month

For AWS EC2, the outgoing data is charged as for Azure: specific tariffs

- for data going out of the current region
- for "intra-region" data going from the current availability zone to another one

The prices are the same for Azure and AWS.

7.1.2.2.5.2 7/7 – 24/24

Alternative Open-Source Components deployed over VM in the Clouds.

Usage	HA Proxy / Nginx or Apache Server / Firewall						
	Deployment on VM over **Azure**			Deployment on VM over **Amazon**			Additional Development
	Process	SSD Disk	Total	Process	SSD Disk	Total	
Weak	121	76	**197**	116	107	**223**	
Medium	242		**318**	233		**340**	
Intensive	484		**560**	466		**573**	

Table 50: Open-Source Gateway "7/7 - 24/24" price

7.1.2.2.5.3 5/7 – 24/24

Alternative Open-Source Components deployed over VM in the Clouds.

Usage	HA Proxy / Nginx or Apache Server / Firewall						
	Deployment on VM over **Azure**			Deployment on VM over **Amazon**			Additional Development
	Process	SSD Disk	Total	Process	SSD Disk	Total	
Weak	88	76	**164**	84	107	**191**	
Medium	176		**252**	169		**276**	
Intensive	352		**428**	338		**445**	

Table 51: Open-Source Gateway "5/7 - 24/24" price

7.1.2.2.5.4 5/7 – 15/24 (from 5h00 to 20h00)

Alternative Open-Source Components deployed over VM in the Clouds.

Usage	HA Proxy / Nginx or Apache Server / Firewall						
	Deployment on VM over **Azure**			Deployment on VM over **Amazon**			Additional Development
	Process	SSD Disk	Total	Process	SSD Disk	Total	
Weak	55	76	**131**	53	107	**160**	
Medium	110		**186**	105		**212**	
Intensive	220		**296**	211		**318**	

Table 52: Open-Source Gateway "5/7 - 15/24 (from 5h00 to 20h00)" price

7.1.2.2.6 All costs summary: Azure, AWS, Open-Source comparison

7.1.2.2.6.1 7/7 – 24/24

Usage	**Volumetry (To/month)**	**Azure**	**Amazon**	**Open Source**	
				VM on Azure	**VM on AWS**
Weak	5	63	90	197	223

Medium	25	371	409	318	340
Intensive	50	1047	808	560	573

Table 53: Azure, AWS, Open-Source Gateway "7/7 - 24/24" price comparison

7.1.2.2.6.2 5/7 – 24/24

Usage	**Volumetry (To/month)**	**Azure**	**Amazon**	**Open Source**	
				VM on Azure	**VM on AWS**
Weak	5	59	79	164	191
Medium	25	340	355	252	276
Intensive	50	923	702	428	445

Table 54: Azure, AWS, Open-Source Gateway "5/7 - 24/24" price comparison

7.1.2.2.6.3 5/7 – 15/24 (from 5h00 to 20h00)

Usage	**Volumetry (To/month)**	**Azure**	**Amazon**	**Open Source**	
				VM on Azure	**VM on AWS**
Weak	5	53	69	131	160
Medium	25	293	313	186	212
Intensive	50	738	596	296	318

Table 55: Azure, AWS, Open-Source Gateway "5/7 - 24/24 (from 5h00 to 20h00)" price comparison

7.1.2.2.7 Conclusion

This in-depth review of the frontal Gateway implementation on a Cloud shows benefits of using native components and services provided by the Clouds in case these services are only little used. These results reinforce the evidence of the facts that the Clouds charge in the way "pay as you go" – even if the services can be reserved for a few years. However, a technical question can be asked: is it really wise (advised) to use the components provided by the Clouds for small volumes and restricted uses, when they give their full potential – they were designed – for intensive usage and large volumes? In fact, it all depends on whether we favour the purely technical aspect or the price aspect... and of course, the best value for money (the best technology for the best price) will be chosen.

Conversely, the results show that it is advantageous to implement an open-source solution for an intensive usage. This comes from that, when you deploy open-source components over a VM of the Cloud, you pay almost exclusively for the quality (the characteristics) of the VM, and not for any Cloud processed services since the open-source components support them. Whether the open-source service is heavily or lightly used by a large number or a small number of data, the price is almost constant and is almost entirely contained in the rental of the chosen VM. The same question comes, as for Cloud-native components: why use an open-source solution for large volumes and intensive use, when there is a Cloud-native solution specially built for it? And the answer is the same: the best value for money must be chosen.

To extend the subject to all the components and services exposed throughout this study, the subtlety of the choice will be made in the positioning of the cursor between the open-source components and the native Cloud components. A combination of both will be necessary and a financial and technical balance will have to be found. Let's not forget that the knowledge (to come, but still very vague at this stage) of the volumetry and the uses will clarify the point and help in the final choice of the components.

This first conclusion on the Gateway could be generalized to all this benchmark study.

7.1.2.3 Application Service

This component allows to create and deploy applications or any API behind a Gateway.

Typically, for DEQ, it would expose REST API or API to upload and download files into or from a BLOB storage. However, the Gateway can provide these API expositions before delegating the treatments directly to an ETL. Moreover, as the Pilot Sites volumetry is not established yet, it is not necessary to dedicate a VM (or a cluster) and a specific component to perform a task which can be managed, for less, by the Gateway.

It is the reason why this component will not be explored further: no cost estimation will be given for it.

7.1.2.4 Logic Function

This component can be very useful to orchestrate workflows, generated with a tool that avoids code, including some logics as loop, parallel runs, conditions, and that must run as distributed applications on the Cloud.

In other terms, it can be presented as an Enterprise Server Bus connecting to any components of the Cloud and launching more or less complex jobs (over a cluster of compute VM, for example).

For DEQ project, this function can be assured by an ETL, as Talend Open Studio Enterprise Server Bus (TOS ESB) or even Azure Data Factory which can perform quite the same tasks. Anyway, it is not easy nowadays to separate the ETL features from the ESB notions: the differences have been erased as they evolved, and their functionalities ended up merging. These products have converged to become one.

It is the reason why this component will not be explored further: no cost estimation will be given for it.

7.1.2.5 ETL Tools

ETL are tools that Extract, Transform, Load large volumes of data, moving data from one location (e.g., data contained into Excel files from a directory) to another location (e.g. a relational database); and in the meantime, the tool processes the data (e.g. controls and transforms them in goals to be inserted into a database).

The most famous and most used tool as a freeware, is Talend.

Azure offers a complete ETL solution in the Cloud: the Data Factory.

Every ETL tool works the same way:

- building a workflow (called pipeline),
- including elementary operations (more or less elementary activities…),
- that consume and produce data (from / to a linked service, accessed through dedicated connectors)

That can be designed according to this schema:

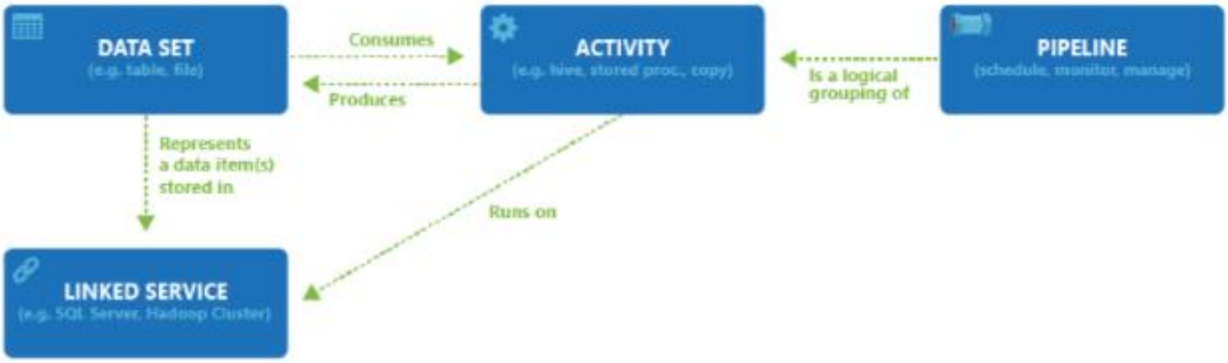

DATA SET
(e.g. table, file)
Consumes
Produces
ACTIVITY
(e.g. hive, stored proc., copy)
Is a logical grouping of
PIPELINE
(schedule, monitor, manage)
Represents a data item(s) stored in
Runs on
LINKED SERVICE
(e.g. SQL Server, Hadoop Cluster)

Figure 48: Azure Data Factory (ADF) design schema

Obviously, the billing of Azure Data Factory is based over these elements that constitute the aim of an ETL:

- pipeline execution and orchestration
- running and debugging the dataflow (volume of data & time of processing data)
- number of operations implemented in the pipeline, including creating and monitoring pipelines

7.1.2.5.1 Metrics Meaning for Azure Data Factory

7.1.2.5.1.1 Azure Pricing Metrics

7.1.2.5.1.1.1 Pipeline Orchestration and Execution

Here are the prices for features related to the pipeline orchestration and the execution in Azure:

- Orchestration refers to activity runs, trigger executions and debug runs.
- Data movement Activity: use of the copy activity to egress data out of an Azure datacenter will incur additional network bandwith charges, which will show up as a separate outbound data transfer line item on the bill.
- Pipeline activities execute on integration runtime. They include Lookup, Get Metadata, Delete and schema operations during authoring (test connection, browse folder list and table list, get schema and preview data)
- External pipeline activities are managed on integration runtime but execute on linked services. External activities include Databricks, stored procedure, HDInsight activities and many more.

Type	Azure Integration Runtime Price	Azure Managed VNET Integration Runtime Price	Self-Hosted Integration Runtime Prime
Orchestration	0,900€ per 1000 runs	0,900€ per 1000 runs	1,350€ per 1000 runs
Data movement Activity	0,225€/DIU-hour	0,225€/DIU-hour	0,090€/hour
Pipeline Activity	0,005€/hour	0,900€/hour (up to 50 concurrent pipeline activities)	0,001800€/hour
External Pipeline Activity	0,000225€/hour	0,900€/hour (up to 800 concurrent pipeline activities)	0,000090€/hour

Table 56: ADF Pipeline Orchestration tariff

For DEQ, the most appropriate use, which best fits to DEQ activities, should be “Azure Integration Runtime”:

- it avoids the teams to manage any VM, what must be done with the self-hosted IR
- Data Factory Pipelines do not require to connect to on-premises local networks (some Pilot Sites forbid it) or to use SSIS (SQL Server Integration Services) to migrate data from a private network (e.g., from on-premises, with SSIS implemented) to the Cloud

7.1.2.5.1.1.2 Data Flow Execution and Debugging

Here are the prices for features related to the Data Flow execution and the Debugging in Azure:

Type	Price	One Year Reserved (% savings)	Three Year Reserved (% savings)
General Purpose	0,259€ per vCore-hour	0,195€ per vCore-hour (~25% savings)	0,169€ per vCore-hour (~35% savings)
Memory Optimized	0,331€ per vCore-hour	0,248€ per vCore-hour (~25% savings)	0,215€ per vCore-hour (~35% savings)

Table 57: ADF Data Flow Execution tariff

Note that Data Factory Data Flows will also bill for the managed disk and blob storage required for Data Flow execution and debugging.

Azure provides a minimum of 8 vCores cluster (1 CPU / Core) to run Data Factory. With these characteristics, the general purpose should be enough for DEQ project.

With reserving a 3-year execution cluster, 35% of the price can be saved.

7.1.2.5.1.1.3 Data Factory Operations

Here are the prices for features related to the Data Factory Operations in Azure:

- Read/Write operations for Azure Data Factory entities include create, read, update, and delete. Entities include datasets, linked services, pipelines, integration runtime and triggers.
- Monitoring operations include get and list for pipeline, activity, trigger, and debug runs.

Type	Price	Examples
Read/Write	0,450€ per 50000 modified/referenced entities	Read/Write of entities in Azure Data Factory
Monitoring	0,450€ per 50000 run records retrieved	Monitoring of pipeline, activity, trigger and debug runs.

Table 58: ADF Pipeline Operations tarification

7.1.2.5.1.2 Qualitative Metrics: Pipelines

This metrics must characterize the Data Factory pipelines against a standard grid (Simple, Medium, Complex).

7.1.2.5.1.2.1 Pipeline Operations

First, enumerate the main operations used by a Data Factory pipeline, operations that will be charged by Azure billing:

Azure Reference Charging	Operations	Comment
Data Factory Operations	Create Linked Service	Linked Service are connectors that establish the connection to the data location to be extracted from, and to data location to be loaded into
Data Factory Operations	Create Datasets	Datasets are the data that are consumed and produced by the pipeline's activities
Data Factory Operations	Create Pipeline	The main pipeline that contains the activities to be processed
Data Factory Operations	Get Pipeline	To run the pipeline, Data Factory must first get it at each run… It could run on multiple instances.
Pipeline Execution and Orchestration	Run Pipeline	Pipeline execution
Pipeline Execution and Orchestration	Copy Data execution time	This item is accounted according to a Data Integration Unit (DIU) base. To copy any data from a pipeline processed by Data Factory, Azure Integration Runtime uses 4 DIU by default. So, if the estimated time is n minutes, 4n is the amount to apply for the billing.
Data Factory Operations	Monitor Pipeline run	Data Factory monitors the pipeline that it executes
Pipeline Execution and Orchestration	Other activities execution time	The time spent executing any external services called by the pipeline processed by Data Factory

Table 59: Pipeline Operations description

7.1.2.5.1.2.2 Pipeline Qualification

Now, let's intuit the amount of operation types that a Simple / Medium / Complex pipeline could use:

<table>
<tr><th colspan="2" rowspan="2">Operations</th><th colspan="3">Pipeline Qualification
(The displayed numbers are an average of each type of qualification)</th></tr>
<tr><th>Simple</th><th>Medium</th><th>Complex</th></tr>
<tr><td colspan="2">Create Linked Service</td><td>2</td><td>4</td><td>8</td></tr>
<tr><td colspan="2"></td><td colspan="3">*In a Simple pipeline, data are extracted from 1 location and loaded to 1 location. In a Medium or Complex pipeline, the data locations are multiple and can be from different types (SQL, NoSQL, WareHouses, Directories…).</td></tr>
<tr><td rowspan="2">Create Datasets</td><td>Nb Datasets</td><td>4</td><td>16</td><td>32</td></tr>
<tr><td>Nb Activities</td><td>1</td><td>6</td><td>12</td></tr>
<tr><td colspan="2"></td><td colspan="3">*Nb read/write entities (= Nb Datasets):
• 1 Dataset per linked service (called Dataset reference)
• 1 Input Dataset & 1 Output Dataset per activity</td></tr>
</table>

Operations		Pipeline Qualification (The displayed numbers are an average of each type of qualification)		
		Simple	Medium	Complex
Create Pipeline		3	5	9
		*At least, for the simplest pipeline, 3 read/write entities: • 1 for pipeline creation • 2 for Dataset references (mapped to 2 linked services)		
Get Pipeline		1	1	1
		*Only 1 instance		
Run Pipeline		2	2	2
		*1, of course, to execute the pipeline activities + 1 eventually to trigger the pipeline		
Execute Activity	Execution time	0,5 mn	3 mn	6 mn
	Nb Activities	1	6	12
		*30s / Activity		
Copy Data execution time		1 mn x 4 DIU	5 mn x 4 DIU	10 mn x 4 DIU
		*DIU: see table above		
Monitor Pipeline run		2	5	13
		*1 for pipeline monitoring + n for monitoring each activity processed by the pipeline		
Other activities execution time		0 mn	1 mn	5 mn
		*Amount of time to execute external services called by the pipeline		

Table 60: Pipeline Qualification

7.1.2.5.1.2.3 Pipeline price per Qualification

Operations			**Pipeline Qualification**		
			Simple	**Medium**	**Complex**
Operation Data Factory	**Read / Write**	Nb Operations	10	26	50
		Formula	N x 0,45 / 50 000		
		Price (€)	**0,00009**	**0,000234**	**0,00045**
	Monitoring	Nb Executions	2	5	13
		Formula	N x 0,225 / 50 000		
		Price (€)	**0,000009**	**0,0000225**	**0,0000585**
	Activity Runs	Nb Runs	2	2	2

Operations			Pipeline Qualification		
			Simple	**Medium**	**Complex**
Pipeline Orchestration **/ Execution**		Formula	N x 0,9 / 1000		
		Price (€)	**0,0018**	**0,0018**	**0,0018**
	Data Movement Activities	Execution Time	1 mn	5 mn	10 mn
		Formula	(N[minutes]/60) x 4 x 0,225		
		Price (€)	**0,015**	**0,075**	**0,15**
	Pipeline Activity **(30s / Activity)**	Execution Time	0,5 mn	3 mn	6 mn
		Formula	(N[minutes]/60) x 0,005		
		Price (€)	**0,0000416**	**0,00025**	**0,0005**
	External Pipeline Activity	Execution Time	0 mn	1 mn	5 mn
		Formula	(N[minutes]/60) x 0,000225		
		Price (€)	**0**	**0,00000375**	**0,00001875**
TOTAL		Execution Time	1,5 mn	9 mn	21 mn
		Price (€)	**0,0169406**	**0,07731025**	**0,15282725**

Table 61: Price per Pipeline type

As the table shows, of all the items whose billing has been costed, one of them is first in front of all the others and takes almost all the charge: it is the execution time of the copy of the data

- from the data source into the pipeline
- from the pipeline to the data target

	% of the cost according to the pipeline type		
	Simple	**Medium**	**Complex**
Data Movement Activities	88%	97%	98%

Table 62: Ratio of the cost of data copy activities into the pipeline global cost

In other terms, the volume of the input and output data to be processed by the pipeline, determines the final price of the pipeline execution. All the rest of the activities is negligible.

Moreover, if the pipeline processes a large volume of input and output data, its execution time increases and therefore, the VM that supports its execution is more used. However, the use of "VM" resources must also be included into the price (see “Data Flow Execution and Debugging” metrics tariff, above): this is discussed in the next chapters.

7.1.2.5.1.3 Qualitative Metrics: Pilot Site activities

This paragraph establishes a metrics of use of the Data Factory Pipelines for a generic Pilot Site.

Usage	% of pipelines type		
	Simple	Medium	Complex
Weak	80%	15%	5%
Medium	65%	25%	10%
Intensive	50%	30%	20%

Table 63: Pilelines type usage determination

7.1.2.5.2 Evaluation Price for Azure Data Factory

Although the efforts to determine a quantity of pipelines processed per day by a generic Pilot Site, it is too difficult to intuit, as is, this number, related to an unknown volume of data processed by the Data Factory.

In fact, do not forget that the whole volume of data processed by the IQS (which could be, why not, found with complicated rules), do not match with the volume of data processed by the Data Factory…

Instead of giving too vague and unclear hypotheses, and false perspectives, it is preferable to give a price for 100 Data Factory Pipelines, which is a simple and understandable base, easy to extrapolate.

7.1.2.5.2.1 Quantitative Pipeline Metrics

Usage	Price (€) for 100-base pipelines								
	per day							per month	
	Simple		Medium		Complex		Total	7/7 (30 days)	5/7 (22 days)
	Nb	Price	Nb	Price	Nb	Price			
Weak	80	1,355	15	1,16	5	0,764	3,28	**98**	**72**
Medium	65	1,1	25	1,93	10	1,53	4,56	**137**	**100**
Intensive	50	0,85	30	2,32	20	3,06	6,23	**187**	**137**

Table 64: Price for 100-base Data Factory Pipelines

7.1.2.5.2.2 "VM support" Pricing: Execution Data Flow

The selected way to process Data Factory avoids to self-manage VM for running pipelines.

However, we have to pay for the VM that Azure uses to process the pipelines, which is charged by an hour-base per v-Core.

It has been estimated, that, for a generic Pilot Site, the data flow execution time should take:

- 3 hours / v-Core / day for a Weak usage
- 5 hours / v-Core / day for a Medium usage
- 8 hours / v-Core / day for an Intensive usage

These metrics are not given for 100-base pipelines, but for the whole Data Factory Pipeline processing of a Pilot Site.

As Azure provides a computing power with no less than 8 v-Cores (using 1 CPU per Core), it comes the following table:

Usage	Price (€) for Execution and Debugging Data Flow			
	per day / 1 v-Core		per month / 8 v-Cores	
	Nb Hours	Price	7/7 (30 days)	5/7 (22 days)
Weak	3	0,507	15,21 x 8 = **122**	11,154 x 8 = **89**
Medium	5	0,845	25,35 x 8 = **203**	18,6 x 8 = **149**
Intensive	8	1,352	40,56 x 8 = **324**	30 x 8 = **240**

Table 65: Price for Execution and Debugging Data Flow

7.1.2.5.2.3 Number of pipelines: an estimation using the execution time

If it is not easy to estimate the data flow volume and the number of pipelines required to process it, it can be deducted from the processing (data flow) total execution time, which is a solid evaluation, based upon real projects.

Reminder the average evaluation:

		Pipeline Qualification		
		Simple	Medium	Complex
Execution time		1,5 mn	9 mn	21 mn
Usage	Weak	80%	15%	5%
	Medium	65%	25%	10%
	Intensive	50%	30%	20%

Table 66: Average Execution Time per pipeline types

Usage	Processing total time per day for 8 v-Cores	Formula (where n = Nb of pipelines processed per day)		
Weak	3h x 8 = 24h = **1440mn**	(1,5 x 80/100 x n) + (9 x 15/100 x n) + (21 x 5/100 x n) = 1440 Rounded **n = 400**		
		Simple	**Medium**	**Complex**
		320	**60**	**20**
Medium	5h x 8 = 40h = **2400mn**	(1,5 x 65/100 x n) + (9 x 25/100 x n) + (21 x 10/100 x n) = 2400 Rounded **n = 450**		
		Simple	**Medium**	**Complex**
		292	**113**	**45**
Intensive	8h x 8 = 64h = **3840mn**	(1,5 x 50/100 x n) + (9 x 30/100 x n) + (21 x 20/100 x n) = 3840 Rounded **n = 502**		
		Simple	**Medium**	**Complex**
		251	**151**	**100**

Table 67: Deducted Nb of pipeline types processed per day

7.1.2.5.2.4 Conclusion: cost evaluation

Usage	Price (€) for processed pipelines into Azure Data Factory												
	per day							per month					
	Simple		Medium		Complex		Total	7/7 (30 days)			5/7 (22 days)		
	Nb	Price	Nb	Price	Nb	Price		Pipeline	Execution	Total	Pipeline	Execution	Total
Weak	320	5,421	60	4,639	20	3,057	13,117	**394**	**122**	**516**	**289**	**89**	**378**
Medium	292	4,947	113	8,736	45	6,877	20,56	**617**	**203**	**820**	**452**	**149**	**601**
Intensive	251	4,252	151	11,674	100	15,283	31,209	**936**	**324**	**1260**	**687**	**240**	**927**

Table 68: Cost evaluation for processed pipelines into Azure Data Factory

7.1.2.5.3 ETL Open-Source solution

To optimize the hot computing of the logical chain, it is recommended to dedicate a VM to the Open Source ETL as Azure does it according to its way. For the beginning, the project can start with a single VM, but can be reinforcing with others if necessary.

The elected product is Talend Open Studio Enterprise Server Bus (TOS ESB). Their behaviour and handling are more flexible than Talend paid version, and the deployment over an exploitation VM, based upon Karaf (light Docker), is relatively easy.

The VM must host some software:

- Java Runtime Environment
- TOS ESB Runtime
- Karaf to deploy the Talend pipelines into the Talend Runtime
- Hawtio, a web console tool to monitor Karaf Container
- Hawtio needs a browser to be executed, as Firefox or Chrome

The price of a VM over Azure Cloud is determined through these metrics:

Price Metrics for deploying and running 1 "Talend" VM on Azure Cloud	
Items	**Comment**
The OS that is installed over the VM	Windows licences must be paid, so the choice will be done among Linux free strong-securized distribution, as: • CentOS • SE Linux • Ubuntu
The number of Cores and CPU of the VM	The selected VM must be optimized for hot computing. Instance Fsv2 Series is a good candidate. **Note that Fsv2 Series contains 2 vCPU per Core.**
VM RAM	The RAM is determined by the chosen Cores of the Fsv2 VM Instance.
The disk storage	1 To SSD should be enough for each DEQ Pilot Site.

Price Metrics for deploying and running 1 "Talend" VM on Azure Cloud	
Items	**Comment**
The number of storage transaction	With a tarification of 0,0018 € per 10 000 transactions, it seems negligible compared to the rest of the price.
The used bandwidth and the outgoing data transfert	This item is not referenced in this tariff: it is already counted with the outgoing bandwidth of the API Gateway.

Table 69: Price Metrics for deploying and running 1 "Talend" VM on Azure Cloud

Obviously, to minimize the cost, we are opting for a 3-year reserved VM Instance.

Usage	Price (€) for using 1 "Talend" VM over Azure Cloud							
	Fsv2 Series Instance				Managed Disk		Storage Transactions	Total Price
	Core	RAM	Temp Storage	Price	Characteristics	Price	Price	
Weak	8	16 Go	64 Go	96	1 To SSD	76	12	**184**
Medium	16	32 Go	128 Go	192			25	**293**
Intensive	32	64 Go	256 Go	383			50	**509**

Table 70: Price for 1 "Talend" VM on Azure Cloud

7.1.2.5.4 Costs Summary

Usage	Azure Data Factory	Talend VM over Azure Cloud
Weak	378	184
Medium	601	293
Intensive	927	509

Table 71: Azure Data Factory / Talend VM price comparison

7.1.2.6 Data Storage

7.1.2.6.1 Compare of storage on Azure and AWS

In the AWS platform, cloud storage is primarily broken down into three services:

- Simple Storage Service (S3). Basic object storage that makes data available through an Internet accessible API.
- Elastic Block Storage (EBS). Block level storage intended for access by a single VM.
- Elastic File System (EFS). File storage meant for use as shared storage for up to thousands of EC2 instances.

In Azure Storage, subscription-bound storage accounts allow you to create and manage the following storage services:

- Blob storage stores any type of text or binary data, such as a document, media file, or application installer. You can set Blob storage for private access or share contents publicly to the Internet. Blob storage serves the same purpose as both AWS S3 and EBS.
- Table storage stores structured datasets. Table storage is a NoSQL key-attribute data store that allows for rapid development and fast access to large quantities of data. Similar to AWS' SimpleDB and DynamoDB services.
- Queue storage provides messaging for workflow processing and for communication between components of cloud services.
- File storage offers shared storage for legacy applications using the standard server message block (SMB) protocol. File storage is used in a similar manner to EFS in the AWS platform.

7.1.2.6.2 BLOB Storage

Azure Blob storage is Microsoft's object storage solution for the cloud. Blob storage is optimized for storing massive amounts of unstructured data. Unstructured data is data that doesn't adhere to a particular data model or definition, such as text or binary data.

Blob storage is designed for:

- Serving images or documents directly to a browser.
- Storing files for distributed access.
- Streaming video and audio.
- Writing to log files.
- Storing data for backup and restore, disaster recovery, and archiving.
- Storing data for analysis by an on-premises or Azure-hosted service.

Users or client applications can access objects in Blob storage via HTTP/HTTPS, from anywhere in the world. Objects in Blob storage are accessible via the Azure Storage REST API, Azure PowerShell, Azure CLI, or an Azure Storage client library. Client libraries are available for different languages: .NET, Java, Node.js, Python, Go, PHP, Ruby

Blob storage now supports the SSH File Transfer Protocol (SFTP). This support provides the ability to securely connect to Blob Storage accounts via an SFTP endpoint, allowing you to leverage SFTP for file access, file transfer, as well as file management.

Total cost of block blob storage depends upon:

- Volume of data stored per month.
- Quantity and types of operations performed, along with any data transfer costs.
- Data redundancy option selected.

Usage	**Azure**	**Amazon**
	BLOB Storage on General purpose storage account v2	Simple Storage Solution (S3)
Weak	4,47 € / month	Storage 4,88€ / month Data transfer 4,16€ / month
Medium	44,98 € / month	54,26€ / month Data transfer 16,65€ / month
Intensive	186,50 € / month	225,96€ / month Data transfer 41.63€ / month

Table 72: Price for Blob storage

Note: *prices estimated with pay as you go option. Azure proposes options to reduce the costs, but they are clearly oriented towards big data consumers: 1-year reserved option by Azure starts at 100 To minimum, 1450 € / month, including 10^4 operations of each kind.*

Note 2: *there is no alternative Open-Source Component for BLOB Storage: BLOB must be stored over the Cloud and this action has a price according to some storage characteristics.*

Here below is described the assessment method which defines what means a weak, medium, or intensive usage (amounts / month).

Usage	**Meaning/Comment**	
	Azure	**Amazon**
Weak	Storage 200 Go, 10^4 writes, 10^6 reads	Storage 200 Go, 10^4 writes, 10^6 reads Data transfer (internet output) 50 Go
Medium	Storage 2 To, 10^6 writes, 10^7 reads	Storage 2 To, 10^6 writes, 10^7 reads Data transfer (internet output) 200 Go
Intensive	Storage 10 To, 10^6 writes, 10^7 reads	Storage 10 To, 10^6 writes, 10^7 reads Data transfer (internet output) 500 Go

Table 73: Blob storage usage assessment

Azure details page and price calculator:

https://azure.microsoft.com/en-us/pricing/details/storage/blobs/

https://azure.microsoft.com/en-us/pricing/calculator/

Amazon details page and price calculator:

https://aws.amazon.com/s3/pricing/

https://calculator.aws/#/createCalculator/S3

7.1.2.6.3 Data Lake storage

Azure Data Lake Storage Gen2 is a set of capabilities dedicated to big data analytics, built on Azure Blob Storage.

Data Lake Storage Gen2 converges the capabilities of Azure Data Lake Storage Gen1 with Azure Blob Storage. For example, Data Lake Storage Gen2 provides file system semantics, file-level security, and scale. Because these capabilities are built on Blob storage, you'll also get low-cost, tiered storage, with high availability/disaster recovery capabilities.

Because Data Lake Storage Gen2 is built on top of Azure Blob Storage, multiple concepts can describe the same, shared things.

The following are the equivalent entities, as described by different concepts. Unless specified otherwise these entities are directly synonymous:

Concept	Top Level Organization	Lower-Level Organization	Data Container
Blobs - General purpose object storage	Container	Virtual directory (SDK only - does not provide atomic manipulation)	Blob
Azure Data Lake Storage Gen2 - Analytics Storage	Container	Directory	File

Table 74: Blob storage / Data Lake equivalence

Usage	Azure	Amazon
	Data Lake Storage Gen2	Elastic Block Storage
Weak	4,60 € / month	34,94€ / month
Medium	36,53 € / month	334,90€ / month
Intensive	178,05 € / month	1637,71€ / month

Table 75: Price for Data Lake storage

Azure Data Lake solution includes Blob storage, and other possibilities (File share, Tables, Queues), at a similar price, considering the same amounts of data and storage conditions. It is our choice over simple Blob storage.

Here below is described the assessment method which defines what means a weak, medium, or intensive usage (amounts / month).

Usage	Meaning/Comment	
	Azure	**Amazon**
Weak	Storage 200 Go, 10^4 writes, 10^6 reads	Storage 200 Go, weekly snapshot, 50 Go modified each snapshot, 730hr
Medium	Storage 2 To, 10^6 writes, 10^7 reads	Storage 2 To, weekly snapshot, 200 Go modified each snapshot, 730hr
Intensive	Storage 10 To, 10^6 writes, 10^7 reads	Storage 10 To, weekly snapshot, 500 Go modified each snapshot, 730hr

Table 76: Data Lake usage assessment

Azure details page and price calculator:

https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-introduction

https://azure.microsoft.com/en-us/pricing/calculator/

Amazon details page and price calculator:

https://aws.amazon.com/ebs/

https://calculator.aws/#/createCalculator/EBS

7.1.2.6.4 File Storage

Azure Files offers fully managed file shares in the cloud that are accessible via the industry standard Server Message Block (SMB) protocol or Network File System (NFS) protocol. Azure Files file shares can be mounted concurrently by cloud or on-premises deployments. SMB Azure file shares are accessible from Windows, Linux, and macOS clients. NFS Azure Files shares are accessible from Linux or macOS clients. Additionally, SMB Azure file shares can be cached on Windows Servers with Azure File Sync for fast access near where the data is being used.

Azure file shares are deployed into storage accounts, which are top-level objects that represent a shared pool of storage. This pool of storage can be used to deploy multiple file shares, as well as other storage resources such as blob containers, queues, or tables. All storage resources that are deployed into a storage account share the limits that apply to that storage account.

There are two main types of storage accounts for Azure Files deployments:

- General purpose version 2 (GPv2) storage accounts: GPv2 storage accounts allow to deploy Azure file shares on standard/hard disk-based (HDD-based) hardware. In addition to storing Azure file shares, GPv2 storage accounts can store other storage resources such as blob containers, queues, or tables.
- FileStorage storage accounts: FileStorage storage accounts allow to deploy Azure file shares on premium/solid-state disk-based (SSD-based) hardware. FileStorage accounts can only be used to store Azure file shares; no other storage resources (blob containers, queues, tables, etc.) can be deployed in a FileStorage account. Only FileStorage accounts can deploy both SMB and NFS file shares.

Usage	**Azure**	**Amazon**	**Amazon**
	File Storage	Windows FS on HDD	Elastic File System (EFS)
Weak	13,55 € / month	6,20€ / month	16,08€ / month
Medium	147,20 € / month	149,29€ / month	164,60€ / month
Intensive	698,28 € / month	412,74€ / month	823,02€ / month

Table 77: Price for File storage

There is no alternative Open-Source Components for File Storage: the Clouds offer a way to synchronize buckets into a Cloud Storage with LAN directories; this action has a price according to some storage characteristics.

Here below is described the assessment method which defines what means a weak, medium, or intensive usage.

Usage	**Meaning/Comment**	
	Azure	**Amazon**
Weak	200 Go	200 Go, 200 Go archive
Medium	2 To, snapshot 100 Go, 10^6 writes, 10^7 reads	2 To, 1 To archive, 50 Mo/s
Intensive	10 To, 10^6 writes, 10^7 reads	10 To, 5 To archive, 100 Mo/s

Table 78: File storage usage assessment

Azure and Amazon pricing pages:

https://azure.microsoft.com/en-us/pricing/details/storage/files/

https://aws.amazon.com/efs/pricing/

7.1.2.6.5 SQL Database

Azure SQL Database is a fully managed platform as a service (PaaS) database engine that handles most of the database management functions such as upgrading, patching, backups, and monitoring without user involvement. Azure SQL Database is always running on the latest stable version of the SQL Server database engine and patched OS with 99.99% availability. PaaS capabilities that are built into Azure SQL Database enable to focus on the domain-specific database administration and optimization activities that are critical for your business.

With Azure SQL Database, one can create a highly available and high-performance data storage layer for the applications and solutions in Azure. SQL Database can be the right choice for a variety of modern cloud applications because it enables to process both relational data and non-relational structures, such as graphs, JSON, spatial, and XML.

Usage	**Azure**	**Amazon**
	Azure SQL Database	Aurora-PostgreSQL with Relational Database Service (RDS)
Weak	3,48 € / month	53,21€ / month
Medium	18,26 € / month	78,98€ / month
Intensive	104,98 € / month	92,39€ / month

Table 79: Price for SQL database

Here below is described the assessment method which defines what means a weak, medium or intensive usage.

Usage	**Meaning/Comment**	
	Azure	**Amazon**
Weak	5 Go data, serverless, 1-8 vCore, 5 Go save, 4 months, 1 year save	5 Go data, E/S 5-100, 40 h peak, serverless, 1 Aurora Capacity Unit
Medium	50 Go data, serverless, 1-8 vCore, 25 Go save, 4 months, 1 year save	50 Go data, E/S 20-500, 40 h peak, serverless, 1 Aurora Capacity Unit
Intensive	200 Go data, serverless, 1-8 vCore, 175 Go save, 4 months, 1 year save	200 Go data, E/S 5-100, 40 h peak, serverless, 1 Aurora Capacity Unit

Table 80: SQL database usage assessment

Azure and Amazon pricing pages:

https://azure.microsoft.com/en-us/pricing/details/azure-sql-database/single/

https://aws.amazon.com/fr/rds/aurora/pricing/

Alternative Open-Source Components deployed over VM in the Clouds. Cost is VM Price plus additional storage if willing to keep data on a separate secured disk.

Usage	PostgreSQL		
	Deployment on VM over **Azure**	Deployment on VM over **Amazon**	Additional Costs
Weak	97,83 € / month	65,56€ / month	VM and database management by an administrator
Medium	101,19 € / month	70,96€ / month	
Intensive	232,25 € / month	150,15€ / month	

Table 81: Price for Open-Source database solution

Here below is described the assessment method which defines what means a weak, medium, or intensive usage.

Usage	Meaning/Comment	
	Azure	**Amazon**
Weak	1 D4s v3, 1 HDD S4, 1 year	1 EC2 t4g.xlarge, 1 EBS HDD 30 Go, 1 year
Medium	1 D4s v3, 1 SSD E6, 1 year	1 EC2 t4g.xlarge, 1 EBS SSD gp2 64 Go, 1 year
Intensive	1 D8s v3, 2 SSD P10, 1 year	1 EC2 t4g.2xlarge, 1 EBS SSD gp3 256 Go, 1 year

Table 82: VM size assessment for Open-Source database deployment

7.1.2.6.6 NoSQL Database

Simple NoSQL databases

Azure Table storage is a service that stores non-relational structured data (structured NoSQL data) in the cloud, providing a key/attribute store with a schemaless design. Because Table storage is schemaless, it's easy to adapt the data as the needs of application evolve.

Amazon SimpleDB is a highly available NoSQL data store that offloads the work of database administration.

Amazon SimpleDB provides a simple web services interface to create and store multiple data sets, query your data easily, and return the results. Data model can be changed on the fly, and data are automatically indexed, making it easy to quickly find the information that you need. There is no need to pre-define a schema or change a schema if new data is added later.

Usage	**Azure**	**Amazon**
	Azure Table Storage	SimpleDB
Weak	0,29 € / month	4,66€ / month
Medium	5,38 € / month	32,98€ / month
Intensive	55,02 € / month	262,36€ / month

Table 83: Price for simple NoSQL database

Here below is described the assessment method which defines what means a weak, medium, or intensive usage.

Usage	Meaning/Comment	
	Azure Table Storage	Amazon Simple DB
Weak	5 Go, 10^6 transactions	50h CPU, 1 Go OUT, 5 Go Store
Medium	100 Go, 10^6 transactions	75h CPU, 5 Go OUT, 100 Go store
Intensive	1 To, 10^8 transactions	75h CPU, 10 Go OUT, 1 To store

Table 84: Size assessment for simple NoSQL databases

Advanced NoSQL databases

Azure Cosmos DB is a fully managed NoSQL database for modern app development. Single-digit millisecond response times, and automatic and instant scalability, guarantee speed at any scale.

Amazon DynamoDB is a fully managed, serverless, key-value NoSQL database designed to run high-performance applications at any scale. DynamoDB offers built-in security, continuous backups, automated multi-Region replication, in-memory caching, and data export tools.

Usage	Azure	Amazon
	Cosmos DB	Dynamo DB
Weak	14,54 € / month	13,80€ / month, 197,60€ initial
Medium	41,25 € / month	39,91€ / month, 197,60€ initial
Intensive	294,31 € / month	293,91€ / month, 197,60€ initial

Table 85: Price for advanced NoSQL database

Here below is described the assessment method which defines what means a weak, medium, or intensive usage.

Usage	Meaning/Comment	
	Azure Cosmos DB	Amazon Dynamo DB
Weak	5 Go, 400 UR/s, 365h/month	5 Go, 1-10 write/s, 50h / month
Medium	100 Go, 400 UR/s, 365h/month	100 Go, 5-50 writes/s, 75h / month
Intensive	1 To, 400 UR/s, 365h/month	1 To, 5-50 writes/s, 75h / month

Table 86: Size assessment for advanced NoSQL database

Open-Source NoSQL solutions

Alternative Open-Source Components deployed over VM in the Clouds. See Table 82 for assessment method.

Usage	MongoDB		
	Deployment on VM over **Azure**	Deployment on VM over **Amazon**	Additional costs
Weak	97,83 € / month	65,56€ / month	VM and database management by an administrator
Medium	101,19 € / month	70,96€ / month	
Intensive	232,25 € / month	150,15€ / month	

Table 87: VM size assessment for Open-Source database deployment

Alternative solutions from Azure Marketplace

ArangoDB is a multi-model NoSQL database that supports documents, graphs and key/values.

Usage	Azure
	ArangoDB
Weak	29,71€ / month
Medium	59,09€ / month
Intensive	118,51€ / month

Table 88: Price for alternative database

Here below is described the assessment method which defines what means a weak, medium, or intensive usage.

Usage	Meaning/Comment
	Azure Arango DB
Weak	DS1V2 1 Core, 3,5 G RAM, 7 G disk, 365h / month
Medium	DS2V2 2 cores, 7 G RAM, 14 G disk, 365h / month
Intensive	DS3V2 4 cores, 14 G RAM, 28 G disk, 365h / month

Table 89: Size assessment for alternative database

Azure and Amazon pricing pages:

https://azure.microsoft.com/en-us/pricing/details/storage/tables/

https://aws.amazon.com/fr/simpledb/pricing/

https://aws.amazon.com/dynamodb/

https://calculator.aws/#/createCalculator/DynamoDB

https://azuremarketplace.microsoft.com/en-us/marketplace/apps/arangodb.arangodb?tab=PlansAndPrice

7.1.2.7 LDAP Servers for Users Authentication/Authorization

7.1.2.7.1 Cloud's Solutions

7.1.2.7.1.1 Description and specific aspects

This resource embeds User authentication and User authorization: after a User is authenticated as being able to access a domain / service / application / resource, the User roles determine which rights are granted to the User for the accessed resource.

The best offer to manage User authentication (i.e., the most secured offer) permits to avoid passing a password in clear mode over the network. To implement a such solution, they use a dedicated protocol as Kerberos, or an own protocol provided by the editor.

Instead of installing the whole technology "by hand" as open-source components on Cloud's VMs – which is not so simple… – it exists all-in-one offers on the Clouds which embed the following components:

- DNS (Domain Server Name): to resolve domain names
- LDAP Protocol to access a LDAP Server (as Active Directory) which centralizes User information
- KDC (Key Distribution Center): to manage keys to be distributed to granted Users, using Kerberos protocol
- Kerberos (or an own protocol): to manage User authentication
- User authorisation management: based upon User roles hosted by LDAP Server

This package is titled

- ADDS (Active Directory Domain Services) on Azure
- AWS Directory Service on AWS

Over Azure, Active Directory must be deployed separately using Azure AD component.

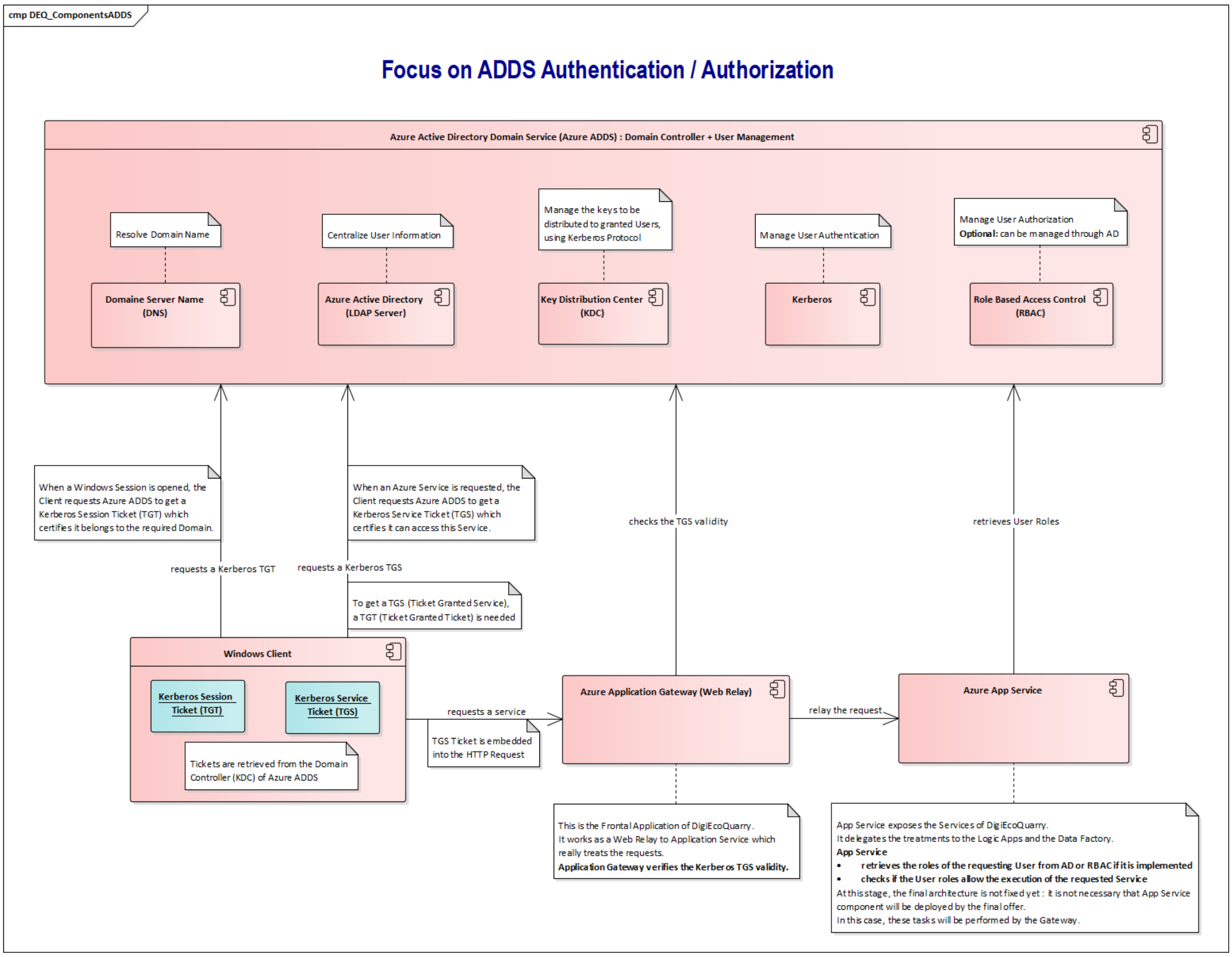
With an Azure subscription, the free version of Azure Active Directory can be used as LDAP Server on the Cloud. It includes MFA (Multi-Factor Authentication) using an authentication mobile application, while premium versions (those you pay) use advanced MFA features or additional protections, useless for Pilot Sites.

Consequently, the free version of Azure AD, added to ADDS, is complete enough and strongly secured to manage User Pilot Sites.

Note that AWS deploys the real version of AD, and obviously not the Azure AD version (reserved to Azure).

At last, it is important to mention that, to secure users access to reports produced by Power BI embedded into Azure, Azure AD should be mandatory.

7.1.2.7.1.2 *DEQ Authentication/Authorization Schema*


cmp DEQ_ComponentsADDS
Focus on ADDS Authentication / Authorization
Azure Active Directory Domain Service (Azure ADDS) : Domain Controller + User Management
Resolve Domain Name
Centralize User Information
Manage the keys to be distributed to granted Users, using Kerberos Protocol
Manage User Authentication
Manage User Authorization
Optional: can be managed through AD
Domaine Server Name (DNS)
Azure Active Directory (LDAP Server)
Key Distribution Center (KDC)
Kerberos
Role Based Access Control (RBAC)
requests a Kerberos TGT
requests a Kerberos TGS
checks the TGS validity
retrieves User Roles
Windows Client
Kerberos Session Ticket (TGT)
Kerberos Service Ticket (TGS)
requests a service
Azure Application Gateway (Web Relay)
relay the request
Azure App Service

When a Windows Session is opened, the Client requests Azure ADDS to get a Kerberos Session Ticket (TGT) which certifies it belongs to the required Domain.

When an Azure Service is requested, the Client requests Azure ADDS to get a Kerberos Service Ticket (TGS) which certifies it can access this Service.

To get a TGS (Ticket Granted Service), a TGT (Ticket Granted Ticket) is needed

Tickets are retrieved from the Domain Controller (KDC) of Azure ADDS

TGS Ticket is embedded into the HTTP Request

This is the Frontal Application of DigiEcoQuarry.
It works as a Web Relay to Application Service which really treats the requests.
Application Gateway verifies the Kerberos TGS validity.

App Service exposes the Services of DigiEcoQuarry.
It delegates the treatments to the Logic Apps and the Data Factory.
App Service

- **retrieves the roles of the requesting User from AD or RBAC if it is implemented**
- **checks if the User roles allow the execution of the requested Service**

At this stage, the final architecture is not fixed yet : it is not necessary that App Service component will be deployed by the final offer.
In this case, these tasks will be performed by the Gateway.

Figure 49: ADDS Operating Overview

7.1.2.7.2 Metrics for an Active Directory Service

This globalizing, totalizing solution is suggested and encouraged because it is the most comprehensive and the most secure on the market, especially because it embeds Kerberos technology. That's the main point that makes it put forward, and this point is assumed by both Clouds.

For the rest, Azure and AWS use the same implementation:

- 2 Domain Controllers to assure the High Availability
- Billing by the hour of use for a given range of use
- Storage capacity
- Backup frequency
- Additional charging for sharing LDAP data: synchronisation with an existing Active Directory on premises
- Additional charging for data transfers out of the current region (replicas set)

The range of use correspond to a number of users registered to the LDAP Server, who can possibly access the LDAP server simultaneously. For Azure or AWS, the ranges are nearly the same; for example, the standard range (the lowest and cheapest one) is:

- for Azure 25 000 impacted LDAP objects (users, user groups, computers, equipment… all what is possible to register to a LDAP Server), that means until 3 000 users according to AWS
- for AWS 30 000 impacted LDAP objects (users, user groups, computers, all what is possible to register to a LDAP Server), that means until 5 000 users according to AWS

It seems it is not necessary to explore a solution beyond this standard metrics for DigiEcoQuarry: each Data Lake per Pilot Site should not exceed the limit of 3 000 users…

Usage	Nb of Users	Nb of Impacted LDAP Objects	Backup Frequency	Storage Capacity	Tarif / hour (€) (for 2 controllers)	Tarif / hour (€) (for Load Balancing)
Azure Standard	3 000	25 000	each 5 days		0,14	negligible
AWS Standard	5 000	30 000		1 Go	0,1188	included in the price

Table 90: Azure and AWS Active Directory Service metrics

The main charge for a Load Balancer is the number of rules used per hour by the Load Balancer. The charge is set to "negligible" because no rules should be written over the Load Balancer… So only the input and output traffic should be charged: 0,005 € / Go. But the requests established over the Active Directory will be light and will not generate a lot of traffic per month.

For the rest, the recommendations are

- to build the LDAP Server "ex-nihilo" without any sharing user data with an existing one (moreover, the Pilot Sites did not give their formal consent to remotely share user information from an existing LDAP Server if they have one…)
- to not use any replicas set: to not transfer user information from Active Directory out of the working region; for example, the backup can be done over a VM of the same datacenter or a datacenter of the same region.

With the respect of these recommendations, the Active Directory Service implementation remains cheap compared to the high securitization and the state of art that it provides.

7.1.2.7.3 Costs Summary

7.1.2.7.3.1 7/7 – 24/24

Usage	Nb hours of use per month	Azure	Amazon	Open Source
Standard	730	102	87	

Table 91: Azure and AWS Active Directory Service price for a "7/7 - 24/24" use

7.1.2.7.3.2 5/7 – 24/24

Usage	Nb hours of use per month	Azure	Amazon	Open Source
Standard	530	74	63	

Table 92: Azure and AWS Active Directory Service price for a "5/7 - 24/24" use

7.1.2.7.3.3 5/7 – 15/24 (from 5h00 to 20h00)

Usage	Nb hours of use per month	Azure	Amazon	Open Source
Standard	330	46	39	

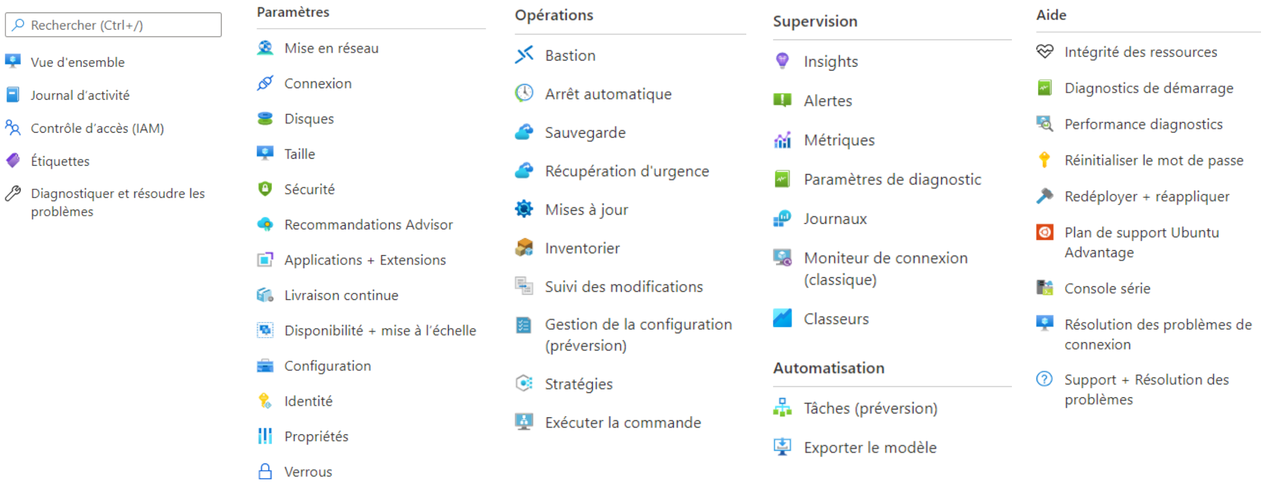
Table 93: Azure and AWS Active Directory Service price for a "5/7 - 15/24 (from 5h00 to 20h00" use

7.1.2.8 Others: VM Monitoring

VM over a Cloud must be monitored. The Clouds allow, through specific HMIs, to natively create and monitor VM. This service is free of charge, it can be accessed with a simple subscription over the Cloud.

7.1.2.8.1 Azure Example

Azure provides all services needed to create and monitor resources, as the menus shows it below:



Rechercher (Ctrl+/)

- Vue d'ensemble
- Journal d'activité
- Contrôle d'accès (IAM)
- Étiquettes
- Diagnostiquer et résoudre les problèmes

Paramètres

- Mise en réseau
- Connexion
- Disques
- Taille
- Sécurité
- Recommandations Advisor
- Applications + Extensions
- Livraison continue
- Disponibilité + mise à l'échelle
- Configuration
- Identité
- Propriétés
- Verrous

Opérations

- Bastion
- Arrêt automatique
- Sauvegarde
- Récupération d'urgence
- Mises à jour
- Inventorier
- Suivi des modifications
- Gestion de la configuration (préversion)
- Stratégies
- Exécuter la commande

Supervision

- Insights
- Alertes
- Métriques
- Paramètres de diagnostic
- Journaux
- Moniteur de connexion (classique)
- Classeurs

Automatisation

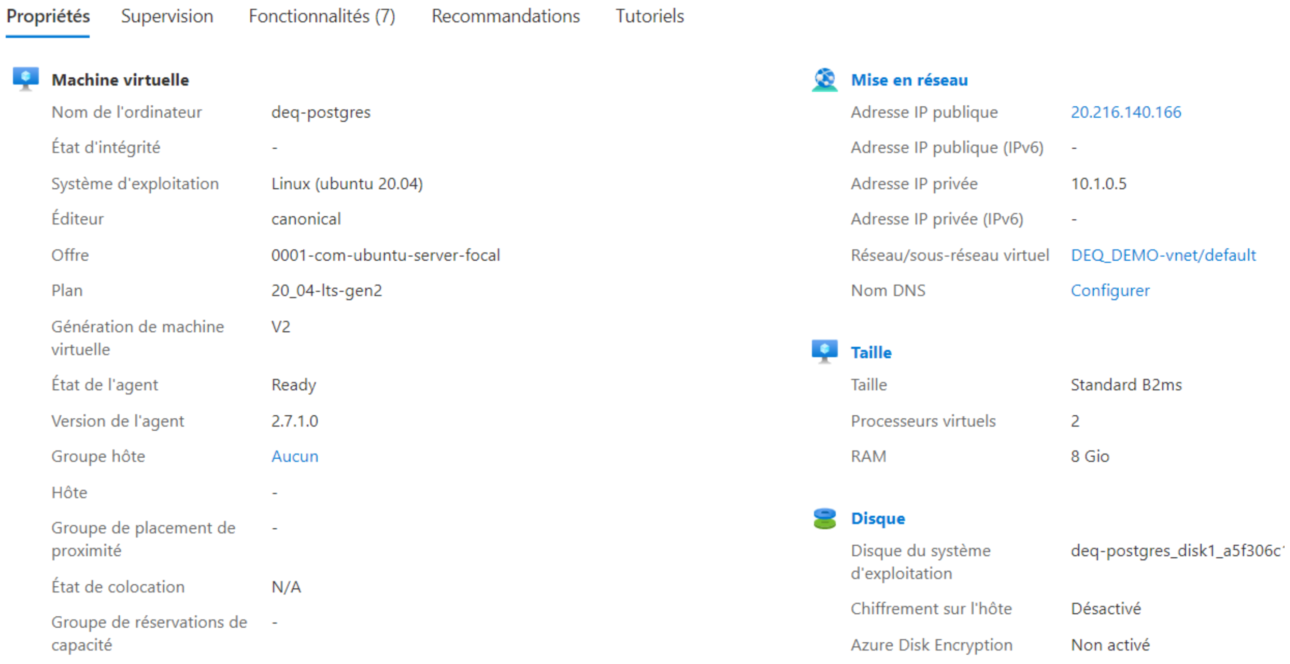
- Tâches (préversion)
- Exporter le modèle

Aide

- Intégrité des ressources
- Diagnostics de démarrage
- Performance diagnostics
- Réinitialiser le mot de passe
- Redéployer + réappliquer
- Plan de support Ubuntu Advantage
- Console série
- Résolution des problèmes de connexion
- Support + Résolution des problèmes

Table 94: Azure monitoring resources menus

Focus on analytic properties of a resource (Overview menu):



Propriétés | Supervision | Fonctionnalités (7) | Recommandations | Tutoriels

Machine virtuelle

Nom de l'ordinateur	deq-postgres
État d'intégrité	-
Système d'exploitation	Linux (ubuntu 20.04)
Éditeur	canonical
Offre	0001-com-ubuntu-server-focal
Plan	20_04-lts-gen2
Génération de machine virtuelle	V2
État de l'agent	Ready
Version de l'agent	2.7.1.0
Groupe hôte	Aucun
Hôte	-
Groupe de placement de proximité	-
État de colocation	N/A
Groupe de réservations de capacité	-

Mise en réseau

Adresse IP publique	20.216.140.166
Adresse IP publique (IPv6)	-
Adresse IP privée	10.1.0.5
Adresse IP privée (IPv6)	-
Réseau/sous-réseau virtuel	DEQ_DEMO-vnet/default
Nom DNS	Configurer

Taille

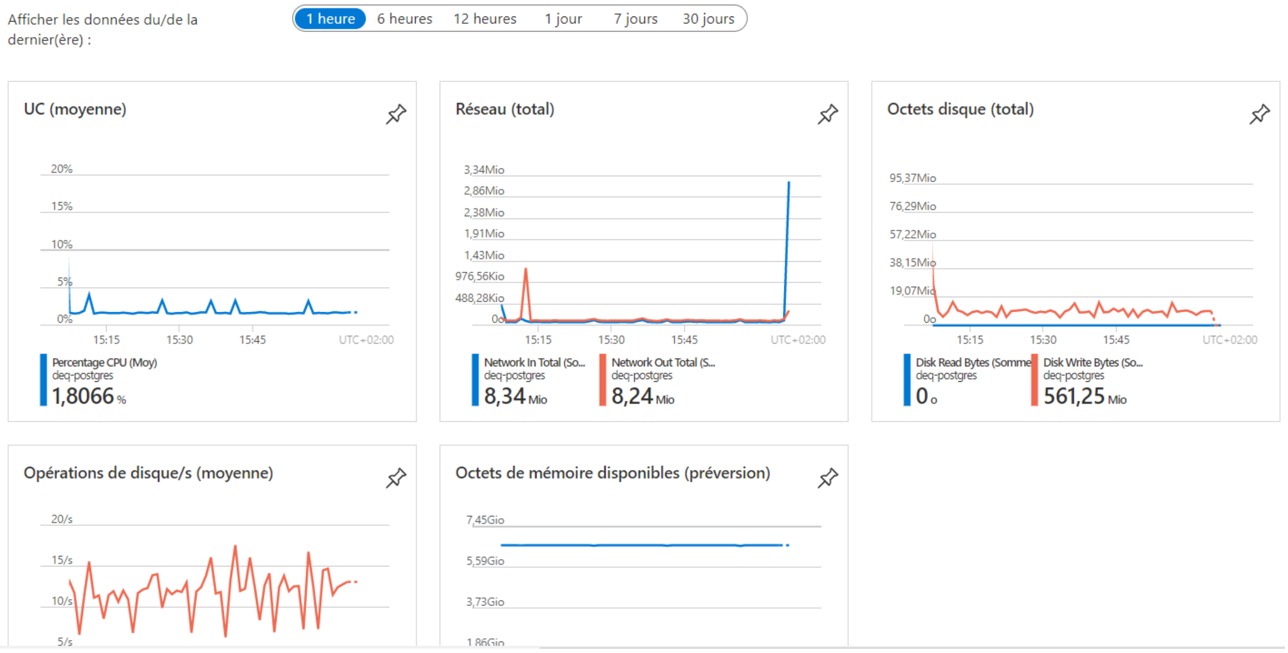
Taille	Standard B2ms
Processeurs virtuels	2
RAM	8 Gio

Disque

Disque du système d'exploitation	deq-postgres_disk1_a5f306c
Chiffrement sur l'hôte	Désactivé
Azure Disk Encryption	Non activé

Figure 50: Azure monitoring IHM - Properties of a resource

Focus on a resource monitoring (Overview menu):



Afficher les données du/de la dernier(ère) :

Figure 51: Azure monitoring IHM - Overview

7.1.2.9 Summary of costs: Data Lake tariff for a selection of components

7.1.2.9.1 Component's Choice

Here is presented a minimalistic design of what must be implemented as Data Lake components.

The IoT components (especially IoT Gateway, IoT treatments and BI components) are not included into this design, they are treated independently in the next chapter.

7.1.2.9.1.1 Design explanation

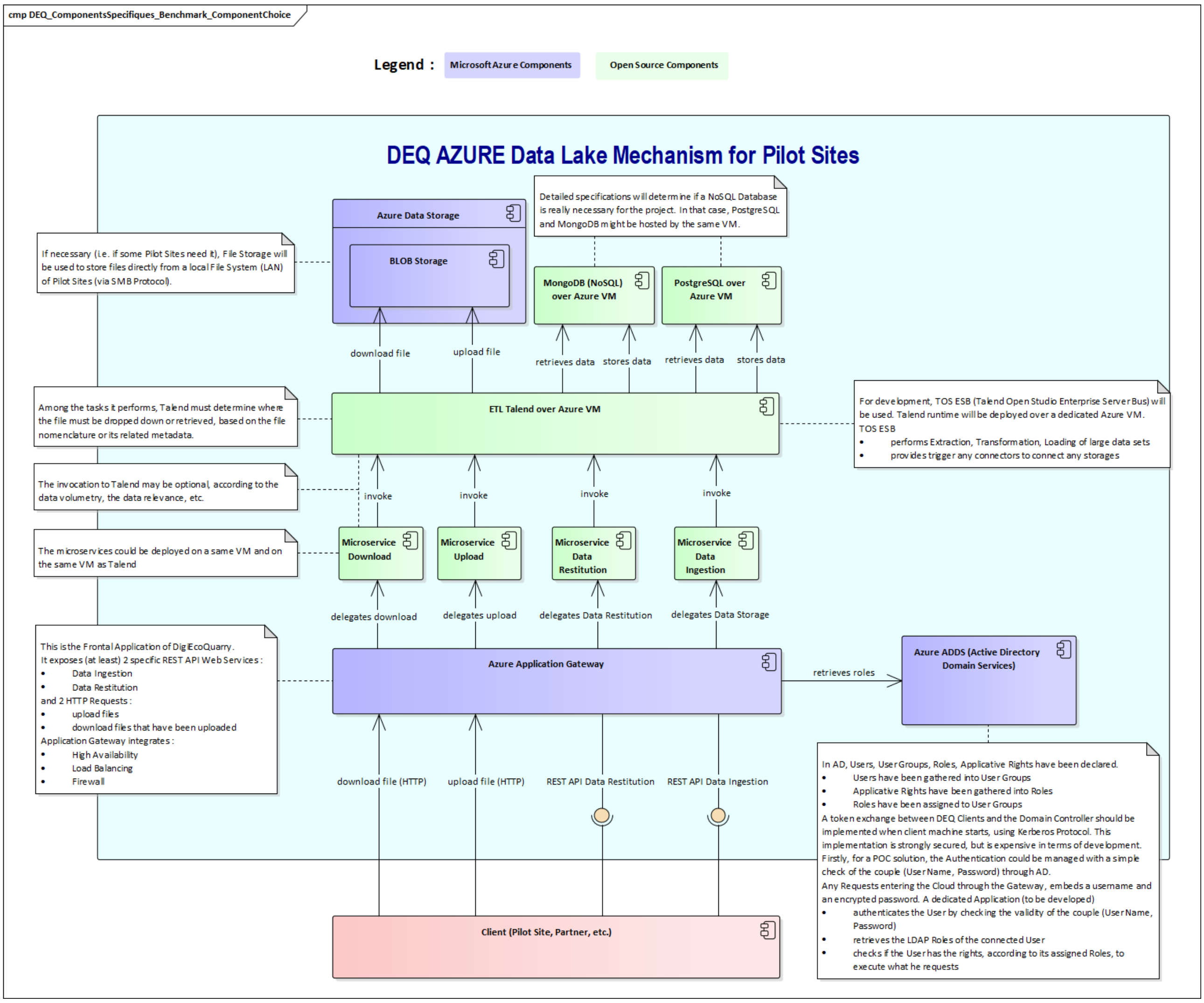
Component	Comment / Justification
Azure Cloud	Azure and AWS are very close in terms of pricing; but, for some components, the Azure pricing seems to be easier to understand (e.g., in terms of volumetry for a Gateway). A successful implementation of the targeted solution will depend on the knowledge of the cloud solution and the availability of the right skill to deploy and run the right feature.
Azure Application Gateway	For a weak use, the price remains low and interesting, especially compared to the features provided. Moreover, the Azure Application Gateway will expose HTTP and REST API requests and trigger the treatments over dedicated computation services: in other terms, and given the expected volumetry, it will replace (advantageously in terms of price) the Azure App Service. Obviously, depending on the volumetry, an open-source solution could be chosen.
Azure ADDS (Active Directory Domain Services)	The user authentication and authorization will be assured by Azure ADDS which embeds the full Domain Controller service (Azure Active Directory + DNS). In a second time, a KDC (Key Distribution Center) with Kerberos protocol should be integrated. Note that we do not plan to implement the Azure RBAC (Role Based Access Control) whose the standard roles that it provides will not add any value in DEQ project: the few roles, mandatory to run each Pilot Site, will be created as usual into Active Directory.
Talend: TOS ESB	Talend Open Studio Enterprise Server Bus offers both-in-one features of an ETL and an ESB; so it will also acts as a workflow orchestrator if needed. It is easy to use it, and it is widely and commonly used by all IT companies: development resources are easy to find. Talend community is very active and easy to join. For the price, according to the volumetry, the appropriate WM will be chosen. For the price simulation, we selected a strong and robust VM with high level characteristics, especially dedicated to the computation (calculation).
BLOB Storage / File Storage	As the Data Lake needs to store collected files of any types, coming from Pilot Sites or Partners, a storage must be implemented. BLOB storage is the perfect candidate for this; in fact, it is not recommended to use a non-specific file system over a dedicated VM when Azure provides a specific BLOB storage, especially designed to store this kind of files: nothing over Azure will not be more efficient than this. Note that, if needed, the Azure File Storage can also be used for DEQ specificities: a reconciliation with the Pilot Sites is necessary to know if this functionality is really expected or not.
Databases	For structured data, PostgreSQL will be deployed on a dedicated VM, especially selected for this kind of works (appropriate for storage, I/O read/write). If needed, MongoDB will be also deployed as a noSQL database.

Component	Comment / Justification
Components Deployement	AKKA recommend installing and deploying open-source components over Azure VM with some Ansible code.
VM Monitoring	Azure Cloud VM will be monitored with Azure native tools (HMI).

Table 95: Data Lake Components choice on Azure DEQ Cloud

7.1.2.9.1.2 Design schema

The previous explanation can be presented with the following schema:



Detailed specifications will determine if a NoSQL Database is really necessary for the project. In that case, PostgreSQL and MongoDB might be hosted by the same VM.

If necessary (i.e. if some Pilot Sites need it), File Storage will be used to store files directly from a local File System (LAN) of Pilot Sites (via SMB Protocol).

Among the tasks it performs, Talend must determine where the file must be dropped down or retrieved, based on the file nomenclature or its related metadata.

For development, TOS ESB (Talend Open Studio Enterprise Server Bus) will be used. Talend runtime will be deployed over a dedicated Azure VM.
TOS ESB
- performs Extraction, Transformation, Loading of large data sets
- provides trigger any connectors to connect any storages

The invocation to Talend may be optional, according to the data volumetry, the data relevance, etc.

The microservices could be deployed on a same VM and on the same VM as Talend

This is the Frontal Application of DigiEcoQuarry.
It exposes (at least) 2 specific REST API Web Services :
- Data Ingestion
- Data Restitution

and 2 HTTP Requests :
- upload files
- download files that have been uploaded

Application Gateway integrates :
- High Availability
- Load Balancing
- Firewall

In AD, Users, User Groups, Roles, Applicative Rights have been declared.
- Users have been gathered into User Groups
- Applicative Rights have been gathered into Roles
- Roles have been assigned to User Groups

A token exchange between DEQ Clients and the Domain Controller should be implemented when client machine starts, using Kerberos Protocol. This implementation is strongly secured, but is expensive in terms of development. Firstly, for a POC solution, the Authentication could be managed with a simple check of the couple (User Name, Password) through AD.
Any Requests entering the Cloud through the Gateway, embeds a username and an encrypted password. A dedicated Application (to be developed)
- authenticates the User by checking the validity of the couple (User Name, Password)
- retrieves the LDAP Roles of the connected User
- checks if the User has the rights, according to its assigned Roles, to execute what he requests

Figure 52: Benchmark Data Lake Components selection

7.1.2.9.2 Selected scenario evaluation

The tariff of this component selection considers the results given by the previous paragraphs, weighted by an amount of processed data (incoming, analysing, outgoing data).

Azure Application Gateway	Azure ADDS	Talend + Microservices on Azure VM	Data Lake Storage Gen2	PostgreSQL*	Deployment over Azure VM by Ansible	Azure VM Monitoring	TOTAL € / month
60	70	180	50	100		-	**< 500**

Table 96: An idea of price for a Data Lake component's choice

*Note that if MongoDB has to be deployed, it might be hosted with PostgreSQL over the same VM. The main point is that this VM must be powerful and robust enough, especially configured and tuned to host database processes.

7.1.3 IoT components comparison

In the context of this benchmark, we will assess the cost of the main components that can be used to build an IoT platform enabling generic IoT applications development.

7.1.3.1 Overview: components presentation to be studied

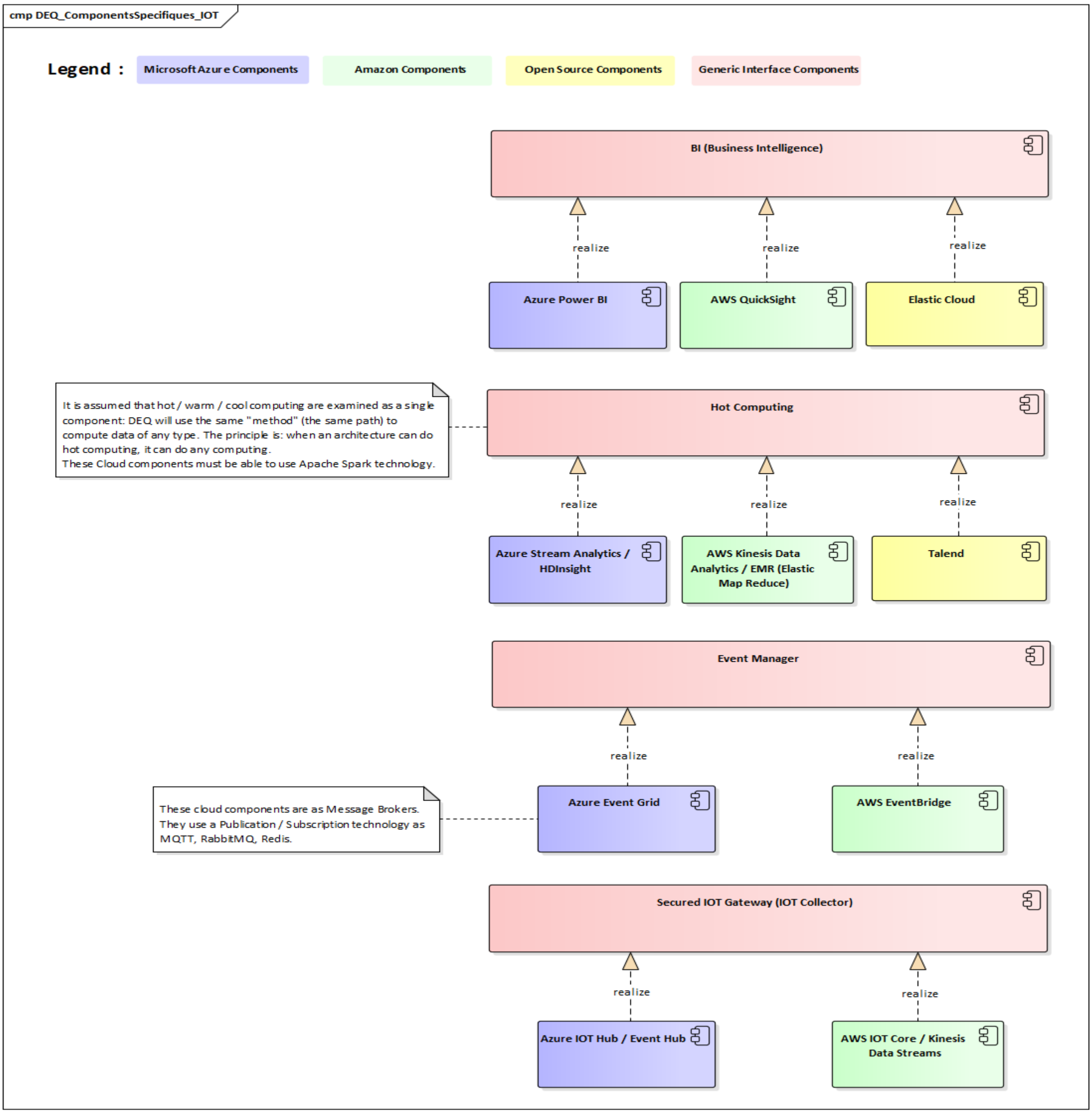

cmp DEQ_ComponentsSpecifiques_IOT
Legend :
Microsoft Azure Components
Amazon Components
Open Source Components
Generic Interface Components
BI (Business Intelligence)
realize
realize
realize
Azure Power BI
AWS QuickSight
Elastic Cloud
It is assumed that hot / warm / cool computing are examined as a single component: DEQ will use the same "method" (the same path) to compute data of any type. The principle is: when an architecture can do hot computing, it can do any computing.
These Cloud components must be able to use Apache Spark technology.
Hot Computing
realize
realize
realize
Azure Stream Analytics / HDInsight
AWS Kinesis Data Analytics / EMR (Elastic Map Reduce)
Talend
Event Manager
realize
realize
These cloud components are as Message Brokers. They use a Publication / Subscription technology as MQTT, RabbitMQ, Redis.
Azure Event Grid
AWS EventBridge
Secured IOT Gateway (IOT Collector)
realize
realize
Azure IOT Hub / Event Hub
AWS IOT Core / Kinesis Data Streams

Figure 53: IoT components to be benchmarked

7.1.3.2 IoT Hub

7.1.3.2.1 IoT Hub Metrics

7.1.3.2.1.1 Generic IoT Hub Metrics for various platforms

	Azure IoT Hub	Amazon IoT Core	Cisco IoT Control Centre	Google cloud IoT Core	IBM Watson IoT Platform
Pricing plan	• Basic tier: 9,31€ to 465,35€ per unit/per month • Standard tier: 23,27€ to 2326,75€ per unit/per month The price within the tier depends on the number of messages exchanged per day (up to 400000, 6 million or 300 million)	• Connectivity: 0,074€ per million minutes of connection • Messaging: 0,65-0,93€ per million messages (the more messages the cheaper) • Devices shadow and registry: 1,16€ per million operations • Rules engines: 0,14€ per million rules triggered/ actions executed	Details are available at request	0,00042- 0,0042€ per MB of data exchanged (the more data the cheaper)	Starts at 465,35€ per unit/per month
Free tier	Up to 8000 messaged per day and up to 500 registered devices	Available for 12 months • 2250000 minutes of connection • 500000 messages • 225000 registry or devices shadow operations • 225000 rules triggered and 225000 actions executed	No free tier	First 250 MB	No free tier
Free tier across additional IoT services	• 12-month free trial of popular Azure services • 186,14€ credit to explore Azure for 30 days • 25+ always free services	Available for 12 months • Device Management: 50 remote actions per month • AWS greengrass: 3 devices • AWS IoT Events: 250000 message evaluations per month • AWS IoT Analytics: 100MB of data processes and 10 GB of data storage	No free tier	• 12-month free trial with 279,21€ credit to spend on any Google Cloud Services • The large suite of always free resources	No free tier

Table 97: IoT Hub Metrics for various platforms

7.1.3.2.1.2 Azure IoT Hub Metrics

https://docs.microsoft.com/en-us/azure/iot-hub/iot-hub-scaling?branch=release-iotbasic

Usage	Azure IoT Hub	
	Azure IoT Hub - Basic	**Azure IoT Hub - Standard**
Weak	B1: 400 000 messages / 4 Ko / day	S1: 400 000 messages / 4 Ko / day
Medium	B2: 6 000 000 messages / 4 Ko / day	S2: 6 000 000 messages / 4 Ko / day
Intensive	**NA**	**NA**

Table 98: Azure IoT Hub Metrics

Based on the number of messages and the type of communication (Bidirectional YES/NO) we can easily select the type of service:

Feature	Basic	Standard / Free
Device-to-cloud telemetry	✓	✓
Per-device identity	✓	✓
Message Routing, Event Grid Integration	✓	✓
HTTP, AMQP, MQTT Protocols	✓	✓
DPS Support	✓	✓
Monitoring and diagnostics	✓	✓
Device Streams^PREVIEW		✓
Cloud-to-device messaging		✓
Device Management, Device Twin, Module Twin		✓
IoT Edge		✓

Table 99: Azure IoT Hub Features

7.1.3.2.1.3 AWS IoT Core Metrics

AWS IoT Core			
Connectivity	**MQTT or HTTP Messaging***	**Device Shadow (DTwin)****	**Rules Engine*****
0,0864 € per million minutes of connection	.1,08 € per million messages for <= 10^9 messages .0,864 € per million messages for the 4x10^9 following messages .0,756 € per million messages for > 5x10^9 messages	1,35 € per million operations that access or modify Device Shadow or Registry data	0,162 € per million rules triggered / per million actions executed

Table 100: AWS IoT Core price metrics

*Max message size: 128 Ko. Messages are metered in 5 Ko increment (e.g., a 6 Ko message is valued as 2 messages). Here are counted both incoming messages (from devices to IoT Core) and outgoing messages (from IoT Core to devices).

**Operations are metered in 1 Ko increment of Device Shadow record size.

***Rules Engine allows to transform device data using arithmetic operations or external functions such as AWS Lambda, and then route the data to an AWS service such as Amazon Simple Storage Service (Amazon S3). Rules Engine use is metered for each time a rule is triggered, and for the number of actions executed within a rule.

7.1.3.2.2 IoT Hub Cost

7.1.3.2.2.1 Azure IoT Hub Price

Usage	**Azure**	
	IoT Hub Basic Device cloud	IoT Hub Standard Bidirectional + DTwin
Weak	11 €	28 €
Medium	56 €	280 €
Intensive	NA	NA

Table 101: Azure IoT Hub Price per month

7.1.3.2.2.2 AWS IoT Core Price

First, determine the amount the messages and minutes of connection for AWS in DEQ context:

Usage	**Azure IoT Core**				
	Nb Devices	**Connectivity**	**Nb Messages**	**Device Shadow**	**Rules Engine**
Weak	50	43 800 mn / month x 50 = 2 190 000	400 000* per day	2 calls x 50 devices x 60 mn x 24 h x 30 days = 4 320 000 calls / month	1 rule x 400 000 messages = 400 000 rules per day 2 actions x 400 000 rules = 800 000 actions par day
Medium	200	43 800 mn / month x 200 = 8 760 000	6 000 000** per day	5 calls x 200 devices x 60 mn x 24 h x 30 days = 43 200 000 calls / month	1 rule x 6 000 000 messages = 6 000 000 rules per day 2 actions x 6 000 000 rules = 12 000 000 actions per day
Intensive	NA	NA	NA	NA	NA

Table 102: AWS IoT Core Metrics for DEQ context

*400 000 messages (5Ko) / day sent by 50 devices ==> a rounded average of 5 to 6 messages / minute sent by each device (1 message sent each 10 sec by each device)

**6 000 000 messages (5Ko) / day sent by 200 devices ==> a rounded average of 21 messages / minute sent by each device (1 message sent each 3 sec by each device)

Then, apply the AWS IoT Core price metrics to DEQ context:

Usage	AWS IoT Core				
	Connectivity	**Nb Messages**	**Device Shadow**	**Rules Engine**	**Total (€)**
Weak	2 190 000 x 0,0864 / 10^6 = **0,2 €**	400 000 x 30 x 1,08 / 10^6 = **12 €**	4 320 000 x 1,35 / 10^6 = **5,8 €**	400 000 x 0,162 / 10^6 = 0,0648 800 000 x 0,162 / 10^6 = 0,1296 (0,0648 + 0,1296) x 30 = **6 €**	Standard: **18** DTwin: **24**
Medium	8 760 000 x 0,0864 / 10^6 = **0,8 €**	6 000 000 x 1,08 x 30 / 10^6 = **195 €**	43 200 000 x 1,35 / 10^6 = **58 €**	6 000 000 x 0,162 / 10^6 = 0,972 12 000 000 x 0,162 / 10^6 = 1,944 (0,972 + 1,944) x 30 = **90 €**	Standard: **285** DTwin: **343**
Intensive	NA	NA	NA	NA	**NA**

Table 103: AWS IoT Core Price per month

7.1.3.2.2.3 IoT Hub Summary

Usage	Azure		Amazon	
	IoT Hub Basic Device cloud	IoT Core Bidirectional + DTwin	IoT Core Basic	IoT Core Bidirectional + DTwin
Weak	11 €	28 €	18 €	24 €
Medium	56 €	280 €	285 €	343 €
Intensive	NA	NA	NA	NA

Table 104: Azure and AWS IoT Hub cost per month comparison

7.1.3.3 IoT Gateway

7.1.3.3.1 IoT Gateway Metrics

7.1.3.3.1.1 Azure Event Hub Metrics

The metrics are based upon:

- the ingress events: the number of events coming from any devices to the Event Hub server
- the capacity: the volumetry of the incoming events
- the capture (optional): the operation which consists in processing the events i.e., recovering them and ingesting them into the storage system

The table below shows the prices charged for these different metrics:

	Basic	Standard	Premium	Dedicated*
Capacity	0,014€/hour per Throughput Unit***	0,027€/hour per Throughput Unit***	1,110€/hour per Processing Unit (PU)	6,854€/hour per Capacity Unit (CU)
Ingree events	0,026€ per million events	0,026€ per million events	Included	Included
Capture		65,683€/month per Throughput Unit***	Included	Included
Apache Kafka		✓	✓	✓
Schema Registry		✓	✓	✓
Max Retention Period	1 day	1 day	90 days	90 days
Storage Retention	84 GB	84 GB	1 TB per PU	10 TB per CU
Extended Retention**			0,11€/GB/month (1 TB included per PU)	0,11€/GB/month (10 TB included per CU)

Table 105: Azure Event Hub Metrics

* Dedicated: Usage will be charged in one-hour increments with a minimum charge for four hours of usage.

** Message retention above the included storage quotas will result in overage charges.

*** Throughput Unit provides 1 MB/s ingress and 2 MB/s egress.

Obviously, at each upgrade of range, Azure provides more services and more volume (storage retention) and more time (retention period).

To be able to compare with IoT Hub, here are used the same volumetry ranges applicable to DEQ.

Usage	**Azure Event Hub metrics applicable to DEQ**	
	Nb Devices	**Azure Event Hub**
Weak	50	400 000 messages / 4 Ko / day
Medium	200	6 000 000 messages / 4 Ko / day
Intensive	**NA**	**NA**

Table 106: Azure Event Hub volumetry for DEQ project

Remark: Note that Event Hub (as IoT Hub) can act as a buffer, gathering a batch of events into a single set; and it is this events' set that is sent to the next component Event Grid. For example, instead of sending 10 messsages weighing 0,4 Ko each, it can be sent 1 message weighing 4 Ko for saving money to reduce Event Grid costs.

7.1.3.3.1.2 AWS Kinesis Metrics

This component was not evaluated since Azure cloud platform was selected.

7.1.3.3.2 IoT Gateway Price

7.1.3.3.2.1 Azure Event Hub Price

First, compute the capacity unit for DEQ volumetry (the incoming unit throughput is 1 Mo/s, and it is assumed that any event volumetry is less than 4 Ko):

Usage	Nb Devices	Nb Events	Throughput (Ko / sec)	Capacity Unit
Weak	50	400 000 (4Ko) / day sent by all devices	400 000 x 4Ko / 24h / 3600s = 18,52 Ko/s	18,52 Ko/s <= 1 Mo/s == > 1 CU
Medium	200	6 000 000 (4Ko) / day sent by all devices	6 000 000 x 4Ko / 24h / 3600s = 277,78 Ko/s	277,78 Ko/s <= 1 Mo/s == > 1 CU
Intensive	NA	NA	NA	NA

Table 107: Throughput and Capacity Unit for DEQ using Azure Event Hub

Usage	Azure Event Hub					
	Capacity		Ingress Events	Capture	Total (€ / month)	
	Basic	Standard	(Basic and Standard)	(only applicable to Standard)	Basic	Standard
Weak	0,014 x 1CU x 24h x 30 days = **10 €**	0,027 x 1CU x 24h x 30 days = **17 €**	400 000 x 30 x 0,026 / 10^6 = **0,31 €**	65,683 x 1CU = **65,683 €**	**10**	**83**
Medium	0,014 x 1CU x 24h x 30 days = **10 €**	0,027 x 1CU x 24h x 30 days = **17 €**	6 000 000 x 30 x 0,026 / 10^6 = **4,68 €**	65,683 x 1CU = **65,683 €**	**15**	**88**
Intensive	NA		NA	NA	NA	

Table 108: Azure Event Hub Price

7.1.3.4 Event Manager

7.1.3.4.1 Event Manager Metrics

7.1.3.4.1.1 Azure Event Grid Metrics

Event Grid is the pub/sub solution for Azure Cloud.

Event Grid Basic tier is priced as pay-per-use based on operations performed.

Operations include

- ingress of events (of 64 Ko) to Domains or Topics,
- advanced matches (using filtering to route to end-points),
- delivery attempt,
- management calls.

Plan pricing includes a monthly free grant of 100,000 operations.

Azure Event Grid Tariffication
0,54 € per million operations*
Free = < 100 000 operations / month

Table 109: Azure Event Grid Tariffication

As Event Grid is set "just behind" the Event Hub component, the same volumetry used for Event Hub is applied:

Usage	Azure Event Grid metrics applicable to DEQ				
	Incoming messages into Event Hub per day	**Publication frequency into Event Grid***		**Operations published into Event Grid per month**	
Weak	400 000 messages / 4 Ko / day	Each incoming messages into Event Hub	Each 10 incoming messages into Event Hub	400 000 x 30 = 12 000 000	/ 10 mes per batch = 1 200 000
Medium	6 000 000 messages / 4 Ko / day			6 000 000 x 30 = 180 000 000	/ 10 mes per batch = 18 000 000
Intensive	NA			NA	

Table 110: Azure Event Grid volumetry for DEQ project

*Because the DEQ volumetry should not be extremely large, the messages can be treated individually. However, it is given two acceptances for this item: single message treated, 10 messages treated per batch. The Event Hub (or IoT Hub) can act as a buffer by constituting a set of 10 (or more) messages before to send it to the Event Grid.

7.1.3.4.1.2 AWS Event Bridge Metrics

This component was not evaluated since Azure cloud platform was selected.

7.1.3.4.2 Event Manager Price

7.1.3.4.2.1 Azure Event Grid Price

It comes that if n is the number of ingress events, 2n is the final number of operations.

Two Important notes:

- It is assumed that the ingress events are <= 64 Ko.
- The delivery (to end-points) is not subject to the 64 Ko rule.

Usage	**Azure Event Grid Price (€ / month)** Frequency: single message treatment	**Azure Event Grid Price (€ / month)** Frequency: 10 messages per batch
Weak	2 x 0,54 x (12 000 000 – 100 000) / 1 000 000 = **12 €**	/ 10 = **1,2 €**
Medium	2 x 0,54 x (180 000 000 – 100 000) / 1 000 000 = **180 €**	/ 10 = **18 €**
Intensive	NA	NA

Table 111: Azure Event Grid Price

7.1.3.5 Computing

For this item, we plan to use Talend (TOS ESB) for computing the messages coming into the Hub (IoT Hub or Event Hub).

The same dedicated VM already discussed on paragraph ETL Tool – Open Source will be used.

7.1.3.6 Business Intelligence

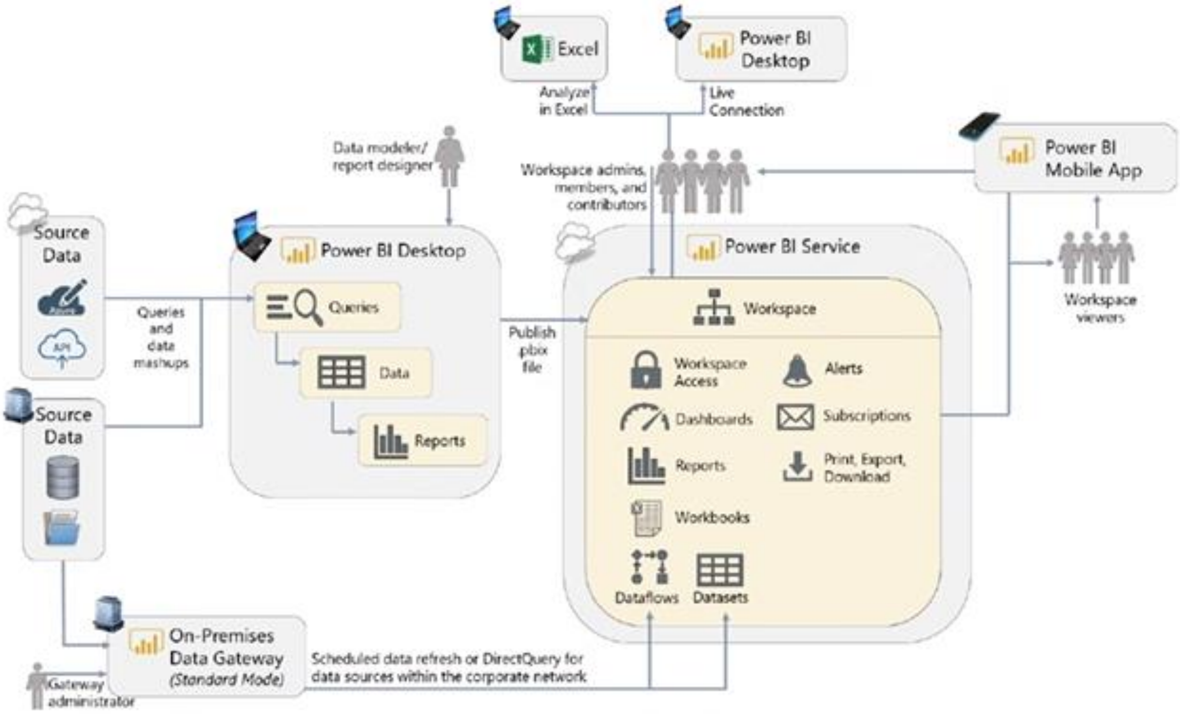
7.1.3.6.1 Generalities

The goal of BI is to make simple, beautiful and, above all, comprehensive, a forest of data that is often dense, sometimes inextricable, always deeply buried into disparate locations, such as data warehouses, databases (structured or not), files in various directories of various machines... which data can be fixed or changing over time.

For this, BI solutions use technologies that know how to process large volumes of disparate data, and which produce structured reports into renderings that are always very polished.

DEQ needs to expose, through BI technologies, a significant number of KPIs. That means that the applications developed under DEQ Project must embed reports, dashboards, and analytics functionalities, to be exposed by DEQ Servers and accessible by Pilot Sites (or Partners if necessary).

BI Solutions chapter has been set in this IoT-Components part because it is involved with IoT data: it computes and transforms IoT data, mainly already refined by Partners, to produce KPIs to be displayed by HMI. However, note that the reports to be generated will be triggered through the frontal Application Gateway described in Data Lake part.



7.1.3.6.2 Solutions for Power BI Enterprise Architecture

Power BI ecosystem has developed too fast in recent years. The entire data solution that has been extended, includes ETL, AI, ML, Synapse cloud data warehouse. All ecological products are in the entire Power BI data platform.

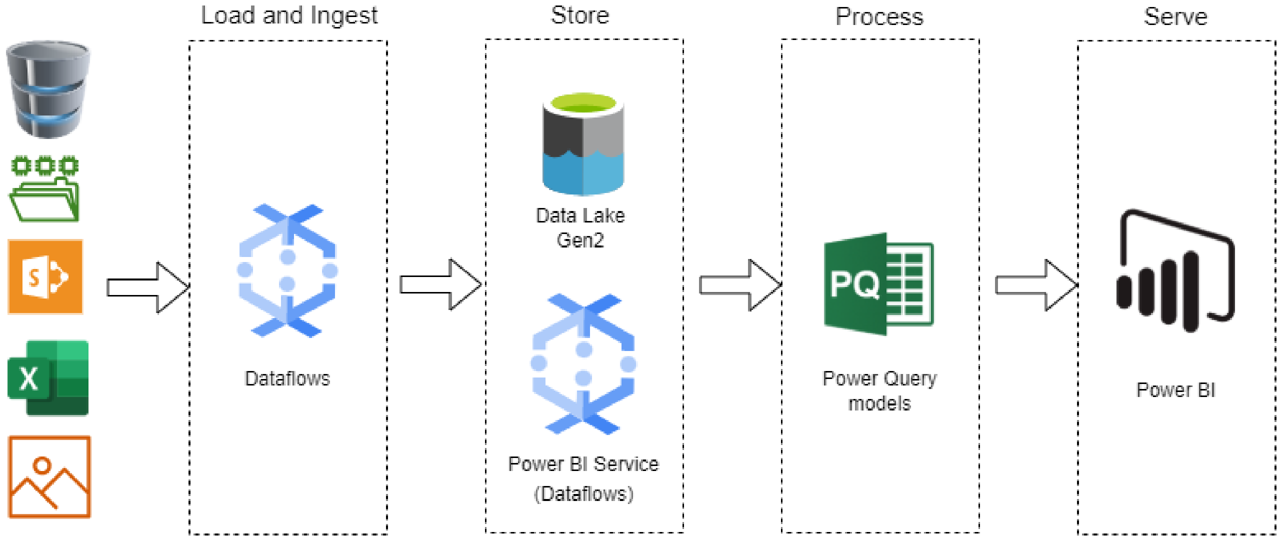
Usually, modern data platforms have 4 steps:

- ETL: Extract, transform and load data from source systems.
- Store: Store the data somewhere (local or cloud) that we can run analytics on it.
- Process: Run analytics on data and plot KPIs, AI, and forecasts.
- Services: Present this data in an easy way for users.

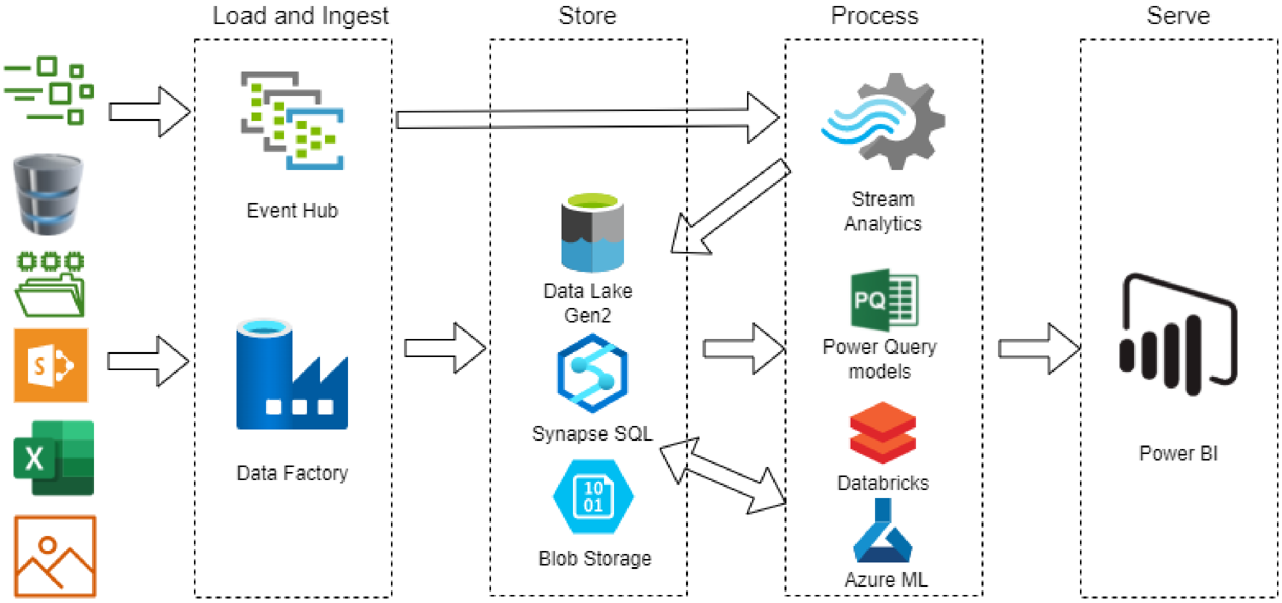
7.1.3.6.2.1 Scale of data scene

Microsoft Power BI is a modern data platform. What is enterprise class? We can simply understand that not only small and medium-sized enterprises, but even huge scales can provide corresponding and matching data scenario application services and ensure stable operation. To simplify the understanding, take "Medium Size" and "Large Size" as examples.

- Medium Size - using Power BI services



- Large Size - using Azure services



7.1.3.6.2.2 Deployment environment

The choice of Power BI deployment environment is related to the publishing, storage, and sharing reports. Its importance is reflected in 3 factors:

- Publish data, reports, and various BI content generated by the enterprise
- Develop an update plan for the data
- Safely and efficiently share data with users

Generally, there are two options for the deployment environment of Power BI:

- Public cloud service (Power BI Service) provided by Azure
- Local report server (Power BI Report Server)

1) Power BI Service

   Power BI Service is a SaaS data analysis reporting service fully hosted on Azure. In terms of architecture, it carries various functions such as data distribution, storage, and management. For end users, Power BI Service is an accessible web port.

   In Excel, users may be accustomed to saving reports on their own computers and publishing them to other users via email or SharePoint. In theory, this approach also works for (.pbix) files generated by Power BI Desktop. The centralized cloud service architecture has the following advantages:

   i) Maintenance cost: users who are not tech-savvy can start using Power BI Service in a short period of time without having to rely on IT for complex deployment planning. Cloud service providers solve management tasks such as server updates and patches, which greatly reduces maintenance costs for users.

   ii) Payment model for cloud services: all licensing agreements can be completed with a monthly payment, saving the upfront investment in software protocols and hardware with traditional server methods. At the same time, Power BI Service allows expansion and addition of users at any time, eliminating the risk of uncertainty in the number of users and data in the early stage of the project, and making architectural decisions more agile.

   iii) Publish and collaborate report service is set up in the cloud, and users can access the server through different terminal devices anytime, anywhere. In this way, users do not have geographical restrictions when digesting data, and can share reports inside and outside the enterprise more safely and effectively.

   iv) Version control: power BI Service serves as the end for developers to publish reports, which can avoid redundancy and lag caused by multiple versions. For service administrators, it is easy to manage and control a centralized service and unify access rights, security, and privacy compliance requirements.

2) Power BI Report Server

   Power BI Report Server is a local replacement service of Power BI Service, which also carries the functions of publishing, storing, and sharing BI content on the server side.

   Power BI Report Server and SQL Server Reporting Service share many functional similarities but note that they are separate in terms of installation and license agreements.

   In deployment environments, we generally consider using PaaS or SaaS services with centralized functions. On-premises deployment scenarios outside of Power BI Service need to be considered only in some special cases. For example, industries with sensitive data or particularly high levels of security (defense).

7.1.3.6.2.3 Power BI Premium service for large deployments

As the name suggests, Power BI Premium offers enterprise-grade premium services to meet the needs of large deployments: lots of read-only users; huge amount of data; fast and instant data updates.

The value-added services of Power BI Premium are reflected in the overall performance of the BI architecture and do not affect the time it takes for a single user to refresh a report.

In terms of working principle, Power BI Premium can be compared to computing resources like virtual machine nodes, providing independent "space" for enterprises using the service. Microsoft has placed many restrictions and bottlenecks on these open "spaces" to ensure the stable operation of the entire shared "space".

Power BI Premium can provide a "higher, faster, stronger" experience in various data processing. We consider the necessity of Premium service from the following perspectives:

- Concurrency of user access
- Concurrency of data updates
- The amount of data queried
- Query complexity
- Data storage mode
- Use of streaming data
- The degree of repetition of calls to the dataset

In deployment process, there are many factors that can determine performance, which is why it is difficult to give specific answers to each consideration. The correct way is to start from finding the problem, test and monitor the changes in the data and user's feedback, put forward new hypotheses and repeatedly verify whether the problem can be solved through the Premium service.

In addition to technical considerations, budget is another decisive factor. We can effectively simulate price models and provide intuitive budget figures by this site https://powerbi.microsoft.com/en-us/pricing/

7.1.3.6.2.4 Power BI Embedded

Power BI Embedded is an embedded service provided by Azure that allows users to embed the Power BI environment as an independent functional unit into an existing application.

Embedded services in the traditional sense are more aimed at decision-making in the software development process. For example, Power BI Embedded is widely used in Microsoft's ISV (Independent Software Vendor) products. They integrated some functions of Power BI into third-party software developed by themselves, thereby increasing the competitiveness of the product in reporting functions. For example, B2C will often customize a system that suits its own situation. When these systems complete complex business logic, they will gradually generate data analysis and reporting requirements. At this time, we can consider embedding Power BI into a known independent system to maintain the consistency of business lines. This situation often requires more powerful technical support to achieve web development and special needs.

7.1.3.6.2.5 Summary

When the previous architectures description is applied to DEQ context, it comes the schema below that compares Power BI Embedded IQS-implementation and Power BI Service IQS-implementation.

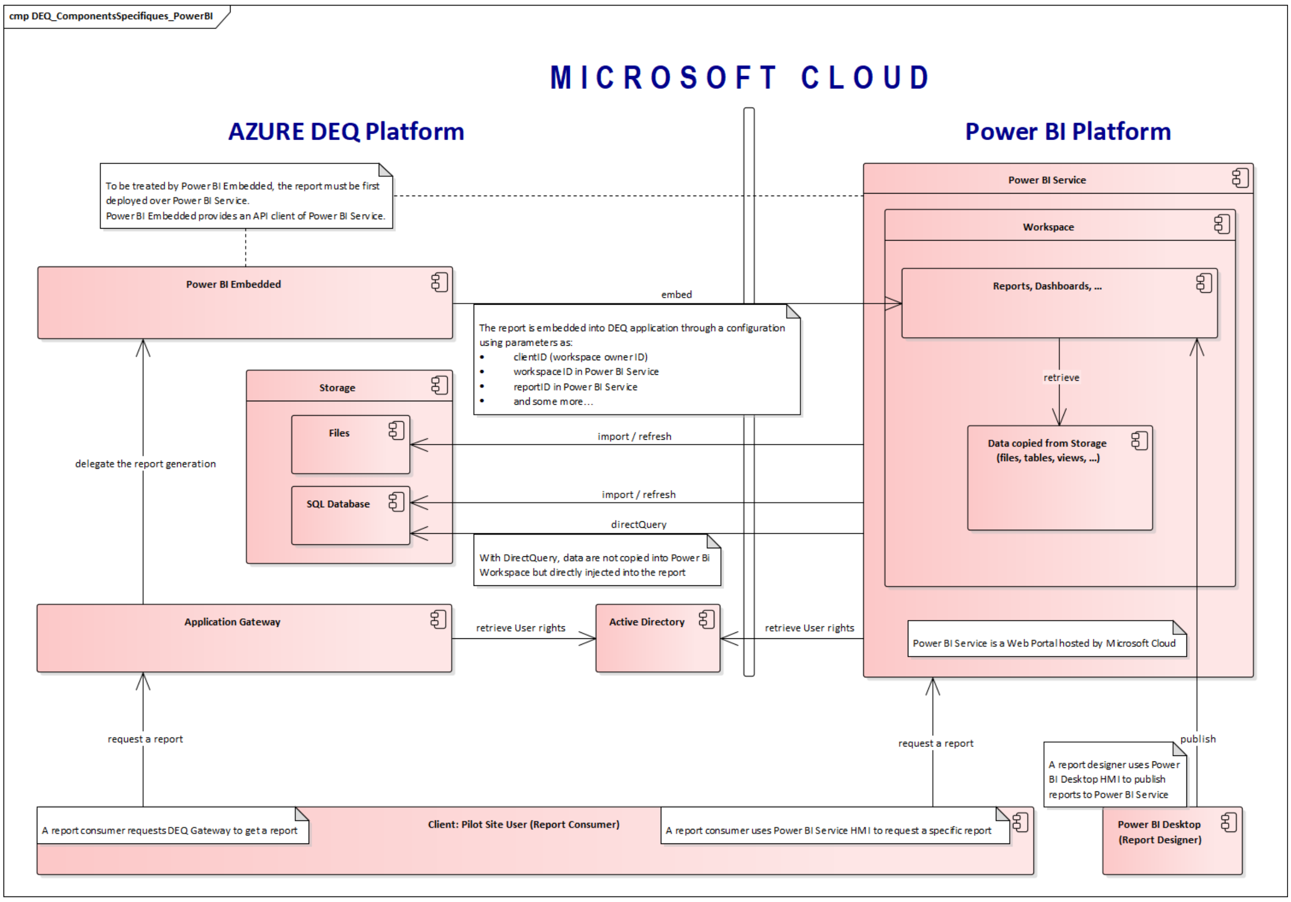

cmp DEQ_ComponentsSpecifiques_PowerBI
MICROSOFT CLOUD
AZURE DEQ Platform
Power BI Platform
To be treated by Power BI Embedded, the report must be first deployed over Power BI Service.
Power BI Embedded provides an API client of Power BI Service.
Power BI Embedded
embed
The report is embedded into DEQ application through a configuration using parameters as:
• clientID (workspace owner ID)
• workspace ID in Power BI Service
• reportID in Power BI Service
• and some more…
Power BI Service
Workspace
Reports, Dashboards, ...
retrieve
Data copied from Storage (files, tables, views, ...)
Storage
Files
SQL Database
import / refresh
import / refresh
directQuery
With DirectQuery, data are not copied into Power Bi Workspace but directly injected into the report
delegate the report generation
Application Gateway
retrieve User rights
Active Directory
retrieve User rights
Power BI Service is a Web Portal hosted by Microsoft Cloud
request a report
request a report
publish
A report designer uses Power BI Desktop HMI to publish reports to Power BI Service
A report consumer requests DEQ Gateway to get a report
Client: Pilot Site User (Report Consumer)
A report consumer uses Power BI Service HMI to request a specific report
Power BI Desktop (Report Designer)

Figure 54: Power BI IQS-implementation - Power BI Embedded / Power BI Service

Power BI Service is a web portal that allows

- designers to publish reports to be visualized by a set of report consumers according to the rights whose they have been granted
- the selection and the visualisation by the consumers, of the reports that have been made available by the designers

And in practise, Power BI Embedded is an API performing the interface between a Cloud specific application (as DEQ Cloud, for example), and Power BI Service that serves the reports. Power BI Embedded cannot be used without using first Power BI Service.

That being given, it is time, now, to evaluate the cost of these two Microsoft Power BI solutions.

7.1.3.6.3 Azure Power BI Embedded

For designing report templates, Microsoft has edited “Power BI Desktop” that produces reports that can be launched and operated as a scenario aggregating real data sets (for example, data from Excel files hosted into a dedicated directory).

Here are the most common uses of Power BI Desktop:

- Connect to data
- Transform and cleanse that data to create a data model
- Create visuals, such as charts, that provide visual representations of the data
- Create reports for collections of visuals, on one or more report pages
- Share reports with other users using the Power BI service

And for implementing the Cloud solution, Azure provides “Power BI Embedded”, an Azure service that exposes, through API reached from a Gateway into the Cloud, templates generated by Power BI Desktop. These templates can be plugged to a data warehouse containing incoming data, or a data flow coming from a database. In other terms, the templates accessed by Clients are always refreshed on the fly with up-to-date data.

Here below are enumerated the requirements for an Azure Power BI Embedded solution:

- an app workspace that hosts the contents to be integrated into the reports-templates generated by Power BI Desktop
- a Power BI Pro license with a unique service account, to proxy Power BI and the API exposing the reports
- the Power BI Pro account must be granted as an administrator of the app workspace
- a workspace capacity as a dedicated resource used to build and execute Power BI reports (a feature allows an administrator to pause the capacity, preventing the BI Servers to be used by any user)
- some code written from Power BI API, to be implemented into a REST API that exposes the requests for generating and rendering BI reports

Power BI Pro license for 1 User: 9€ / month

7.1.3.6.3.1 Metrics for a Power BI Embedded solution

AZURE				
Usage	**Virtual Core**	**Memory RAM**	**Frontend Core / Backend Core**	**Pricing**
Weak	1	3 Go	0,5 / 0,5	0,9071 € / hour
Medium	2	5 Go	1 / 1	1,8069 € / hour
Intensive	4	10 Go	2 / 2	3,6209 € / hour

Table 112: Azure Power BI metrics

The billing is performed according to the availability of the platform, not the real use.

7.1.3.6.3.2 Costs Summary

7.1.3.6.3.2.1 7/7 24/24

Usage	Nb hours of availability per month	Hourly rate	Price (€)
Weak	730	0,9071	**662**
Medium		1,8069	**1 319**
Intensive		3,6209	**2 643**

Table 113: Azure Power BI "7/7 24/24" price

Note: Add negligible 9 € at each price, for 1 User account Power BI Pro licence

7.1.3.6.3.2.2 5/7 24/24

Usage	Nb hours of availability per month	Hourly rate	Price (€)
Weak	530	0,9071	**481**
Medium		1,8069	**958**
Intensive		3,6209	**1 919**

Table 114: Azure Power BI "5/7 24/24" price

Note: Add 9 € per month at each price, for 1 User account Power BI Pro licence

7.1.3.6.3.2.3 5/7 – 15/24 (from 5h00 to 20h00)

Usage	Nb hours of availability per month	Hourly rate	Price (€)
Weak	330	0,9071	**299**
Medium		1,8069	**596**
Intensive		3,6209	**1 195**

Table 115: Azure Power BI "5/7 15/24" price

7.1.3.6.3.2.4 5/7 – 5/24

To become acceptable in price terms, the platform must only run a few hours per day. It becomes interesting below 5 hours a day. It is the reason why this paragraph has been added.

Usage	Nb hours of availability per month	Hourly rate	Price (€)
Weak	110	0,9071	**100**
Medium		1,8069	**200**
Intensive		3,6209	**400**

Table 116: Azure Power BI "5/7 5/24" price

7.1.3.6.4 Power BI Service

7.1.3.6.4.1 Free License

A free Power BI Licence exists, but it is not appropriate for DEQ project.

This licence only allows a user

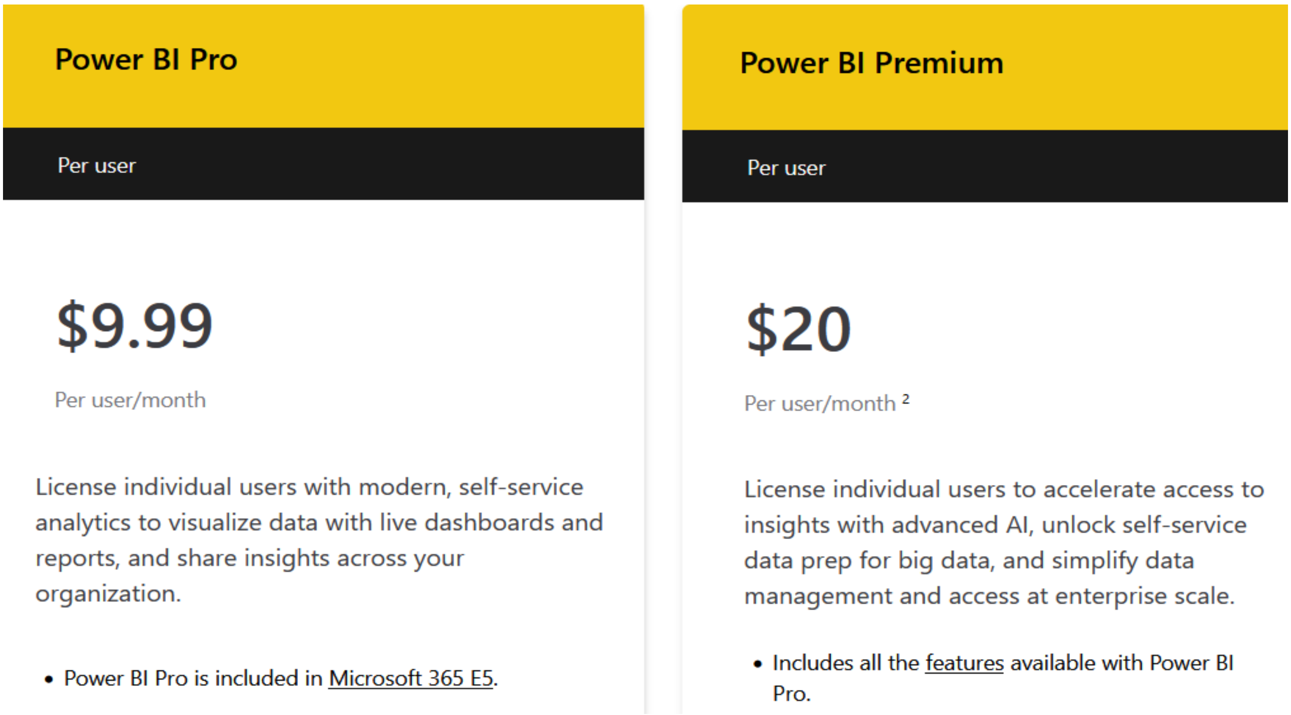
- to create its own workspace into Power BI Service
- to publish reports into its own workspace
- to connect to any kind of data
- to retrieve reports from its own workspace

The user with a free licence

- cannot share anything with anybody else
- and nobody – except him – can retrieve the reports he produced

This license is like a blind tube only lighted for a single user. It allows a developer to make tests and evaluate the solution.

7.1.3.6.4.2 Commercial License



Power BI Pro

Per user

$9.99

Per user/month

License individual users with modern, self-service analytics to visualize data with live dashboards and reports, and share insights across your organization.

- Power BI Pro is included in Microsoft 365 E5.

Power BI Premium

Per user

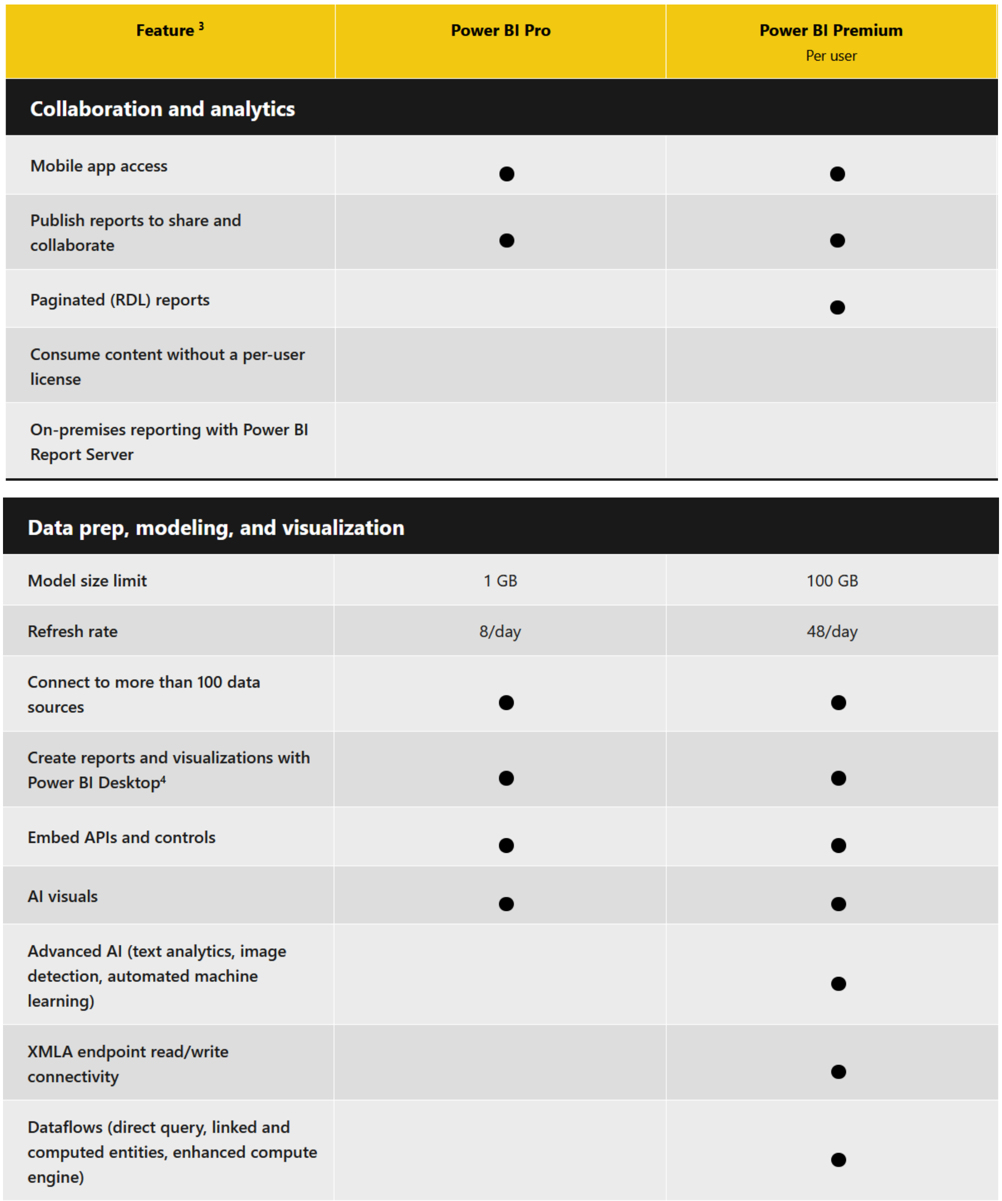
$20

Per user/month [2]

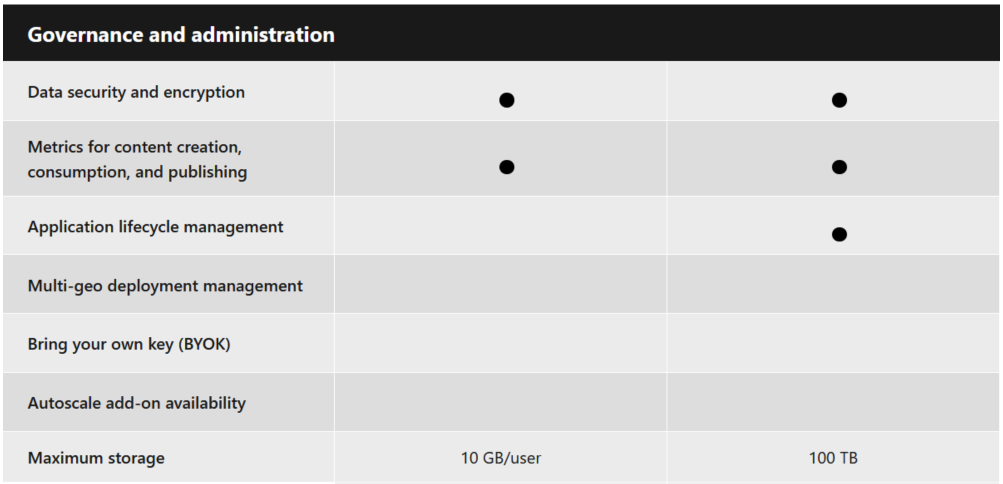
License individual users to accelerate access to insights with advanced AI, unlock self-service data prep for big data, and simplify data management and access at enterprise scale.

- Includes all the features available with Power BI Pro.

Figure 55: Power BI Licences Price



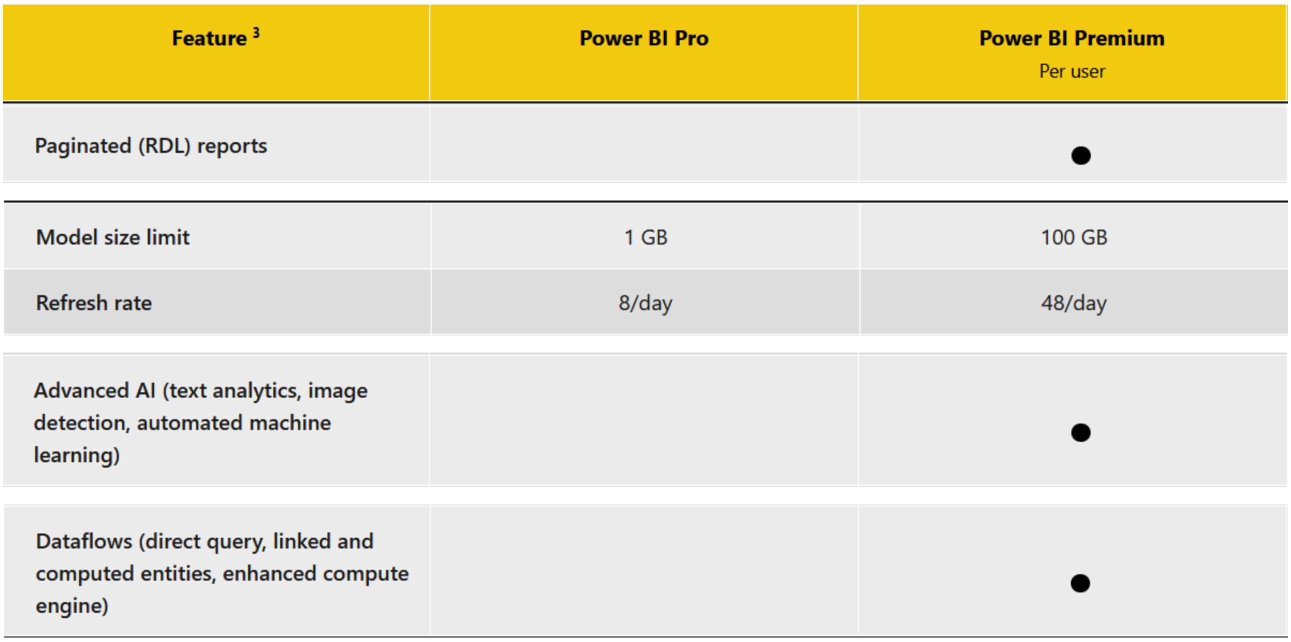
Feature [3]	Power BI Pro	Power BI Premium Per user
Collaboration and analytics		
Mobile app access	●	●
Publish reports to share and collaborate	●	●
Paginated (RDL) reports		●
Consume content without a per-user license		
On-premises reporting with Power BI Report Server		
Data prep, modeling, and visualization		
Model size limit	1 GB	100 GB
Refresh rate	8/day	48/day
Connect to more than 100 data sources	●	●
Create reports and visualizations with Power BI Desktop[4]	●	●
Embed APIs and controls	●	●
AI visuals	●	●
Advanced AI (text analytics, image detection, automated machine learning)		●
XMLA endpoint read/write connectivity		●
Dataflows (direct query, linked and computed entities, enhanced compute engine)		●



Governance and administration		
Data security and encryption	●	●
Metrics for content creation, consumption, and publishing	●	●
Application lifecycle management		●
Multi-geo deployment management		
Bring your own key (BYOK)		
Autoscale add-on availability		
Maximum storage	10 GB/user	100 TB

Figure 56: Power BI Licences Features (figures captured from Microsoft web site)

Below are focused the main features for making the choice between Pro and Premium Licences per User.

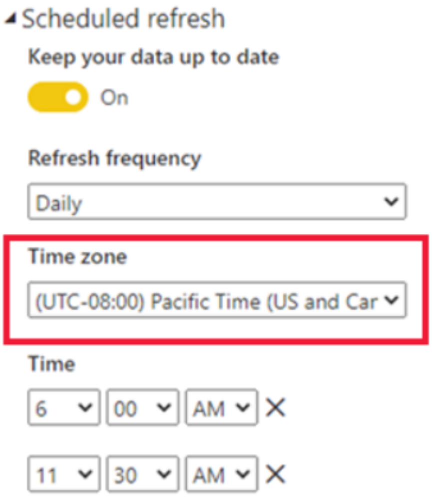


Feature [3]	Power BI Pro	Power BI Premium Per user
Paginated (RDL) reports		●
Model size limit	1 GB	100 GB
Refresh rate	8/day	48/day
Advanced AI (text analytics, image detection, automated machine learning)		●
Dataflows (direct query, linked and computed entities, enhanced compute engine)		●

Figure 57: Main differences between Pro and Premium Licences, impacting DEQ choice

Important remarks:

- The refresh rate must be understood as 8 or 48 times a day per data set. The data refresh can be set by a configuration HMI.



Scheduled refresh

Keep your data up to date

On

Refresh frequency

Daily

Time zone

(UTC-08:00) Pacific Time (US and Car

Time

6 : 00 AM

11 : 30 AM

Figure 58: Refresh Dataset configuration in Power BI

- DirectQuery allows to directly connect the data set that fills the report, to a database. In that case, the data set is not imported into the user workspace. Note that Direct Query is only available with Power BI Premium.

The Licences must be chosen according to each Pilot Site usage. Some usages might require a Premium Licence; for others, a Pro Licence might be enough. Same remark for the number of licences to purchase: 1 licence for 1 user.

Licence Type	Number of Users	Unit Price	Price
Power BI Pro	15	9 €	135 €
Power BI Premium	8	17 €	136 €

Figure 59: Power BI License prices for a specific number of Users

7.1.3.6.5 BI Open-Source solution

An efficient and strongly used open-source solution is the Elastic Suite or ELK Suite for ElasticSearch, Logstash, Kibana.

- ElasticSearch is the search engine and data indexer (columns as types, rows as documents, index as a collection of documents for a same type).
- Kibana is the visualisation layer for HMI (producing dashboards, tables, with pies, histograms, etc.).
- Logstash is the ETL for computing data through a pipeline including 3 main steps: input, filter, output according to its terminology. The pipelines must be created “by hand”, for example with a Logstash Editor plugin for Visual Studio Code (free), but a graphical tool exists, the Pipeline Viewer, that renders a graphical renderer of an existing pipeline.

ELK runs over a Java Runtime.

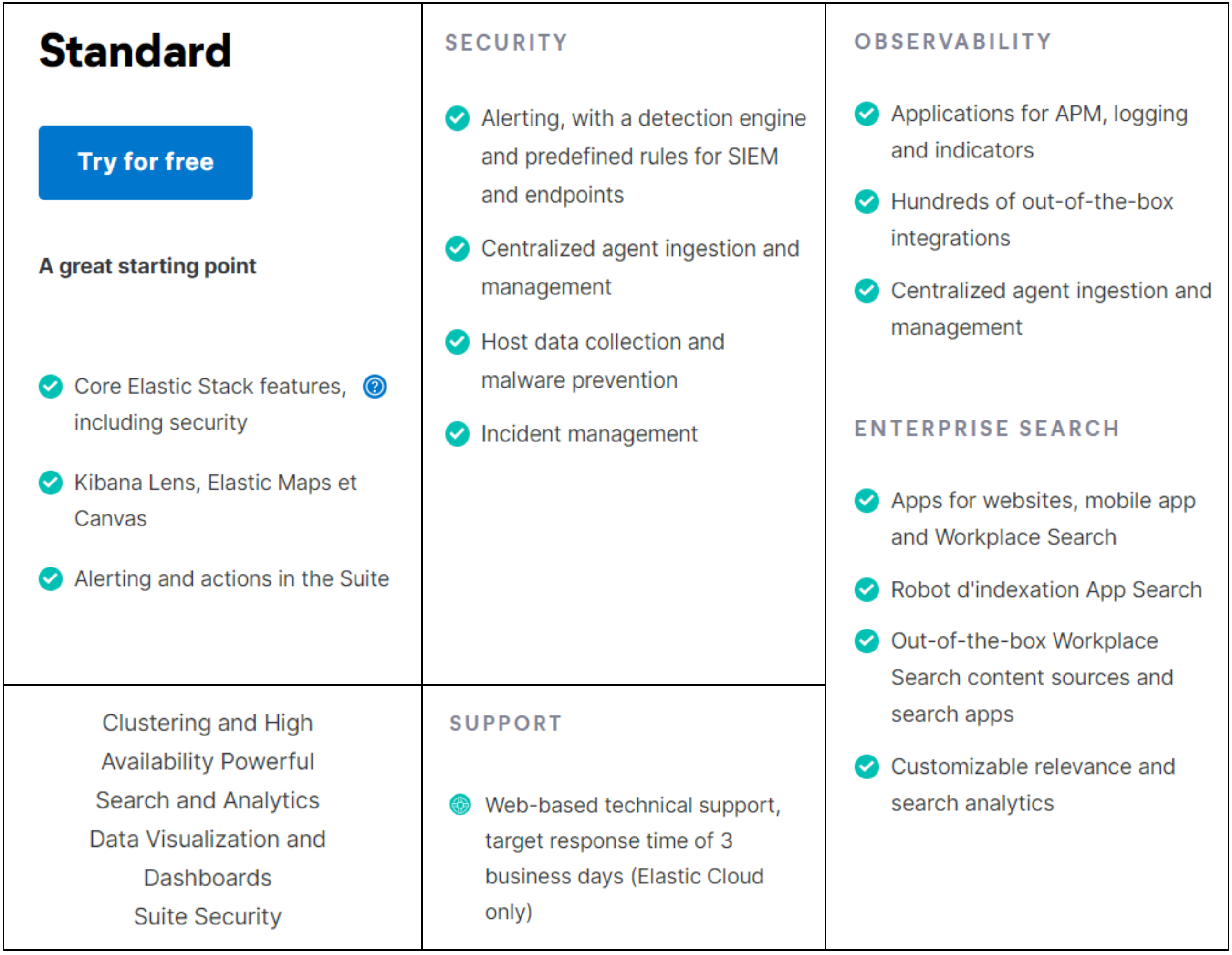
ELK is free but the Cloud part of Elastic must be paid.

The price of Elastic Cloud is the price of the deployment of Elastic and its availability over the Cloud. The price mainly depends upon the number of hours of the platform availability: you pay as soon as the platform is on, ready to accept activity; if the platform is started but not used (no activity on it), you pay anyway. The price includes the VMs cost, where are deployed the Elastic compute nodes.

Elastic Cloud embeds all the components needed for managing and securitizing the deployed platform.

It exists several Elastic Cloud offers (Standard, Gold, Platinum, Enterprise), but the Standard one includes all what DEQ project needs to run efficiently, even in terms of related features as monitoring and securitization.

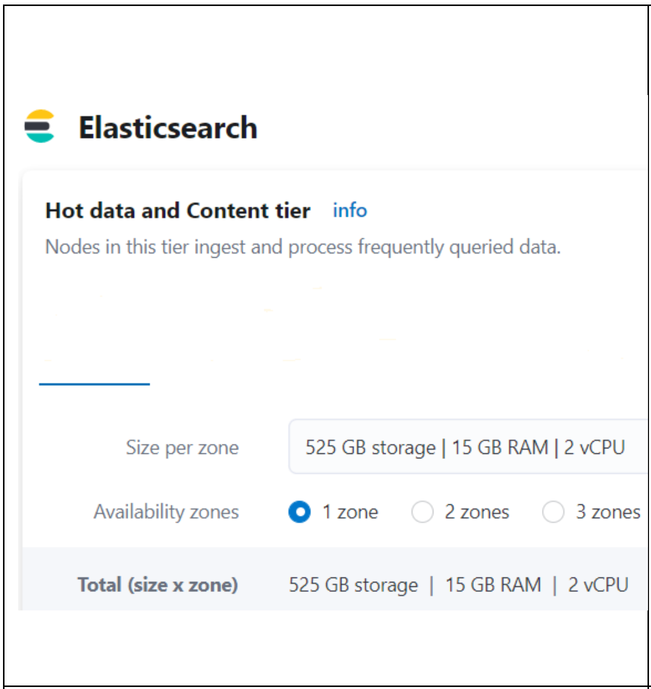
7.1.3.6.5.1 Elastic Cloud features for standard solution



Standard Try for free **A great starting point** ✓ Core Elastic Stack features, ? including security ✓ Kibana Lens, Elastic Maps et Canvas ✓ Alerting and actions in the Suite	SECURITY ✓ Alerting, with a detection engine and predefined rules for SIEM and endpoints ✓ Centralized agent ingestion and management ✓ Host data collection and malware prevention ✓ Incident management	OBSERVABILITY ✓ Applications for APM, logging and indicators ✓ Hundreds of out-of-the-box integrations ✓ Centralized agent ingestion and management ENTERPRISE SEARCH ✓ Apps for websites, mobile app and Workplace Search ✓ Robot d'indexation App Search ✓ Out-of-the-box Workplace Search content sources and search apps ✓ Customizable relevance and search analytics
Clustering and High Availability Powerful Search and Analytics Data Visualization and Dashboards Suite Security	SUPPORT Web-based technical support, target response time of 3 business days (Elastic Cloud only)	

Table 117: Elastic Cloud features for standard solution

7.1.3.6.5.2 Elastic Cloud metrics for standard solution over Azure

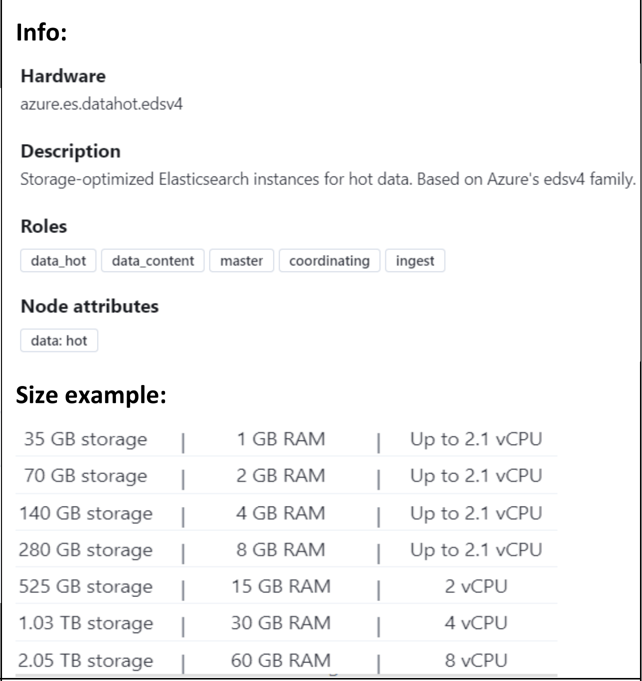


Elasticsearch

Hot data and Content tier info

Nodes in this tier ingest and process frequently queried data.

Size per zone	525 GB storage | 15 GB RAM | 2 vCPU
Availability zones	● 1 zone ○ 2 zones ○ 3 zones
Total (size x zone)	525 GB storage | 15 GB RAM | 2 vCPU



Info:

Hardware

azure.es.datahot.edsv4

Description

Storage-optimized Elasticsearch instances for hot data. Based on Azure's edsv4 family.

Roles

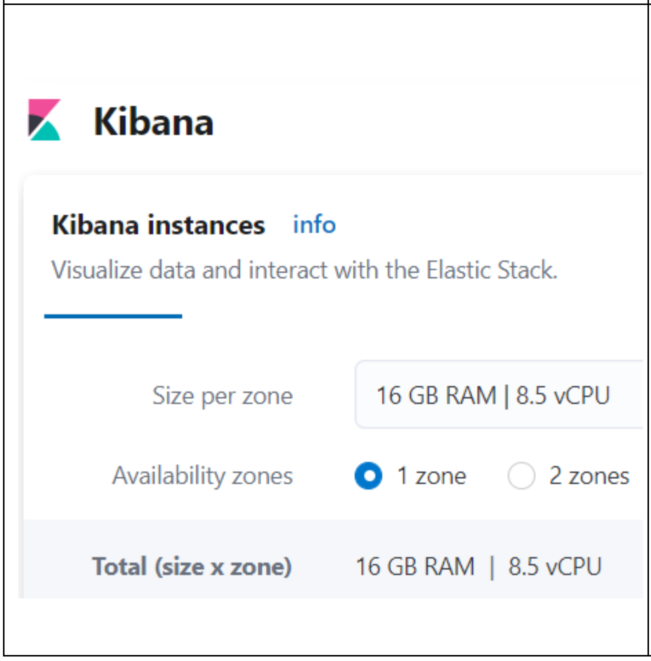
data_hot data_content master coordinating ingest

Node attributes

data: hot

Size example:

35 GB storage	1 GB RAM	Up to 2.1 vCPU
70 GB storage	2 GB RAM	Up to 2.1 vCPU
140 GB storage	4 GB RAM	Up to 2.1 vCPU
280 GB storage	8 GB RAM	Up to 2.1 vCPU
525 GB storage	15 GB RAM	2 vCPU
1.03 TB storage	30 GB RAM	4 vCPU
2.05 TB storage	60 GB RAM	8 vCPU

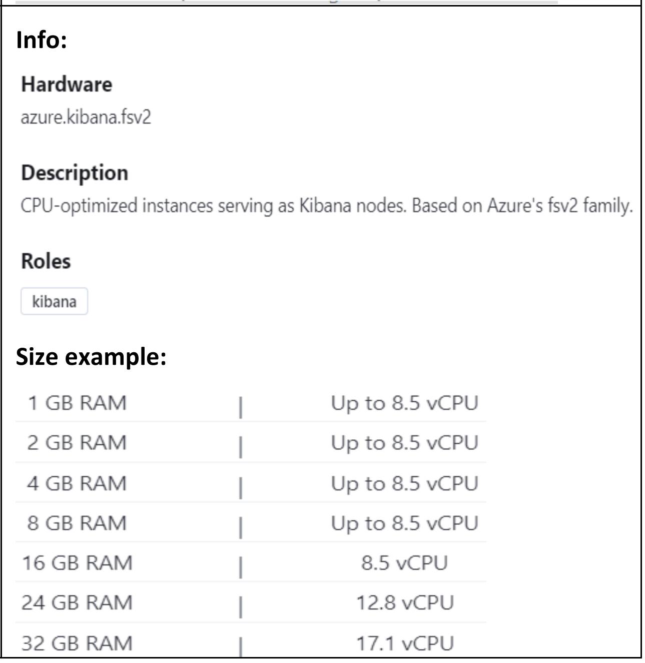


Kibana

Kibana instances info

Visualize data and interact with the Elastic Stack.

Size per zone	16 GB RAM | 8.5 vCPU
Availability zones	● 1 zone ○ 2 zones
Total (size x zone)	16 GB RAM | 8.5 vCPU



Info:

Hardware

azure.kibana.fsv2

Description

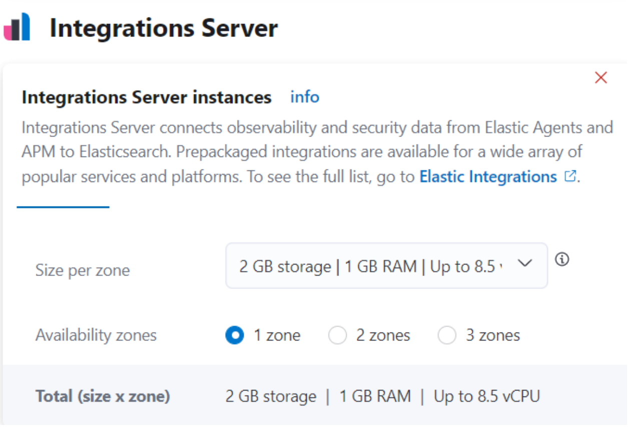
CPU-optimized instances serving as Kibana nodes. Based on Azure's fsv2 family.

Roles

kibana

Size example:

1 GB RAM	Up to 8.5 vCPU
2 GB RAM	Up to 8.5 vCPU
4 GB RAM	Up to 8.5 vCPU
8 GB RAM	Up to 8.5 vCPU
16 GB RAM	8.5 vCPU
24 GB RAM	12.8 vCPU
32 GB RAM	17.1 vCPU



Integrations Server

Integrations Server instances info

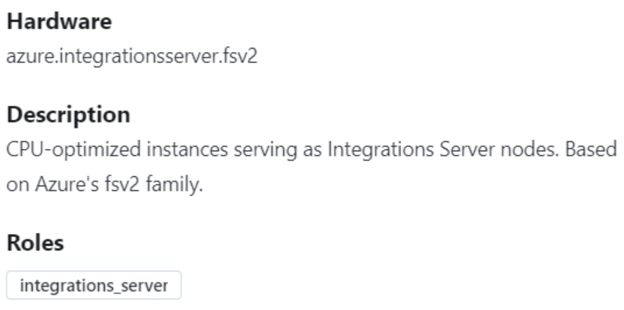
Integrations Server connects observability and security data from Elastic Agents and APM to Elasticsearch. Prepackaged integrations are available for a wide array of popular services and platforms. To see the full list, go to Elastic Integrations.

Size per zone: 2 GB storage | 1 GB RAM | Up to 8.5 ·

Availability zones: ● 1 zone ○ 2 zones ○ 3 zones

Total (size x zone) 2 GB storage | 1 GB RAM | Up to 8.5 vCPU

Info:



Hardware

azure.integrationsserver.fsv2

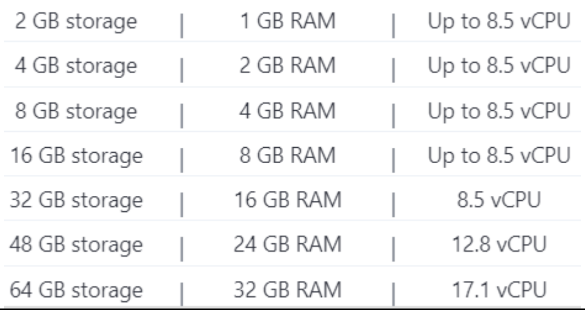
Description

CPU-optimized instances serving as Integrations Server nodes. Based on Azure's fsv2 family.

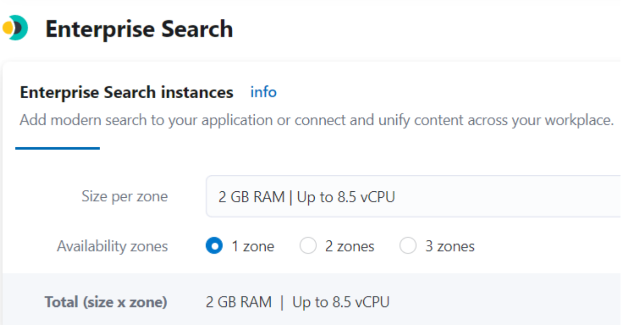
Roles

integrations_server

Size example:



2 GB storage	1 GB RAM	Up to 8.5 vCPU
4 GB storage	2 GB RAM	Up to 8.5 vCPU
8 GB storage	4 GB RAM	Up to 8.5 vCPU
16 GB storage	8 GB RAM	Up to 8.5 vCPU
32 GB storage	16 GB RAM	8.5 vCPU
48 GB storage	24 GB RAM	12.8 vCPU
64 GB storage	32 GB RAM	17.1 vCPU



Enterprise Search

Enterprise Search instances info

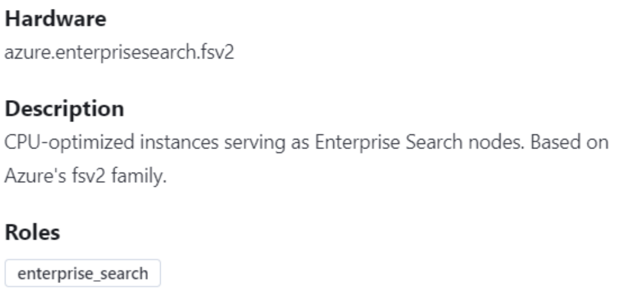
Add modern search to your application or connect and unify content across your workplace.

Size per zone: 2 GB RAM | Up to 8.5 vCPU

Availability zones: ● 1 zone ○ 2 zones ○ 3 zones

Total (size x zone) 2 GB RAM | Up to 8.5 vCPU

Info:



Hardware

azure.enterprisesearch.fsv2

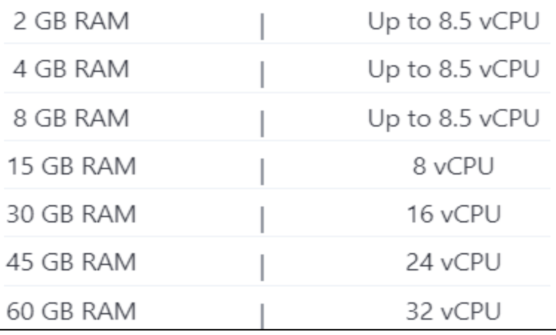
Description

CPU-optimized instances serving as Enterprise Search nodes. Based on Azure's fsv2 family.

Roles

enterprise_search

Size example:



2 GB RAM	Up to 8.5 vCPU
4 GB RAM	Up to 8.5 vCPU
8 GB RAM	Up to 8.5 vCPU
15 GB RAM	8 vCPU
30 GB RAM	16 vCPU
45 GB RAM	24 vCPU
60 GB RAM	32 vCPU

Table 118: Elastic Cloud component characteristics for standard solution over Azure

	Elastic Cloud tariffication for standard solution over Azure		
	Weak	Medium	Intensive
ELASTICSEARCH			
Hot storage	280 GB	525 GB	525 GB
Hot memory	8 GB	15 GB	15 GB
Hourly rate	0,23544€	0,44145€	0,44145€
INTEGRATIONS SERVER			
Memory	1 GB	1 GB	1 GB
Hourly rate	Free	Free	Free
KIBANA			
Memory	8 GB	8 GB	16 GB
Hourly rate	0,26208€	0,26208€	0,52416€
APM (Application Performance Monitoring)			
Memory	1 GB	1 GB	1 GB
Hourly rate	Free	Free	Free
ENTERPRISE SEARCH			
Memory	2 GB	2 GB	2 GB
Hourly rate	Free	Free	Free
TOTAL			
Total storage	280 GB	525 GB	525 GB
Total memory	20 GB	27 GB	35 GB
Hourly rate	0,4975€	0,7035€	0,9656€

Table 119: Elastic Cloud metrics for standard solution over Azure

Besides these prices above regarding the availability of the Elastic Cloud Service, some more fees must be added.

They concern:

- the storage size*: 0,0297 €/Go per month with 100 Go/month free
- the storage API requests**: 0,00162 € per 1 000 API calls (1,62 € per million API calls) with 100 000 API calls free

*This storage size does not impact the ElasticSearch storage (i.e., when the ES indexes are filled), but any storages out of the Elastic Cloud deployment (BLOB/File storage, databases storage on other VM, etc.).

**Note that the API calls that extract data from a storage hosted out of the Elastic cluster, are free of charge.

Moreover, some fees must be paid for the data transfer:

- Data sent out of the Elastic cluster (for example, to Internet): 0,032 €/Go per month with 100 Go/month free
- Data transferred inside the Elastic cluster (for example, to Kibana nodes for data rendering): 0,016 €/Go per month with 100 Go/month free
- Data transferred into the Elastic cluster (incoming data) are free of charge

7.1.3.6.5.3 Elastic Cloud prices for standard solution over Azure

7.1.3.6.5.3.1 DEQ storage fees

Recall: This storage does not concern the ElasticSearch indexes populating.

Consequence: As in DEQ project, Elastic suite is only used for generating BI dashboards and tables, very little data coming from Elastic cluster will be inserted into BLOB storage or database storage. Since 100 Go are free of charge per month, the cost of this storage can be declared as negligible for DEQ case.

Obviously, the same reasoning must be applied to the storage API requests (with 100 000 calls free).

7.1.3.6.5.3.2 DEQ data transfer fees

Usage	Outside Data Transfer		Inside Data Transfer		Total Price
	Metrics	Price (€)	Metrics	Price (€)	
Weak	50 Go	< 100 Go ==> **0**	75 Go	< 100 Go ==> **0**	**0 €**
Medium	100 Go	= 100 Go ==> **0**	150 Go	(150 – 100) x 0,016 = **1**	**1 €**
Intensive	200 Go	(200 – 100) x 0,032 = **3**	300 Go	(300 – 100) x 0,016 = **3**	**6 €**

Table 120: Data transfer fees for Elastic Cloud for standard solution over Azure

7.1.3.6.5.3.3 7/7 – 24/24

Usage	Deployment / Availability	Data Storage	Data Transfert	Total Price (€)
Weak	730 hours x 0,4975 = 363	–	0	**363**
Medium	730 hours x 0,7035 = 513	–	1	**514**
Intensive	730 hours x 0,9656 = 704	–	6	**710**

Table 121: Standard Elastic Cloud "7/7 - 24/24" price over Azure

7.1.3.6.5.3.4 5/7 – 24/24

Usage	Deployment / Availability	Data Storage	Data Transfer	Total Price (€)
Weak	530 hours x 0,4975 = 264	–	0	**264**
Medium	530 hours x 0,7035 = 373	–	1	**374**
Intensive	530 hours x 0,9656 = 512	–	6	**518**

Table 122: Standard Elastic Cloud "5/7 - 24/24" price over Azure

7.1.3.6.5.3.5 5/7 – 15/24 (from 5h00 to 20h00)

Usage	Deployment / Availability	Data Storage	Data Transfer	Total Price (€)
Weak	330 hours x 0,4975 = 164	–	0	**164**
Medium	330 hours x 0,7035 = 232	–	1	**233**
Intensive	330 hours x 0,9656 = 318	–	6	**324**

Table 123: Standard Elastic Cloud "5/7 - 15/24" price over Azure

7.1.3.6.5.3.6 5/7 – 8/24 (for example, from 8h30 to 12h30, then from 13h30 to 17h30)

Since a BI service may not have time constraints other than office hours, it could be possible to make it active only 8 hours a day. The tariffs would then come as follows:

Usage	Deployment / Availability	Data Storage	Data Transfert	Total Price (€)
Weak	176 hours x 0,4975 = 88	–	0	**88**
Medium	176 hours x 0,7035 = 124	–	1	**125**
Intensive	176 hours x 0,9656 = 170	–	6	**176**

Table 124: Standard Elastic Cloud "5/7 - 8/24" price over Azure

7.1.3.6.6 BI Solutions comparison

For Power Bi Embedded to have the same computing power in terms of CPU, RAM and storage, compared to the described Elastic Cloud solution, the price would be out of proportion.

At least, for 16 vCores (8 front-end, 8 back-end), 50 Go RAM, Azure charges 14,5 € per hour.

For 8 hours use a day with 5/7 days per month (means 176 hours per month), it comes 2 550 € per month.

On the other hand, Power BI Service is price competitive compared to Elastic Cloud, if a small number of licences must be purchased i.e., if only a few users request reports.

7.1.3.7 Summary of costs: IoT tariff for a selection of components

7.1.3.7.1 Component's Choice

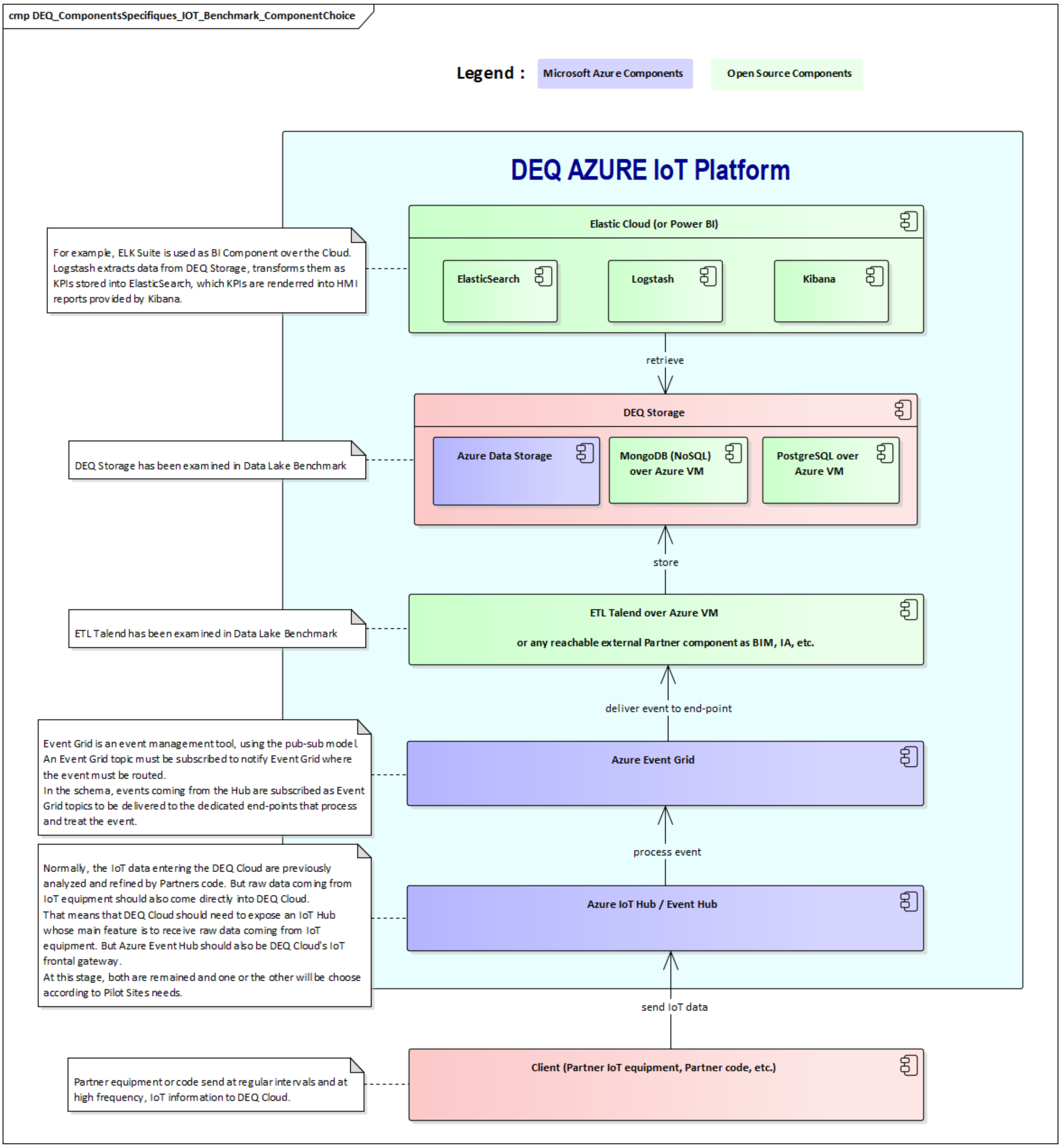
Here is presented a minimalistic design of what must be implemented as IoT components.

7.1.3.7.1.1 Design explanation

Component	Comment / Justification
Azure IoT Hub / Azure Event Hubs	The two services are similar in that they both support data ingestion with low latency and high reliability, but they are designed for different purposes. IoT Hub has been developed for connecting IoT devices to the Azure Cloud, while the Event Hubs service has been designed for streaming Big Data (mainly for hot computing). According to Pilot Sites needs, one or the other should be used.
Azure Event Grid	Event Grid is an event management tool, using the publish-subscribe model. An Event Grid topic must be subscribed to notify Event Grid where the event must be routed. For example, events coming from the Hub are subscribed as Event Grid topics to be delivered to the dedicated end-points that process and treat the event. Note that end-points can be hosted out of DEQ Cloud, as Partner components (BIM, AI, etc.)
ETL Talend	Talend is the ETL selected to transform any data coming into DEQ Cloud, before to be stored. For more information, see ETL Talend – Data Lake design explanation.
Elastic Cloud as BI Platform	It is assumed that Kibana can render efficiently and without restriction any KPI required for DEQ. Especially since all the KPIs do not have to be computed and displayed by this BI platform: many reports will be directly generated by Partners existing tools; even if these reports (or some of them) will be finally exposed by the frontal DEQ Gateway, they will not necessarily have been built by Elastic Stack.
Power BI Service as BI Platform	Power BI Service is well known and so mostly used to generate reports, that it cannot be dismissed and set it aside.

Table 125: IoT Components choice on Azure DEQ Cloud

7.1.3.7.1.2 Design schema


cmp DEQ_ComponentsSpecifiques_IOT_Benchmark_ComponentChoice
Legend :
Microsoft Azure Components
Open Source Components
DEQ AZURE IoT Platform
Elastic Cloud (or Power BI)
ElasticSearch
Logstash
Kibana
retrieve
DEQ Storage
Azure Data Storage
MongoDB (NoSQL) over Azure VM
PostgreSQL over Azure VM
store
ETL Talend over Azure VM
or any reachable external Partner component as BIM, IA, etc.
deliver event to end-point
Azure Event Grid
process event
Azure IoT Hub / Event Hub
send IoT data
Client (Partner IoT equipment, Partner code, etc.)

For example, ELK Suite is used as BI Component over the Cloud. Logstash extracts data from DEQ Storage, transforms them as KPIs stored into ElasticSearch, which KPIs are renderred into HMI reports provided by Kibana.

DEQ Storage has been examined in Data Lake Benchmark

ETL Talend has been examined in Data Lake Benchmark

Event Grid is an event management tool, using the pub-sub model. An Event Grid topic must be subscribed to notify Event Grid where the event must be routed.
In the schema, events coming from the Hub are subscribed as Event Grid topics to be delivered to the dedicated end-points that process and treat the event.

Normally, the IoT data entering the DEQ Cloud are previously analyzed and refined by Partners code. But raw data coming from IoT equipment should also come directly into DEQ Cloud.
That means that DEQ Cloud should need to expose an IoT Hub whose main feature is to receive raw data coming from IoT equipment. But Azure Event Hub should also be DEQ Cloud's IoT frontal gateway.
At this stage, both are remained and one or the other will be choose according to Pilot Sites needs.

Partner equipment or code send at regular intervals and at high frequency, IoT information to DEQ Cloud.

Figure 60: Benchmark IoT Components selection

7.1.3.7.2 Selected scenario evaluation

The price of this component selection is an aggregation of the tariffs computed by the previous paragraphs. The components that have already been charged in Data Lake Components (as ETL Talend) are not counted in this simulation.

IoT Frontal*	Event Grid	BI	TOTAL € / month
IoT Hub	40	**Elastic Cloud**	**< 250**
55		125	
Event Hub		**Power BI Service**	
15**		< 140	

Table 126: An idea of price for an IoT component's choice

*See the explanation "IoT Hub / Event Hub" at the previous paragraph to understand why both are maintained yet.

**The "Capture" feature of the Event Hub is not considered because it is redundant with the Event Grid which performs the events notifying ETL Talend for (hot) computing.

7.1.4 References

Document Resource ID	Document Resource name and reference
DR1	EU Grant Agreement n°101003750
DR2	D1.3 Requirements for Quarry full digitalisation (for Smart Sensors, Automation &Process Control, and for ICT solutions, BIM and AI report

List of Figures

List of Tables