The Open Group Guide

Information Architecture
Business Intelligence & Analytics and Metadata Management Reference Models

The Open Group Guide

Information Architecture: Business Intelligence & Analytics and Metadata Management Reference Models

ISBN: 1-947754-46-1

Document Number: G201

Published by The Open Group, January 2020.

Comments relating to the material contained in this document may be submitted to:

The Open Group, Apex Plaza, Forbury Road, Reading, Berkshire, RG1 1AX, United Kingdom

or by electronic mail to:

ogspecs@opengroup.org

Contents

Preface

The Open Group

The Open Group is a global consortium that enables the achievement of business objectives through technology standards. Our diverse membership of more than 700 organizations includes customers, systems and solutions suppliers, tools vendors, integrators, academics, and consultants across multiple industries.

The mission of The Open Group is to drive the creation of Boundaryless Information Flow™ achieved by:

- Working with customers to capture, understand, and address current and emerging requirements, establish policies, and share best practices
- Working with suppliers, consortia, and standards bodies to develop consensus and facilitate interoperability, to evolve and integrate specifications and open source technologies
- Offering a comprehensive set of services to enhance the operational efficiency of consortia
- Developing and operating the industry's premier certification service and encouraging procurement of certified products

Further information on The Open Group is available at www.opengroup.org.

The Open Group publishes a wide range of technical documentation, most of which is focused on development of Standards and Guides, but which also includes white papers, technical studies, certification and testing documentation, and business titles. Full details and a catalog are available at www.opengroup.org/library.

This Document

This document is The Open Group Guide to Information Architecture. It has been developed and approved by The Open Group.

This document is structured as follows:

- Chapter 1 highlights the pressure upon traditional companies to adapt to a changing IT landscape in which citizens increasingly expect a more central and autonomous role in accessing products and services, leading to more emphasis on data and its management
- Chapter 2 presents the ArchiMate® modeling notation as a key asset to effectively develop good Data Architecture practice, based upon a common modeling language
- Chapter 3 presents the TOGAF® ADM as the basis for a step-by-step approach for companies to follow in order to achieve their goal of becoming data-driven

- Chapter 4 outlines, using the Business Intelligence & Analytics Reference Model, the objectives and goals that an organization would target in order to become data-driven
- Chapter 5 provides a list of the application functions that compose Business Intelligence & Analytics and Metadata Management capabilities
- Chapter 6 further describes the functions of the Business Intelligence & Analytics Reference Model
- Chapter 7 ensures ease-of-use by placing the functions, with more detail added, into alphabetical order

The following typographical convention is used to highlight key terminology and/or concepts:

Key Term/Concept
An explanation of the key term/concept.

Trademarks

Acknowledgements

The Open Group gratefully acknowledges the contribution of the following people in the development of this Guide:

AXA Data Architecture Community

This Guide is inspired from Data Architecture deliverables donated to The Open Group by AXA. They are the results of the Data Architect community's work led by the group and with the contribution of affiliates mainly in France, Germany, Belgium, United Kingdom, Japan, US, Hong Kong, Spain, Switzerland, and Italy from 2015 to 2019.

AXA is a global insurer (www.axa.com).

Céline Lescop – Lead Author

Céline holds a Master's Degree (1996) in Computer Science (www.ensiie.fr). She was a Data Project Leader in the pharmaceutical industry from 1999 until 2012 (Galderma & Astrazeneca), an Architect and TOGAF trainer from 2012 to 2015 at Arismore (The Open Group, France), and has been an Enterprise Architect at AXA leading the Data Architecture practice since 2016.

Guillaume Corve – Contributor

Guillaume holds a Master's Degree (1998) in Mechanical Design (ENSAM: https://artsetmetiers.fr/). He has been a Solution Architect at AXA since 2015.

Guillaume Hervouin – Contributor

Guillaume holds a Bachelor's Degree (2002) in Statistics and Business Intelligence. He was a Solution Architect from 2002 to 2015 at Capgemini, and has been a Data Architect at AXA since 2016.

Jean-Baptiste Piccirillo – Contributor

Jean-Baptists holds a Master's Degree (2010) in Computer Science (www.ensiie.fr). He was an Architect until 2015 for Arismore (The Open Group, France), a Data Scientist for AXA from 2015 to 2018, and he is currently a Data Architecture and Data Management Consultant for RHAPSODIES CONSEIL (www.rhapsodiesconseil.fr) supporting AXA.

Referenced Documents

The following documents are referenced in this Guide.

(Please note that the links below are good at the time of writing but cannot be guaranteed for the future.)

- ArchiMate® 3.1 Specification, a standard of The Open Group (C197), published by The Open Group, November 2019; refer to: www.opengroup.org/library/c197
- ArchiMate® Model Exchange File Format for the ArchiMate Modeling Language, Version 3.1, a standard of The Open Group (C19C), published by The Open Group, November 2019; refer to: www.opengroup.org/library/c19c
- DAMA DMBOK V2: Data Management Body of Knowledge, Version 2, published by Technics Publications, July 2017; refer to: https://dama.org/content/body-knowledge
- The Open Group IT4IT™ Reference Architecture, Version 2.1, a standard of The Open Group (C171), published by The Open Group, January 2017; refer to: www.opengroup.org/library/c171
- The TOGAF® Standard, Version 9.2, a standard of The Open Group (C182), published by The Open Group, April 2018; refer to: www.opengroup.org/library/c182

1 Introduction

1.1 Web Giants and their Impact

Since the 1980s the usage of digital technologies has increased. This expansion in use is perceived as essential for economic and social development.

The digital transition is said to be critical for countries and companies with digital objects and interfaces gradually becoming part of every aspect of our social life. For both states and organizations it is considered a key capability in the worldwide market.

During the last decade, this digitalization trend of society was reinforced by the web giants' new services offer. They have disrupted sectors like advertisement, retail, human communities, cell-phone market and people communication channels, hospitality services, ride services, and video distribution. This disruption has definitively increased the autonomy and expectation for autonomy of citizens in their access to services provided by organizations. It impacts on traditional companies that must shift from a product-centered business model to a customer-centric one. Customer data that used to be spread across systems must be managed to enable personalized interactions and services. This data transformation significantly impacts both on operational processes dealing with customers and on the IT landscape itself.

Physical persons or organizations, also called moral persons, exist by themselves with their data becoming more and more subject to regulations, such as for instance General Data Privacy Regulations (GDPR) in Europe. To provide them with a completely secure set of services, traditional corporations must organize partnerships and a way to interoperate in a multi-corporation scope. The data of individuals, physical or moral, must flow as a coordinated service chain across organizations, public or private, that provide services.

To deliver services at worldwide scale, these web giants have also had to extend the capabilities of available IT platforms to extend storage, computing, and availability. These innovations have also disrupted the IT landscape by making cloud services relying on distributed architectures available on the market, providing almost unlimited capabilities.

Traditional organizations, industries, services, and administrations must adapt to that new context. They must leverage their customer relations and data assets to be able to serve the new expectations of their current customers, protect their market share from new competitors, and maintain an ability to conquer new markets.

This happens in a context where their information system:

- Was initiated in the 1980s and includes several hundreds of applications and data flows using different generations of technologies and is concerned by technical debt
- Was built to serve siloed services with dedicated data centers in dedicated networks
- Is now supporting almost all processes of the organization

This information system must shift towards:

- Interoperability internally and externally
- Worldwide platforms on the cloud
- Ability to scale up and down
- Ability to leverage digital innovations

Web Giants

The term refers to companies leveraging the worldwide web to provide global services platforms that drive large societal change, rather than just the largest tech companies.

They have disrupted sectors like advertisement, retail, human communities, cell-phone market and people communication channels, hospitality services, ride services, and video distribution.

1.2 From an Application to a Data-Driven Architecture

Architects must tackle this complexity to support the right decisions about the architecture of an organization in a context where technology is currently changing massively and where providers are diversifying:

- 1980-1990s was the time of specific development
- 2000-2015 was the time of application-driven architecture relying on software approaches

  It has allowed democratization of digital, fast digitalization of business processes with out-of-the-box packages, but it has created siloes at the same time.
- Since 2015 and the rise of Big Data, a shift towards data-driven architecture has been started by traditional organizations that want to become data-driven companies

  This comes with a will to break the application siloes, leveraging transversal data management systems like Master Data Management (MDM) that manage data shared across the business lines, or data warehouse and data lake that are supposed to gather all of the data from an organization to feed analytics.

Those who have worked on master data or Business Intelligence (BI) know that data persists even when technologies evolve. Data, when considering the main business objects managed by a given organization, are few compared to applications, processes, or technologies. They provide architects with a stable and concentrated point of view on the architecture across the information system. For historical reasons, Enterprise Architecture practices have not exploited that much this point of view and have mainly concentrated on Application or Technology Architecture. For a while, Data Architects have not gone beyond BI or MDM siloes. Today, applications and technologies evolve very quickly and are now owned by external companies that provide infrastructure or application services – data happens to be the unique, stable part of the architecture owned by the organization.

Data is also a key asset to better serve customers and allow employees, agents, and partners to adapt their capabilities and provide new customer experiences. Hence organizations now aiming to become data-driven.

1.3 Focus on Data Management Capabilities

For most organizations of a given size, all components of the value chain have been digitalized and are supported by an information system. Figure 1 describes the set of capabilities provided by Information Technology (IT) departments to support business operations – divided into six areas and organized like a value chain, splitting support activities and primary activities:

- **Internal IT capabilities:**

  Core business primary activities supporting the value chain: includes the key capabilities that support the organization value chain from receptions of orders to processing and delivery. This is where the income is captured.

  Engagement: digital interactions with the value chain. This includes the key capabilities that are used to interact with the customer, business users, and partners. It covers the digital, service center, distribution, and partner channels Business to Business to Consumer (B2B2C). This category also includes the Customer Relationship Management (CRM) marketing, servicing, and sales as key enablers for this layer.

  Data management support: becoming a data-driven company means a shift to a data-driven architecture approach. It leads to a specific focus on data management capabilities that support the business and that are the responsibility of the Chief Data Officer. These data management capabilities either concentrate a large amount of data or play a key role in data management processes. They are quoted in the DAMA DMBOK V2 (see Referenced Documents).

  Business support: includes other supporting capabilities that are required to support the value chain, like human resources, finance, the IT4IT™ Reference Architecture, workplace, etc.

- **External IT capabilities:**

  External data providers capabilities: data providers that can be open data or data acquired externally that the organization is leveraging.

  External partners capabilities: services with a digital front provided by other organizations on the market that must be combined with the digital services of the organization to deliver a more complete experience to customers.

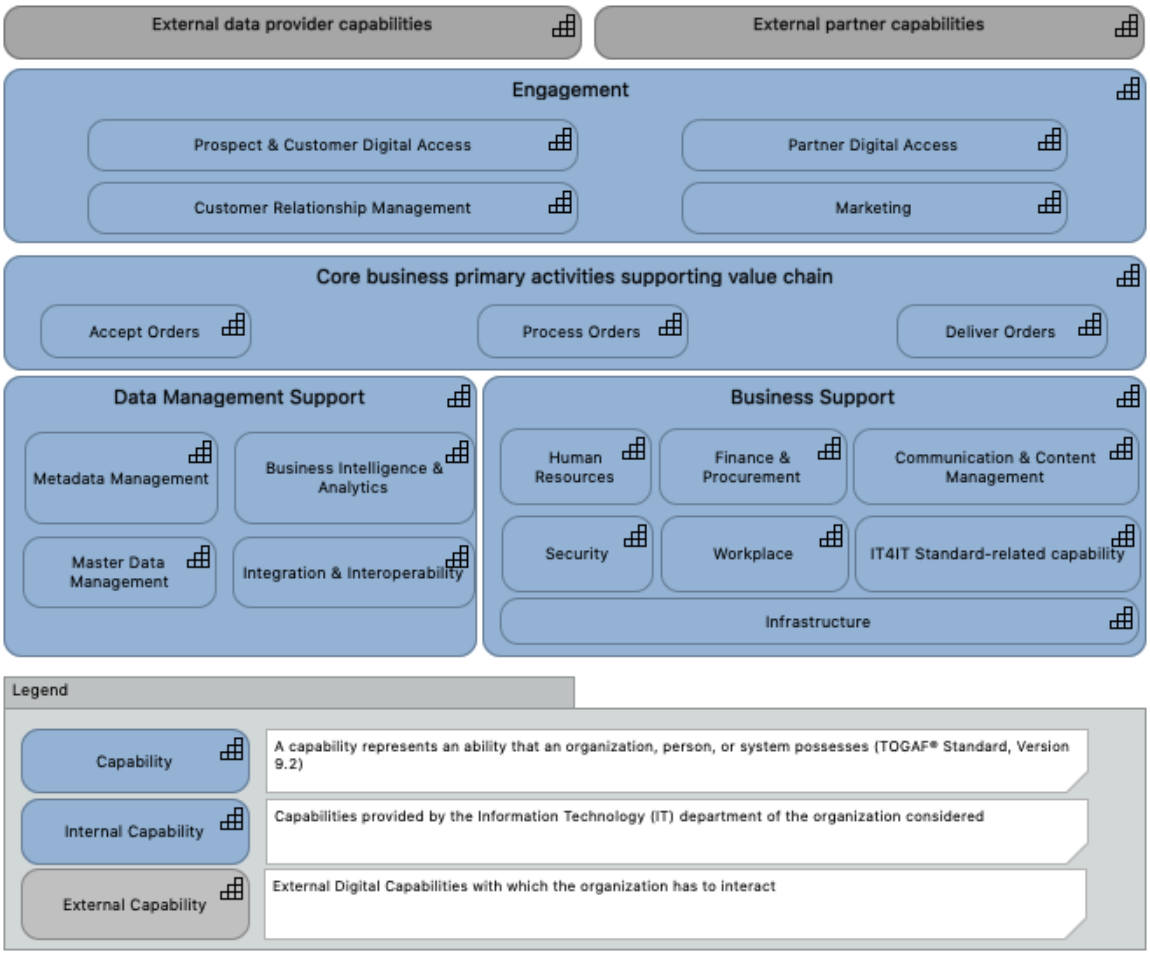


Figure 1: Generic Description of the Capabilities Delivered by the IT Department

As shown in Figure 2, the IT department capabilities are mapped with the DAMA-DAMBOK roles (see Referenced Documents).

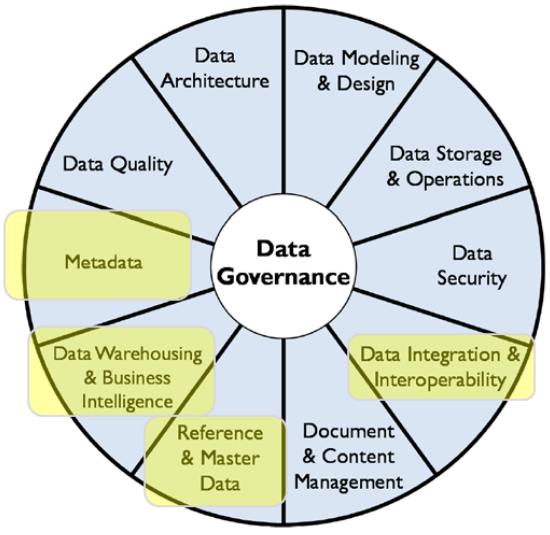


Figure 2: Mapping of the "Data Management Support" Capabilities to the DAMA DMBOK V2 Data Management Framework Knowledge Areas (aka DAMA Wheel)
[Source: DAMA DMBOK V2]

BI & Analytics

A capability where all of the data of a given organization is made available to anyone removing operational application siloes, constraints, and making decisions relying on facts. It covers the capabilities of data warehouses, data marts, BI, data lake, and data sciences. Indeed, these systems propose to make available all the data related to the organization, including the current and historical values in a single place. It aims at serving data to stakeholders needing reports for monitoring or regulatory purposes, queries, or more advanced processing like statistics or data science.

Master & Reference Data Management

These are capabilities that refer to processes, organizations, people, and systems put together to measure and improve continuously the value of master and reference data for an organization, including business value, data quality, availability, and security. They deliver the true customer data to relevant use-cases.

Master Data

This is data shared by several teams across the organization. Master data is related to the objects that interact when a transaction occurs. It can include products, employee, customers, or assets.

Reference data is any kind of data that is used solely to categorize other data found in a database, or solely for relating data in a database to information beyond the boundaries of the enterprise.

Reference Data

This defines the set of permissible values to be used by other data fields. Reference data is often defined by standards organizations (such as country codes as defined in ISO 3166-1: The International Standard for Country Codes and Codes for their Subdivisions. (Refer to https://www.iso.org/iso-3166-country-codes.html.)

Data Integration & Interoperability

Data Integration & Interoperability (DII) describes processes related to the movement and consolidation of data within and between data stores, applications, and organizations. Integration consolidates data into consistent forms, either physical or virtual. Data interoperability is the ability for multiple systems to communicate.

Tools that support this capability include: Enterprise Service Bus (ESB), Extract, Transform, Load (ETL), and Application Programming Interface (API).

Metadata Management

Metadata is data about data. Metadata describes the data itself (e.g., databases, data elements, and data models), the concepts the data represents (e.g., business processes, application systems, software code, and technology infrastructure), and the connections (relationships) between the data and concepts. Metadata data management can also be called knowledge about data management.

In this document we will focus on a reference model with application functions to be covered by BI & Analytics and Metadata Management capabilities from graphic representations to requirements. It aims to become a stable and common language to describe the functions that those platforms should cover. It can be used:

- To describe the functional coverage of a given provider
- To compare the functional coverage of two providers
- Assess the as-is architecture of a given organization
- Design the to-be architecture of a given organization
- To evaluate the maturity of a capability
- In Architecture Governance to decide on a scenario for the Target Architecture

This reference model provides descriptions of functions in the application layer to be covered for these capabilities.

2 Methodology

The ArchiMate modeling notation (see Referenced Documents) is a key asset for a rigorous Data Architecture practice. The architecture artifacts proposed in this document use it as a common modeling language.

An ArchiMate tool compliant with the ArchiMate Model Exchange File Format (see Referenced Documents) was chosen to create ArchiMate models.

Artifacts of different levels of detail are available through ArchiMate diagrams. They provide a consistent reference view that allows architects to zoom in or out depending on the context and the stakeholders involved in the architecture design.

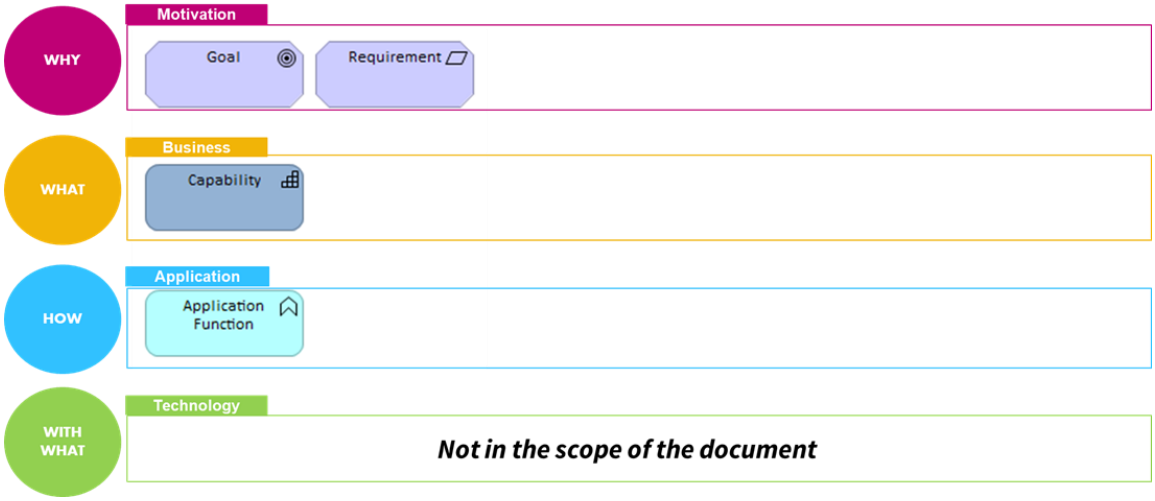


Figure 3: The ArchiMate Modeling Notation Elements Used in this Document

To make it easier for the reader, a selection of ArchiMate definitions is provided below:

Goal

A goal represents a high-level statement of intent, direction, or desired end state for an organization and its stakeholders.

Requirement

A requirement represents a statement of need that must be met by the architecture.

Capability

A capability represents an ability that an organization, person, or system possesses.

Application Function

An application function represents automated behavior that can be performed by an application component.

3 Data Architecture and The Open Group Standards

3.1 The TOGAF Architecture Development Method (ADM)

The Boundaryless Information Flow vision of The Open Group is about promoting standards to allow the flow of data across and between organizations.

The TOGAF standard (see Referenced Documents) dedicates a full chapter to Data Architecture: "Phase C: Information Systems Architectures – Data Architecture".

Data Architecture (TOGAF Definition)

A description of the structure and interaction of the enterprise's major types and sources of data, logical data assets, physical data assets, and data management resources.

This covers two domains of the DAMA DMBOK V2:

Data Architecture (DAMA DMBOK V2 Definition)

Data Architecture identifies the data needs of the enterprise (regardless of structure), and designs and maintains the master blueprints to meet those needs. Master blueprints are used to guide data integration, control data assets, and align data investments with business strategy.

Data Modeling

Data modeling is the process of discovering, analyzing, and scoping data requirements, and then representing and communicating these data requirements in a precise form called the data model. This process is iterative and may include a conceptual, logical, and physical data model.

Organizations aiming to become "data-driven" will have to consider data at each step of the ADM process.

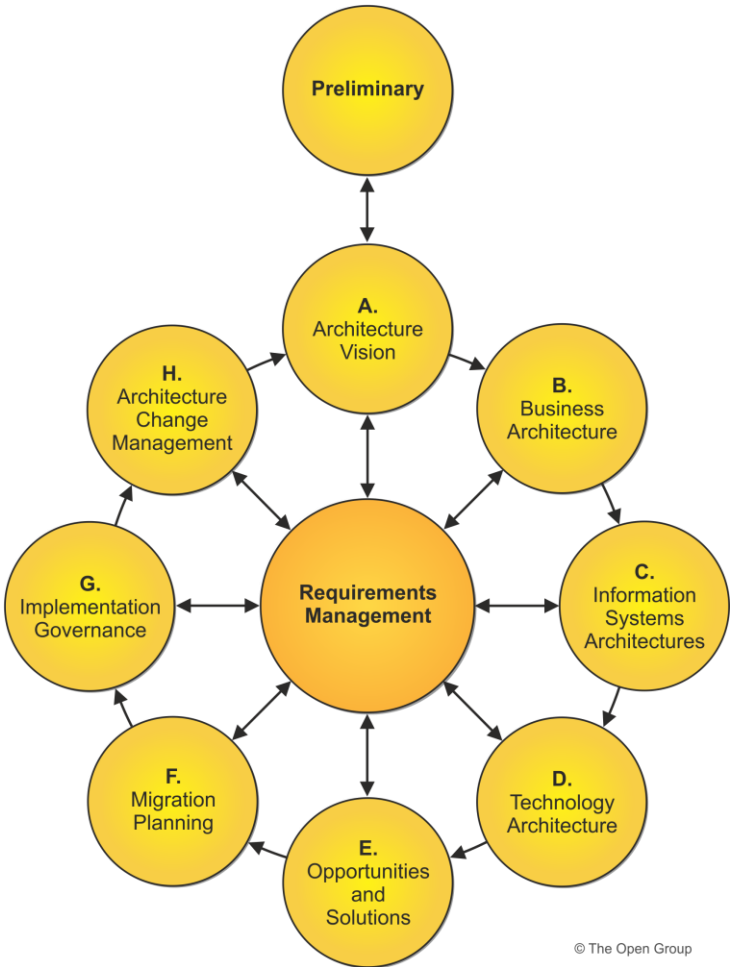


Figure 4: The TOGAF Architecture Development Method

- **Preliminary Phase**:

  — Choose a framework for the Data Architecture practice; possible frameworks include DAMA DMBOK V2, and the TOGAF and ArchiMate standards

  — Choose a reference data model; for example, ACORD® (insurance sector)

  — Agree on data principles

- **Phase A – Architecture Vision**:

  — Define and get consensus on the data management vision and target data management capability

- **Phase B – Business Architecture**:

  — Business Information Model (BIM), data classifications

  — Requirements related to data management

- **Phase C – Information Systems Architectures – Application Architecture**:

  — Identify major applications to support the data-driven company

— Focus on applications that play a key role for the data management in an organization: MDM for the management of data that is shared across the organization, and BI & Analytics that aim to make the data of operational systems available for reporting, analytics (statistics, data science), DII, and metadata

- **Phase C – Information Systems Architectures – Data Architecture**:

  — A logical data model is derived from the BIM detailing keys, main business fields, and main links; cardinality and a full list of fields are not required at this stage

  — Related Content Metamodel building block:

    — Data Entities Components: an encapsulation of data that is recognized by a business domain expert as a thing

    — Logical Data Components: a data structure composed of logically-related data entities

    — Physical Data Components: a data structure that realizes related logical data components represented in the format or schema required by a particular technology

  — Data security, integrity, and quality are evaluated, and design decisions taken

- **Phase D – Technology Architecture**:

  — Storage mode: Relational Database Management System (RDBMS), document databases, and graph databases, etc.

  — Network integration, especially in the case of an architecture relying on multiple data centers hosted by the organization or the cloud provider

  — Description of technical services for data management: encryption, backup/restore

When practicing Data Architecture, the level of detail should be adjusted to deal with architecture decisions that have to be taken before starting the projects. Architecture decisions are those that have a major impact on the architecture: requirement covered, the budget, and the change impact on the business. For the sake of architecture process efficiency, architecture descriptions should not go into detail that is not required for the architecture decision process. Data modeling practices that are built to support the design of software are going into too much detail for the architecture description. The Data Architect can use them but must choose the right level of detail. Indeed, modeling data in too much detail will delay the architecture process and block the project with designs that may not fit. For most architecture decisions, business object models that list the major data entities to be managed, their key, definitions, and main links should be enough. Physical data models are most of the time too detailed. Nevertheless, as always, the right level of detail must be adapted depending on the context.

3.2 The ArchiMate Modeling Notation

The ArchiMate standard provides the following elements to support Data Architecture descriptions. They allow the representation of data objects at business, application, and technical levels but fields of data objects or links between objects cannot be easily represented.

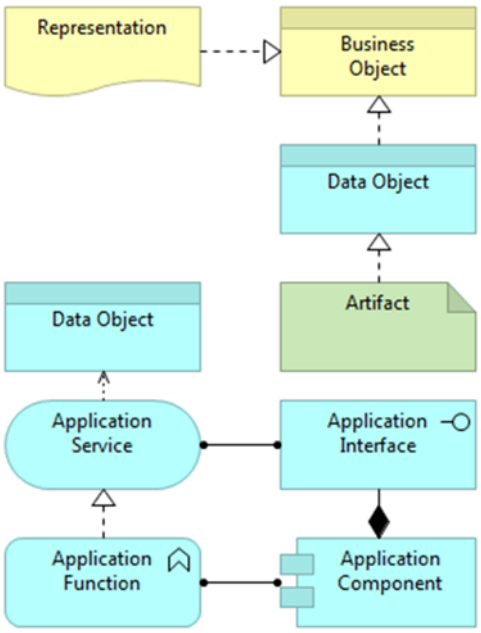


Figure 5: The ArchiMate Modeling Notation

Business Layer/Passive Structure/Business Object

A business object represents a concept used within a particular business domain.

Business Layer/Passive Structure/Representation

A representation represents a perceptible form of the information carried by a business object.

Application Layer/Passive Structure/Data Object

A data object represents data structured for automated processing.

Technology Layer/Passive Structure/Artifact

An artifact represents a piece of data that is used or produced in a software development process, or by deployment and operation of an IT system.

4 BI & Analytics Reference Model – Goals

In this chapter we describe the objectives that the organization is targeting with BI & Analytics. Those objectives are then used when describing requirements in Chapter 6.

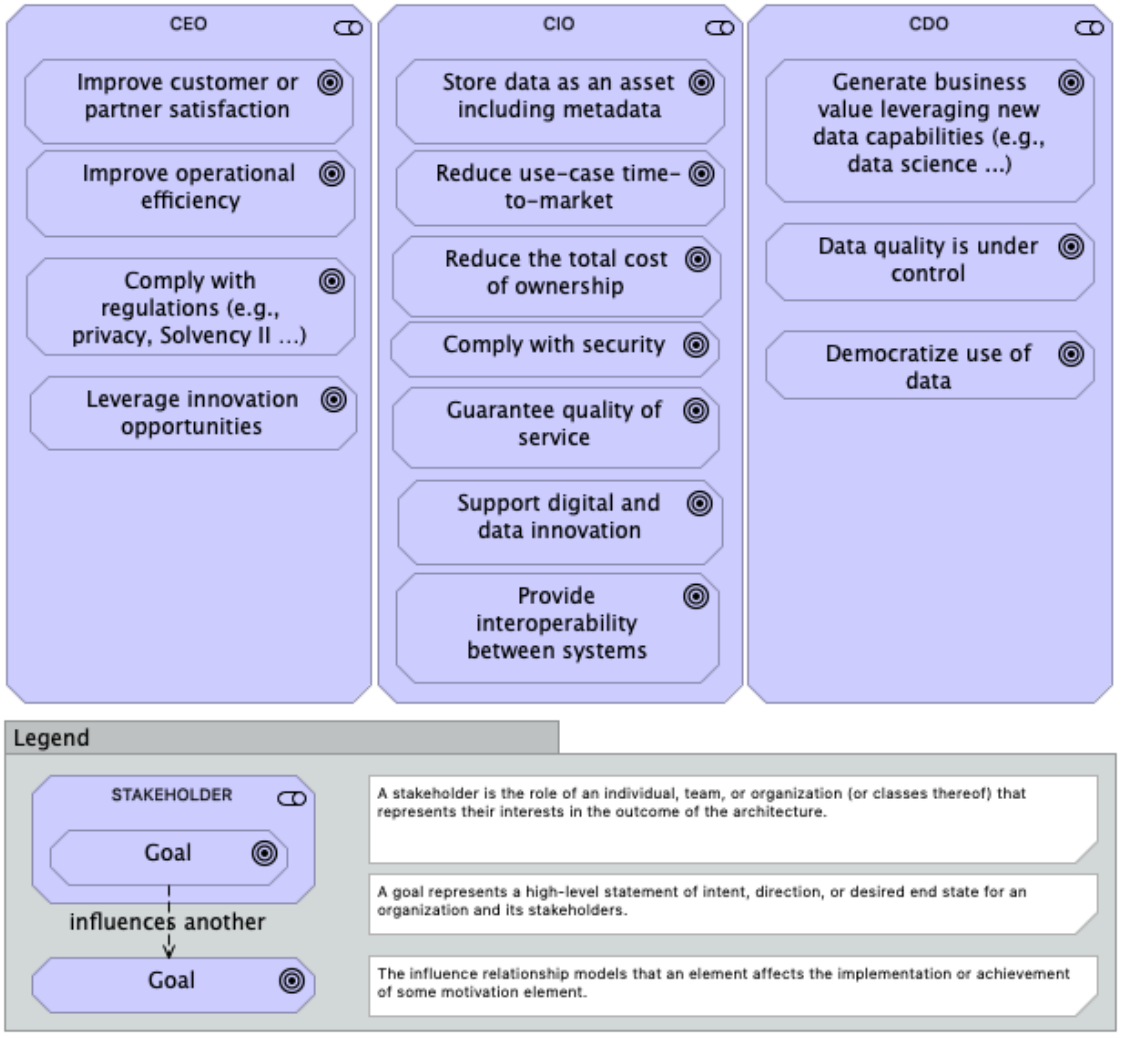


Figure 6: Goals Targeted by Stakeholders with BI & Analytics Capability

Table 1: Objectives and Goals of the BI & Analytics Reference Model

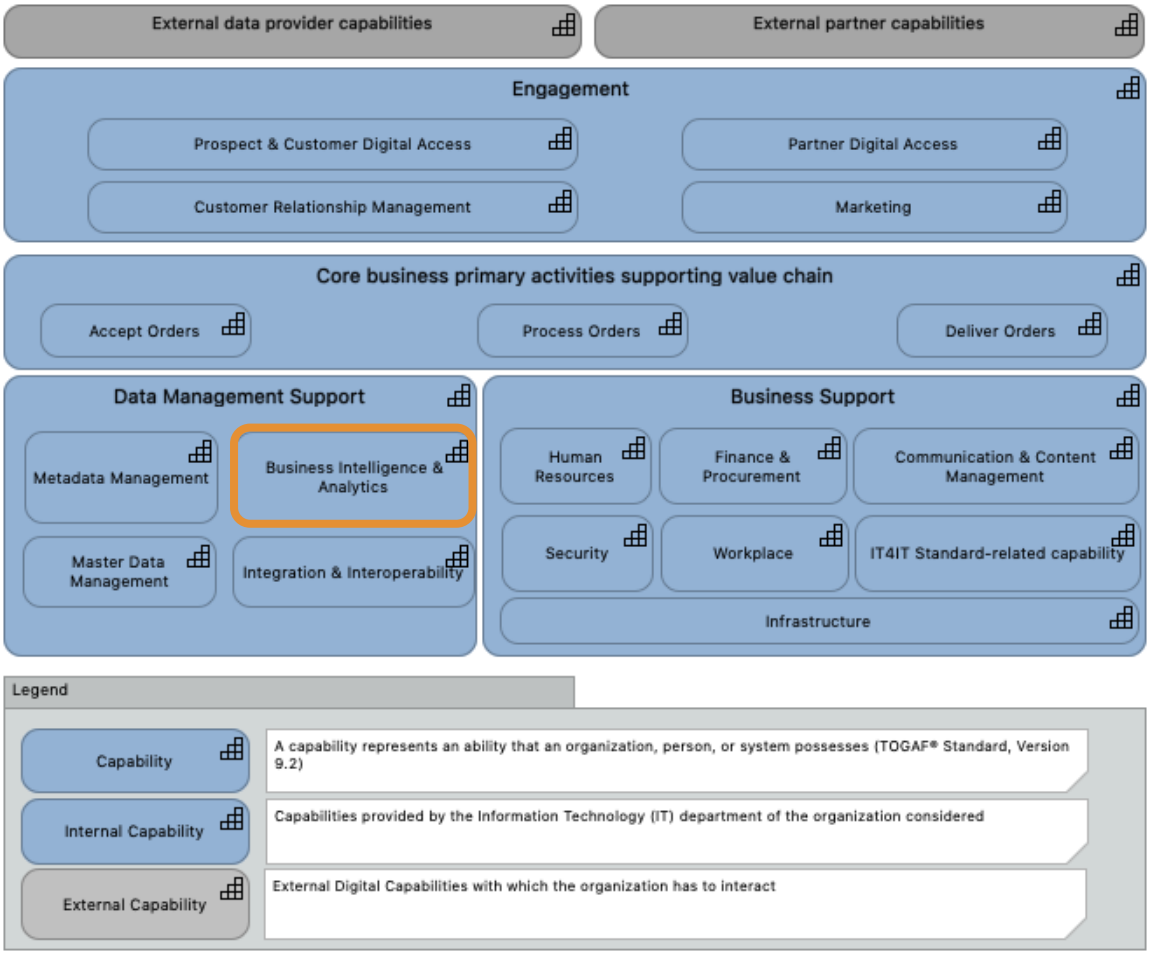
ID	Goals	Description	KPI	Number of Related Requirements (see Chapter 7)
1	Comply with regulations (e.g., privacy, Solvency II, etc.)	Some regulations impact data management: • Data privacy regulation aims to guarantee customers' data ownership – all personal data needs to be sufficiently controlled to ensure its necessity, protection, restitution, and deletion (audits should be runnable on the system); in Europe this refers to European GDPR • Solvency II concerns the amount of capital that European Union insurance companies must hold to reduce the risk of insolvency	↑ number of datasets tagged with the level of privacy ↓ data leakages	5
2	Comply with security – data classification and encryption	Data is classified as secret, confidential, and public depending on its sensitivity and the evaluation of associated risk. It is then encrypted with encryptions keys that are managed depending on this classification.	↓ security breaches	48
3	Comply with security – data leakage	Data should be protected from eavesdropping, leakage, or theft according to its classification.	↓ security breaches	14
4	Comply with security – identity and access management	Management of electronic or digital identities, including the organizational policies for managing digital identity as well as the technologies needed to support identity management.	↓ security breaches	26
5	Comply with security – infrastructure baseline	Minimal security baseline to be provided for the infrastructure (administrator password changes on material delivered, port management, etc.).	↓ security breaches	34
6	Data quality is under control	Data quality should remain at a suitable level to ensure the operational performance of the organization.	↑ number of datasets monitored by business quality rule	21

ID	Goals	Description	KPI	Number of Related Requirements (see Chapter 7)
7	Democratize use of data	Users should have access to shared semantics and data required to perform their duties through tools appropriate to their analytics requirements – from ready-to-read supports to advanced queries or analytics. This principle of data democratization will continually "bump up against" the principle of data security. Under no circumstances will the data sharing principle cause confidential data to be compromised.	↑ number of decisions taken under data inputs	17
8	Guarantee quality of service	Includes availability; data should be available when needed and as appropriate.	↑ improve system time availability	50
9	Improve customer or partner satisfaction	Use-cases and services that aim to increase customer or partner satisfaction.	↑ improve Net Promoter Score (NPS)	2
10	Improve operational efficiency	Use-cases and services that aim to increase agility in the organization, from innovation to operations.	↑ automated functions	75
11	Reduce the Total Cost of Ownership (TCO)	Increase cost efficiency.	↓ cost per user	11
12	Reduce use-case time-to-market	Setting up new use-cases in an agile manner to provide competitive advantage. This includes self-service for different stakeholders.	↓ delay from prototype validation to production running	32
13	Store data as an asset, including metadata	Valuable data owned by AXA™ should be kept in safe storage and with appropriate documentation, and be available, understandable, and usable (qualitative, meaningful, query-able, process-able, etc.) for transversal analytics.	↑ number of permanent connections	68
14	Generate business value leveraging new data capabilities (e.g., data science, etc.)	New data capabilities come with the promise of creating new business value and new markets. Organizations have to be able to compete in the international market leveraging new data capabilities. CDO positions have been created to make it happen.	↑ value generated by data use-cases	5
15	Provide interoperability between systems	Allow break of organizational silos by ensuring application interoperability	↑ number of APIs available and used	27

ID	Goals	Description	KPI	Number of Related Requirements (see Chapter 7)
16	Support digital and data innovation	Provide the digital capabilities to support innovation.	↑ innovation deployed at scale with P&L impact	7

5 BI & Analytics Reference Model – High-Level

This artifact is consistent with the ArchiMate metamodel and provides a high-level vision of the application functions that compose BI & Analytics and Metadata Management capabilities integrated in the full landscape. The purpose is to have on a single page the full landscape description with a focus on BI & Analytics application services.



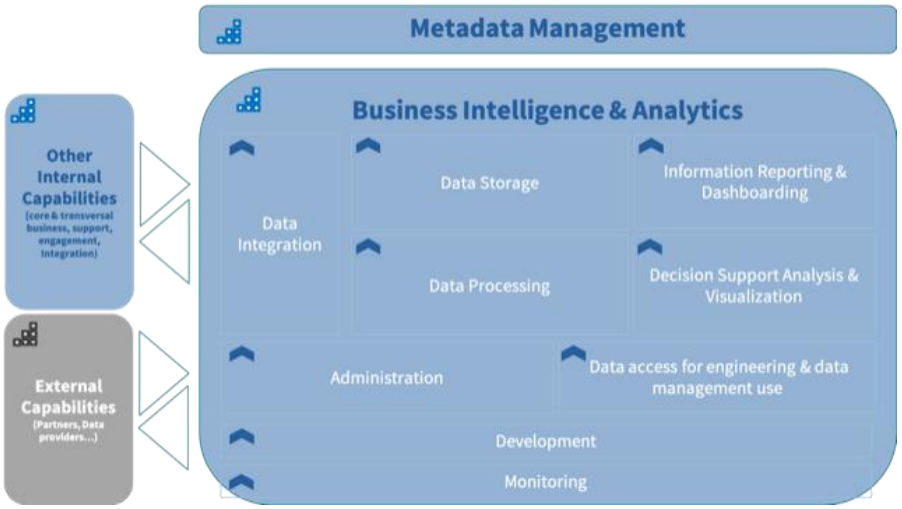


Figure 7: Focus on BI & Analytics Capability Functions

The BI & Analytics scope includes the following high-level application functions:

- **Administration**: administration tasks required to support BI & Analytics capabilities, such as access management, scheduling, data processing, etc.
- **Data access for engineering and data management use**: direct access to the data of BI & Analytics for advanced data stewards, data engineers, data scientists, data administrators, etc.
- **Data integration**: services that allow integration and BI & Analytics in the full landscape, including batch, streaming, interfaces, etc.
- **Data processing**: services that allows data transformation, including basic preparations or more advanced features, calculations, leveraging statistics, or data science
- **Data storage**: services that allow the persistent storage of structured or unstructured data
- **Decision support, analysis, and visualization**: support data discovery and visualization using data stored in BI & Analytics
- **Information reporting and dashboarding**: support reporting and dashboarding activities using data stored in BI & Analytics
- **Development**: development capabilities, including development integrated environment and continuous delivery chain for data scientists, data engineers, etc.
- **Monitoring**: monitoring services for infrastructure, network, and applications to guarantee the service level agreed

6 BI & Analytics Reference Model – Detailed Level

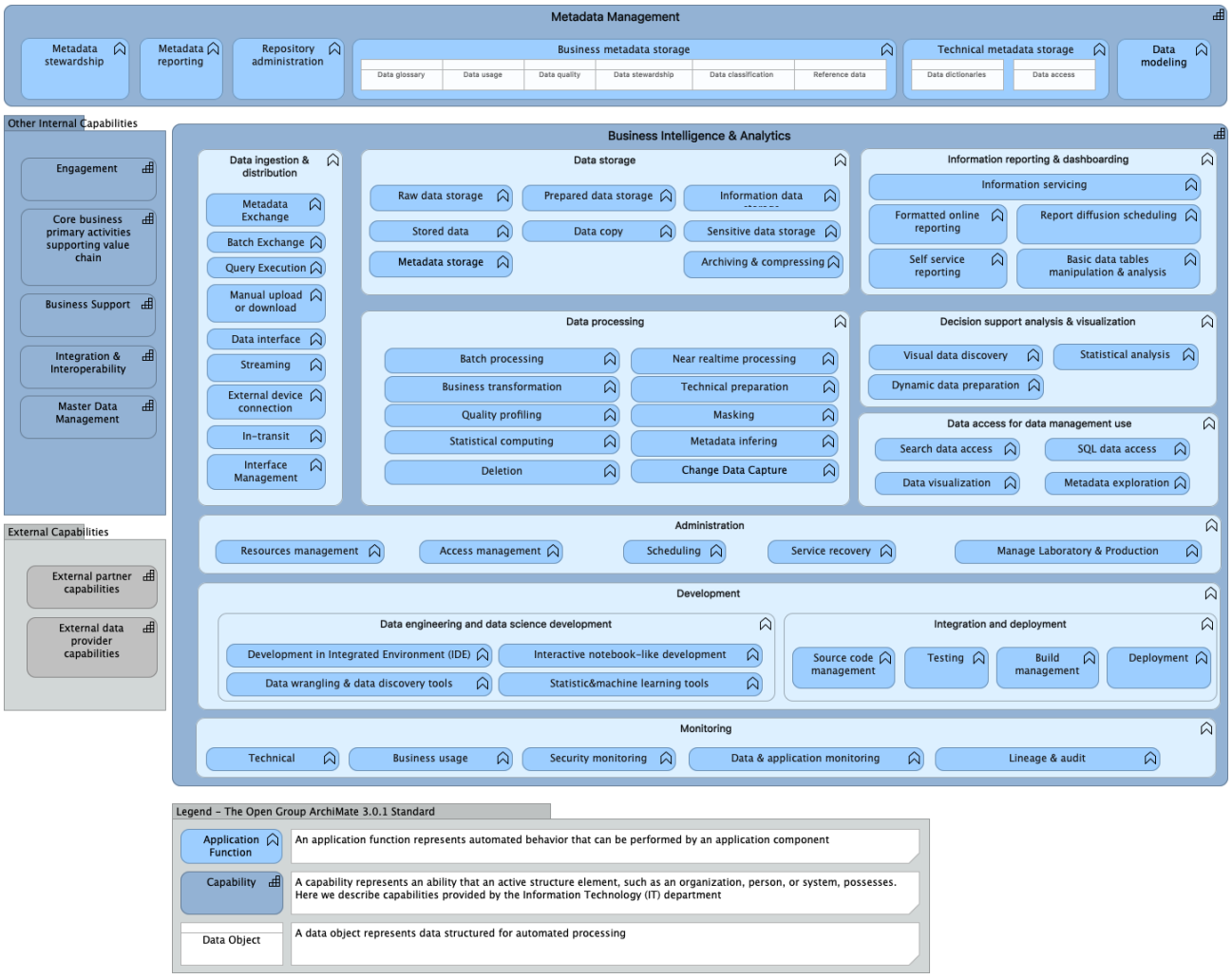


Figure 8: BI & Analytics Reference Model – Detailed Level

Table 2: Functions of the BI & Analytics Reference Model

Title	Description	Number of Requirements (see Chapter 7)
BI & Analytics	Capabilities applicable to the full scope of BI & Analytics.	17
BI & Analytics – Administration	Administration tasks required to support BI & Analytics capabilities such as access management, scheduling data processing, etc. Requirements listed here apply to the full scope.	1

Title	Description	Number of Requirements (see Chapter 7)
Administration – Access Management	Identity and access management, mainly including authentication and authorization services for users and administrators.	14
Administration – Manage Laboratory and Production Environment	Management of different environments: production, test, development.	3
Administration – Resources Management	Resources management of BI & Analytics.	4
Administration – Scheduling	Job scheduler administration.	5
Administration – Service Recovery	Recovery of the service after a disaster.	7
BI & Analytics – Data Access for Data Management	Direct access to the data of BI & Analytics for advanced data stewards, data engineers, data scientists, data administrators, etc. Requirements listed here apply to the full scope.	0
Data Access for Data Management – Data Visualization	Ability to visualize data directly on BI & Analytics storage.	3
Data Access for Data Management – Metadata Exploration	Ability to explore the metadata of BI & Analytics.	7
Data Access for Data Management – Search Data Access	Ability to search for data through a search engine directly on BI & Analytics storage.	4
Data Access for Data Management – SQL Data Access	Ability to perform Structured Query Language (SQL) queries directly on BI & Analytics storage.	1
BI & Analytics – Data Engineering and Data Science Development	Capabilities applicable to the developer's workstation.	13
Data Engineering and Data Science Development – Data Wrangling and Data Discovery Tools	Data preparation tooling for the data engineer or data scientist.	3
Data Engineering and Data Science Development – Integrated Development Environment	Integrated Development Environment (IDE) for the data engineer or data scientist.	11

Title	Description	Number of Requirements (see Chapter 7)
Data Engineering and Data Science Development – Interactive Notebook-like Development	Notebook environment (for data analytics in early exploration phase, algorithm comparison) for data scientists, data engineers, and graphical presentation of data.	4
Data Engineering and Data Science Development – Statistic and Machine Learning Tools	Statistic and machine learning tools for data scientists and data engineers.	1
BI & Analytics – Data Ingestion and Distribution	Services that allow integration and BI & Analytics in the full landscape including batch, streaming, interfaces, etc. Requirements listed here apply to the full scope.	5
Data Ingestion and Distribution – Batch Exchange	Scheduled asynchronous distribution of files.	5
Data Ingestion and Distribution – Data Interface	Web service integration, APIs, etc.	6
Data Ingestion and Distribution – External Device Connection	Integration with an external device connection.	1
Data Ingestion and Distribution – Interface Management	Capabilities applicable to full “Interface Management” scope to provide a central point to monitor and configure interfaces in a self-service mode for entities.	12
Data Ingestion and Distribution – In-transit Encryption	Protection of data transport through encryption of the data in motion (transit) on the network.	4
Data Ingestion and Distribution – Manual Upload or Download	Integration of a file provided by an end user or download of a set of data.	10
Data Ingestion and Distribution – Metadata – Exchange	Metadata ingestion and distribution in parallel from data.	6
Data Ingestion and Distribution – Query Execution	Synchronous Java® Database Connectivity (JDBC) or Open Database Connectivity (ODBC) database connection.	10
Data Ingestion and Distribution – Streaming	Streaming through a messaging system that transports and secures data flow back pressure.	2

Title	Description	Number of Requirements (see Chapter 7)
BI & Analytics – Data Processing	Services that allow data transformation including basic preparations or more advanced features calculations leveraging statistics or data science. Requirements listed here apply to the full scope.	0
Data Processing – Batch Processing	Daily, weekly, monthly, and yearly data transfers performed by batch.	17
Data Processing – Business Transformation	Business processing (filtering, aggregation, joining, and de-duplication).	8
Data Processing – Change Data Capture	Ingestion or distribution of change data capture flows.	1
Data Processing – Deletion	Deletion of data from BI & Analytics (i.e., GDPR).	10
Data Processing – Masking	Anonymizing or pseudonymizing BI & Analytics data.	21
Data Processing – Metadata Inferring	Process of deducing properties of data-such as type, classification, metadata, etc. through the analysis of the data itself.	5
Data Processing – Near-real-time Processing	Low-latency data processing – less than 24 hours/day (includes micro batch and near-real time, with minute or second latency).	12
Data Processing – Quality Profiling	Profiling of data quality through statistical techniques.	4
Data Processing – Statistical Computing	Data science and statistical processing (classification, clustering of data using statistics, data science modeling, including machine learning).	2
Data Processing – Technical Preparation	Business-agnostic technical processing common to all use-cases (i.e., type format, cleansing, integration).	10
BI & Analytics – Data Storage	Services that allow persistent storage of structured or unstructured data. Requirements listed here apply to the full scope.	17
Data Storage – Information Data Storage	Storage of data that has been prepared for a specific business use-case: insights, features calculated by data science, aggregations. Includes the capability to execute low-latency queries for streaming jobs.	3
Data Storage – Prepared Data Storage	Storage of data that has been processed applies standard technical rules that are business-agnostic: type formats, management of null value, generation of unique keys, etc.	2
Data Storage – Raw Data Storage	Storage of source data without any change.	8

Title	Description	Number of Requirements (see Chapter 7)
Data Storage – Archiving and Compression	Archiving and compression functions to manage the data lifecycle and optimize storage resources.	2
Data Storage – Data Copy	Function that allows copying of datasets between different environments.	0
Data Storage – Metadata Storage	Storage of the metadata required by BI & Analytics (schema).	0
Data Storage – Sensitive Data Storage	Storage of sensitive data; e.g., personal, financial, strategy, business, etc.	5
Data Storage – Stored Data Encryption	Protection of data within a node through encryption of the data at rest.	7
BI & Analytics – Decision Support Analysis and Visualization	Support data discovery and visualization using data stored in BI & Analytics (e.g., TIBCO Spotfire®, Tableau®).	2
Decision Support Analysis and Visualization – Dynamic Data Preparation	Data dynamic preparation, with no persistence, to allow statistical processing.	0
Decision Support Analysis and Visualization – Statistical Analysis	Apply dynamic statistical "on-the-shelf" models to information available online.	1
Decision Support Analysis and Visualization – Visual Data Discovery	Dynamic drill down to navigate the data and find the right chart, graphic, etc.	2
BI & Analytics – Information Reporting and Dashboarding	Support reporting and dashboarding activities using data stored in BI & Analytics (e.g., Microsoft® BI or SAP® BI).	2
Information Reporting and Dashboarding – Formatted Online Reporting	Ready-to-use reports and charts that can be refreshed interactively to use data with no specific data management skills.	5
Information Reporting and Dashboarding – Information Servicing	Storage of information data in a format adapted to queries (cube, in memory, view, etc.).	2
Information Reporting and Dashboarding – Report Diffusion Scheduling	Ready-to-use static reports, charts generated periodically, usually paginated and in a printable format.	3
Information Reporting and Dashboarding – Self-service Reporting	Build reports and refresh them with data. Ability for a business user – with no SQL skill, and a drag-and-drop UI – to create a report from available information and share it with other users.	2

Title	Description	Number of Requirements (see Chapter 7)
BI & Analytics – Integration and Deployment	Continuous delivery chain, including versioning, deployment, packaging, etc.	5
Integration and Deployment – Build Management	Prepare the package to be deployed: run the test and generate binaries.	6
Integration and Deployment – Deployment	Deploy a package – new software release – on a given environment (development, integration, pre-production, production).	9
Integration and Deployment – Source Code Management	Versioning of development code (trunk) and release (branch) for developers or data scientists.	5
Integration and Deployment – Testing	Execution of tests on the new code release (non-regression).	1
BI & Analytics – Metadata Management	Capabilities applicable to full "Metadata Management" scope, such as storage, stewardship, reporting, and administration of metadata.	3
Metadata Management – Business Metadata Storage	Storage of business metadata: non-technical information about data. Capabilities applicable to full "Business Metadata Storage" scope.	1
Metadata Management – Business Metadata Storage – Data Classification	Storage of data classification: privacy, security, valuable assets, etc.	3
Metadata Management – Business Metadata Storage – Data Glossary	Storage of business metadata, such as business name, definition. Uniqueness should be respected in the data glossary.	7
Metadata Management – Business Metadata Storage – Data Quality	Storage of business data quality rules and measurement.	2
Metadata Management – Business Metadata Storage – Data Stewardship	Storage of information, such as the stakeholders in charge of data management.	2
Metadata Management – Business Metadata Storage – Data Usage	Storage of data usage activities, including processing identification of personal data.	2
Metadata Management – Business Metadata Storage – Reference Data	Storage of reference data required as examples for reporting.	1
Metadata Management – Data Modeling	Capability to design data models.	4

Title	Description	Number of Requirements (see Chapter 7)
Metadata Management – Metadata Reporting	Generation of reports on Metadata Management.	11
Metadata Management – Metadata Stewardship	Addition of business metadata related to data.	12
Metadata Management – Repository Administration	Administration of the Metadata Management (workflow, authorization, etc.).	10
Metadata Management – Technical Metadata Storage	Storage of technical metadata: information about the technical details of the data (schema, links, access, etc.).	0
Metadata Management – Technical Metadata Storage – Data Access	Storage of data access permissions.	1
Metadata Management – Technical Metadata Storage – Data Dictionaries	Storage of data dictionaries (databases, tables, columns).	5
BI & Analytics – Monitoring	Monitoring services for infrastructure, network, and application to guarantee the service level agreed.	8
Monitoring – Business Usage Monitoring	Monitoring of the business services of the platform (supervise data science model drifting, etc.).	2
Monitoring – Data and Application Monitoring	Monitoring of the data (e.g., quality) and application layer (e.g., availability of services).	14
Monitoring – Lineage and Audit	Capture of the data lifecycle, including the data origin and where it moves over time.	10
Monitoring – Security Monitoring	Monitoring of the security layer.	1
Monitoring – Technical Monitoring	Monitoring of the network layer and infrastructure layer.	5
Other Internal Capabilities	Other capabilities available within the organization.	0
Other Internal Platforms – Core Business Supporting Value Chain	Property & Casualty (P&C) (retail/commercial), Life & Savings (L&S) (individuals/group), health (individuals/group), transversal businesses, asset management, and core banking.	0
Other Internal Platforms – Business Support	Human resources, finance, and procurement.	0
Other Internal Platforms – Engagement Platforms	CRM (supporting sales and services, including lead management).	0

Title	Description	Number of Requirements (see Chapter 7)
Other Internal Platforms – Integration and Interoperability	BI & Analytics gathers and distributes data through the integration platforms (ESB, ETL, API).	0
External Capabilities	Capabilities provided by actors external to the organization.	0
External Platforms – External Partners	Data exposed on the Internet through the Representational State Transfer (REST) API.	0
External Platforms – External Data Providers	Data exposed through a website URL on the Internet.	0

7 BI & Analytics Reference Model – Requirements

The architecture requirements listed in this chapter provide some details about the application functions that are described in the Reference Architecture.

It should probably not be read as a whole: you can refer to it to get some details on some functions.

The level of detail is not completely homogeneous, and the content is probably less universal. Nevertheless, this content provides an example of how requirements engineering can support more formal architecture descriptions and the use of the Reference Architecture. Where requirements contain metrics, these should be seen as examples and the specific values may need to be changed to meet the specific needs of an enterprise.

Architecture Requirements

The Enterprise Architecture systemic approach is applied here. The whole scope of work is a system composed of subsystems. Each function constitutes a subsystem. Thereby, the requirement wording "the system" refers to the function of the requirement.

They are provided in Chapter 6 by the functions listed and here, in this chapter, they are provided in alphabetical order.

The architecture requirements represent formal statements of need, as expressed by stakeholders, which must be met by the architecture.

Architecture requirements capture what is architecturally significant or, in other words, they should capture what has an impact on the architecture decisions. They should not be as detailed as specification or test requirements but can be an input to writing them.

By formal statements, we mean:

- Explicit and understandable need
- Not solution-oriented, no tooling name
- Atomic
- Unique
- Verifiable

The architecture requirements are synthetic but detailed enough to allow:

- Clear understanding of the capabilities needed and delivered
- Roadmap prioritization
- Securing of the implementation of the BI & Analytics platform
- Architecture decisions
- Precise understanding of the capabilities provided by IT suppliers

The formalization of requirements uses the following structure: “The system should [verb][object][stakeholder]”.

7.1 Administration

7.1.1 Administration – Access Management

Identity and access management, mainly includes authentication and authorization services for users and administrators.

Table 3: Administration – Access Management

ID	Description	Objective
920	The system should manage data authorization and access at container and tenant level.	Comply with security – identity and access management
921	The system should enable the Creation, Read, Update, and Deletion (CRUD) of user and group folders, and give authorizations depending on roles (including batch or user role) in self-service mode for end users.	Comply with security – identity and access management
922	The system should log every access for 10 years.	Comply with security – infrastructure baseline
923	The system should use Single Sign On (SSO) to access the platform. Rationale: Access issues control.	Comply with security – identity and access management
924	The system should provide a user role with deletion rights, including for raw data.	Comply with security – identity and access management
925	The system should log every single user accessing data through an external application UI. Rationale: Access issues control.	Comply with security – infrastructure baseline
926	The system should log every single user accessing data. Rationale: Access issues control.	Comply with security – infrastructure baseline
927	The system should manage an access log containing personal data, with access control respecting security requirements.	Comply with security – data classification and encryption
928	The system should provide a support service for user access control.	Guarantee quality of service
929	The system should manage data authorization and access at data processing at file level.	Comply with security – identity and access management
930	The system should manage data authorization and access at data processing at table/field level.	Comply with security – identity and access management

ID	Description	Objective
931	The system should implement personal data access control, capturing documentation of access rights justification.	Comply with security – identity and access management
932	The system should provide data access workflow, where the data owner is part of the validation process.	Comply with security – identity and access management
933	The system should provide a user role with full deletion rights on all the containers of a given entity.	Comply with security – identity and access management

7.1.2 Administration – Manage Laboratory and Production Environment

Management of different environments: production, test, development.

Table 4: Administration – Manage Laboratory and Production Environment

ID	Description	Objective
960	The system should automatically purge old data of the sandbox storage, according to the requirements of the data scientist or developer. Rationale: Entity efficiency.	Improve operational efficiency
961	The system should replicate data from production storage to sandbox storage (one shot, scheduled, full or selective) in a self-service mode. Rationale: Entity efficiency.	Reduce use-case time-to-market
962	The system should manage the lifecycle of datasets in a non-production environment and delete datasets at the end of the retention period.	Reduce the TCO

7.1.3 Administration – Resources Management

Resources management of BI & Analytics.

Table 5: Administration – Resources Management

ID	Description	Objective
980	The system should arbitrate resources (Central Processing Unit (CPU), memory, disk, network) distribution in the cluster.	Guarantee quality of service
981	The system should allow reporting on resources (CPU, memory, disk, network) usage in the cluster.	Guarantee quality of service
982	The system should provide fair allocation of available resources.	Guarantee quality of service
983	The system should provide fair allocation of available resources on the non-production environment.	Guarantee quality of service

7.1.4 Administration – Scheduling

Job scheduler administration.

Table 6: Administration – Scheduling

ID	Description	Objective
1000	The system should allow job scheduling to process data.	Improve operational efficiency
1001	The system should allow event trigger data processing.	Improve operational efficiency
1002	The system should allow orchestration of scheduled jobs across the system from end-to-end. Rationale: Overall governance of the System Information (SI).	Guarantee quality of service
1003	The system should log job scheduling. Rationale: In case of error handling.	Improve operational efficiency
1004	The system should allow scheduling of table reading.	Improve operational efficiency

7.1.5 Administration – Service Recovery

Recovery of the service after a disaster.

Table 7:Administration – Service Recovery

ID	Description	Objective
1020	The system should provide a Recovery Point Objective (RPO) value of 24H00.	Comply with security – infrastructure baseline
1021	The system should permit choosing data to backup.	Comply with security – infrastructure baseline
1022	The system should provide a second location for backup data.	Comply with security – infrastructure baseline.
1023	The system should keep data backups for up to 30 days.	Comply with security – infrastructure baseline
1024	The system should keep metadata (Access Control Lists (ACL), owner, etc.) while processes are active.	Comply with security – infrastructure baseline
1025	The system should provide encryption for data backups.	Comply with security – data classification and encryption
1026	The system should allow to backup data from BI & Analytics.	Comply with security – infrastructure baseline

7.2 BI & Analytics

Capabilities applicable to the full scope of BI & Analytics.

Table 8: BI & Analytics

ID	Description	Objective
1	The system should be able to rely on hybrid technical services – private on-premise, public cloud, and Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS). Rationale: Make choices compatible with cloud hosting.	Reduce use-case time-to-market
2	The system should provide a catalog of standard available services and delay to execute. Rationale: Easy to operate for entities.	Improve operational efficiency
3	The system should provide non-standard services on-demand in negotiated delay. Rationale: Easy to operate for entities.	Guarantee quality of service
4	The system should provide a service contact (advisor) to support onboard at data project set-up. Rationale: Change management.	Reduce use-case time-to-market
5	The system should be driven by business capability-based planning.	Improve operational efficiency
6	The system should provide the availability as agreed in the Service-Level Agreement (SLA). Rationale: Enabling operational and analytical use-cases.	Guarantee quality of service
7	The system provides a transparent cost model. Rationale: Cost adapted to use-cases requirement; Return on Investment (ROI) optimization.	Reduce the TCO
8	The system should comply with the organization's security requirements.	Comply with security – infrastructure baseline
9	The system should be multi-tenanted to support different organizations.	Reduce the TCO
10	The system should be able to perform upgrades without downtime. Rationale: Easy evolution, loose-coupling, and interoperability.	Guarantee quality of service
11	The system should provide user documentation. Rationale: Usability.	Democratize use of data

ID	Description	Objective
12	The system should support multi-language (English, French, German, Spanish, Italian, Greek, Japanese, Arabic. etc.). Rationale: Usability.	Democratize use of data
13	The system should be able to recover from a disaster in an approved SLA. Rationale: Business continuity.	Guarantee quality of service
14	The system should provide guidance, support, and recommendation on the implementation of data projects.	Reduce use-case time-to-market
15	The system should publish the evolution roadmap of the platform.	Improve operational efficiency
16	The system should provide a forecast process to evaluate the cost of the platform for the next year, including a model to evaluate disk storage requirements, processing, and end costs.	Improve operational efficiency
17	The system should anticipate and meet effectively the increasing needs of system usage in terms of computation.	Improve operational efficiency

7.2.1 BI & Analytics – Information Reporting and Dashboarding

Support reporting and dashboarding activities using data stored in BI & Analytics (e.g., Microsoft BI, SAP BI).

Table 9: BI & Analytics – Information Reporting and Dashboarding

ID	Description	Objective
1500	The system should propose different architecture pattern to serve analytics use-cases (from basic to advanced users).	Democratize use of data
1502	The system should allow resources management, access management, scheduling, and service recovery.	Improve operational efficiency

7.2.2 BI & Analytics – Administration

Administration tasks required to support BI & Analytics capabilities such as access management, scheduling data processing, etc. Requirements listed here apply to the full scope.

Table 10: BI & Analytics – Administration

ID	Description	Objective
900	The system should communicate on version upgrades three (3) months beforehand.	Guarantee quality of service

7.2.3 BI & Analytics – Data Engineering and Data Science Development

Capabilities applicable to the developer's workstation.

Table 11: BI & Analytics – Data Engineering and Data Science Development

ID	Description	Objective
1100	The system should propose data science and data engineer training. Rationale: Change management.	Support digital and data innovation
1101	The system should enable self-service installation of non-standard tools by data engineers and data scientists without impacting on the work of other data engineers. Rationale: Allow trials and tests and trials of new tools. Benefit from market innovation.	Support digital and data innovation
1102	The system should provide Internet access to developers and data scientists.	Support digital and data innovation
1103	The system should provide an adequate service level (concurrent use, latency, etc.).	Guarantee quality of service
1104	The system should allow processing of big datasets (several tera, above peta, etc.).	Generate business value leveraging new data capabilities (e.g., data science, etc.)
1105	The system should provide access to BI & Analytics metadata (raw, information, use-case) by data scientists and data engineers.	Support digital and data innovation
1106	The system should provide data anonymization self-service to data scientists in their environment.	Comply with security – data leakage
1107	The system should provide guidelines to help data scientists and data engineers use anonymization libraries in their working environments.	Comply with security – data leakage
1108	The system should provide streaming messaging in open or private channels across the group. Rationale: Synchronous collaboration.	Improve operational efficiency
1109	The system should provide task management with workflow capability for project teams. Rationale: Community efficiency.	Improve operational efficiency
1110	The system should provide content-centric collaboration between work groups. Rationale: Community efficiency.	Improve operational efficiency
1111	The system should guarantee security clearance for any collaboration capability. Rationale: Developers are meant to collaborate.	Improve operational efficiency

ID	Description	Objective
1112	The system should provide one-to-one streaming messaging across the group. Rationale: Synchronous collaboration.	Improve operational efficiency

7.2.4 BI & Analytics – Data Ingestion and Distribution

Services that allow integration and BI & Analytics in the full landscape including batch, streaming, interfaces, etc. Requirements listed here apply to the full scope.

Table 12: BI & Analytics – Data Ingestion and Distribution

ID	Description	Objective
100	The system should decouple data extraction and data ingestion: operations of integrated systems should not impact each other. Rationale: Protect operational databases of intrusion and workload.	Guarantee quality of service
101	The system should be able to ingest or distribute file structured data; for example, Extensible Markup Language (XML), JavaScript Object Notation (JSON), Comma-Separated Values (CSV), etc. Rationale: Variety of data sources.	Provide interoperability between systems
102	The system should enable the establishment of a permanent connection with a data source where data is captured. Rationale: Ingestion industrialization.	Provide interoperability between systems
103	The system should enable the establishment of a permanent connection with a data consumer. Rationale: Use-case industrialization.	Provide interoperability between systems
104	The system should allow data transfer with low latency in the outbound distribution.	Provide interoperability between systems

7.2.5 BI & Analytics – Data Storage

Services that allows persistent storage of structured or unstructured data. Requirements listed here apply to the full scope.

Table 13: BI & Analytics – Data Storage

ID	Description	Objective
700	The system should provide storage compatible with data residency restrictions.	Reduce the TCO

ID	Description	Objective
701	The system should propose different storage mediums with different pricing models. Rationale: Cost efficiency.	Reduce the TCO
702	The system should scale to enable exceptional growth of data volume, above 5TB within an agreed delay (2 months). Rationale: Easy evolution, loose-coupling, and interoperability.	Guarantee quality of service
703	The system should allow sorting of data at storage time, allowing grouping of data retrieved together on the same node in order to provide good performance at retrieval time. Rationale: Performance.	Guarantee quality of service
704	The system should allow file storage.	Store data as an asset, including metadata
705	The system should allow table creation.	Store data as an asset, including metadata
706	The system should allow table alteration.	Store data as an asset, including metadata
707	The system should separate data storage according to the stage of the data lifecycle (raw, cleansed, prepared, use-case).	Store data as an asset, including metadata
708	The system should log any event happening on the storage layer (file operations, create, alter, drop, insert, update, delete).	Comply with security – infrastructure baseline
709	The system should store any unstructured data. Rationale: Final use-cases might need unstructured data.	Store data as an asset, including metadata
710	The system should enable linear scalability in terms of data query execution performance up to 30 simultaneous users. Rationale: Ensuring performance.	Guarantee quality of service
711	The system should be able to store open data.	Store data as an asset, including metadata
712	The system should propose a standard file arborescence (raw, prepared, use-case) while leaving the ability to customize it for authorized end users. Rationale: Easy-to-find personal data.	Data quality is under control
713	The system should segregate data according to its classification (public, private, confidential, secret). Rationale: Ensure security of data with different levels of security.	Comply with security – data classification and encryption
714	The system should allow read, create, and delete operations on stored data.	Improve operational efficiency

ID	Description	Objective
715	The system should allow the management of zones on stored data managed by different rules in terms of anonymization, service level, and security.	Comply with security – identity and access management
716	The system should be able to distinguish data which is coming from different environments (for data copies). Rationale: Ensure data integrity in case of data copies.	Data quality is under control

7.2.6 BI & Analytics – Decision Support Analysis and Visualization

Support data discovery and visualization using data stored in BI & Analytics (e.g., TIBCO Spotfire, Tableau).

Table 14: BI & Analytics – Decision Support Analysis and Visualization

ID	Description	Objective
1501	The system should allow the restitution of data on geographical maps.	Democratize use of data
1600	The system should provide administration services required for visual discovery: resources management, access management, scheduling, service recovery.	Improve operational efficiency

7.2.7 BI & Analytics – Integration and Deployment

Continuous delivery chain including versioning, deployment, packaging, etc.

Table 15: BI & Analytics – Integration and Deployment

ID	Description	Objective
1200	The system should enable Continuous Integration/Continuous Delivery (CI/CD) chains on BI & Analytics. Rationale: Limit manual login process with system account to the platform – secure data and operations on the platform. Automate the deployment of new applications/releases.	Reduce use-case time-to-market
1201	The system should allow the automated synchronous delivery of code on different platforms; i.e., hybrid cloud (on-premise, private cloud, public cloud, etc.). Rationale: End-to-end deployment of business services.	Reduce use-case time-to-market
1202	The system should provide a technical delivery chain separated from the business project chain to secure their planning. Rationale: Avoid impact of technical upgrades on project.	Improve operational efficiency

ID	Description	Objective
1203	The system should organize each development with a proper lifecycle model.	Improve operational efficiency
1204	The system should package releases on container images (deployment simplification).	Improve operational efficiency

7.2.8 BI & Analytics – Monitoring

Monitoring services for infrastructure, network, and application to guarantee the agreed service level.

Table 16: BI & Analytics – Monitoring

ID	Description	Objective
1300	The system should provide support on the overall BI & Analytics ecosystem (from ingestion to distribution) for agreed hours (worldwide from 7.30am until 6.00pm).	Guarantee quality of service
1301	The system should allow mail gateway access to configure alerts to send exceptions, notifications, or reminders with predefined senders, and broadly defined recipient lists. Rationale: Alert on incident.	Improve operational efficiency
1302	The system should allow exposition of platform monitoring information to allow cross-system monitoring. Rationale: Monitor Data Architecture.	Improve operational efficiency
1303	The system should provide monitoring of the service level 24 hours per day. Rationale: Monitor Data Architecture.	Improve operational efficiency
1304	The system should monitor the services to detect or anticipate service interruption or degradation of data capabilities according to defined service levels. Rationale: Enable action on interruption and prevention.	Improve operational efficiency
1305	The system should provide a Graphical User Interface (GUI) to monitor the data platform (SLAs, usage, billing, etc.). Rationale: Monitor Data Architecture.	Improve operational efficiency
1306	The system should guarantee the service level agreed (SLA) on input data – file or stream – arrival. Rationale: Preparation of data for analysis.	Guarantee quality of service
1307	The system should provide an SLA tracking service. Rationale: Ensure minimum service level for business performance.	Improve operational efficiency

7.3 Data Access for Data Management

7.3.1 Data Access for Data Management – Data Visualization

Ability to visualize data directly on BI & Analytics storage.

Table 17: Data Access for Data Management – Data Visualization

ID	Description	Objective
320	The system should provide a set of capabilities to search, analyze, and visualize data directly from BI & Analytics storage according to the access rights. Rationale: Data access.	Store data as an asset, including metadata
321	The system should allow browsing files directly on the BI & Analytics storage layer when appropriate access rights have been given. Rationale: Data access.	Store data as an asset, including metadata
322	The system should provide full-text search on metadata (*aka* Google®). Rationale: Increase knowledge and mastership of data.	Store data as an asset, including metadata

7.3.2 Data Access for Data Management – Metadata Exploration

Ability to explore the metadata of BI & Analytics.

Table 18: Data Access for Data Management – Metadata Exploration

ID	Description	Objective
340	The system should store "lineage"; i.e., the data's origins, what happens to it, and where it moves over time as operational metadata. Rationale: Knowledge of data statistics for data enhancement.	Comply with security – infrastructure baseline
341	The system should store "data access log" operational metadata for six (6) months, as required by European authorities. Rationale: Knowledge of data statistics for data enhancement.	Comply with security – infrastructure baseline
342	The system should store "data quality log" operational metadata. Rationale: Knowledge of data statistics for data enhancement.	Comply with security – infrastructure baseline
343	The system should allow the storage of a business data glossary. Rationale: Locate the data.	Store data as an asset, including metadata
344	The system should store "data privacy classification".	Comply with security – data classification and encryption

ID	Description	Objective
345	The system should be able to identify through technical metadata the link between tables.	Store data as an asset, including metadata
346	The system should expose metadata through a GUI.	Data quality is under control

7.3.3 Data Access for Data Management – Search Data Access

Ability to search for data through a search engine directly on BI & Analytics storage.

Table 19: Data Access for Data Management – Search Data Access

ID	Description	Objective
360	The system should allow searches on unstructured data.	Store data as an asset, including metadata
361	The system should allow near-real-time indexation.	Store data as an asset, including metadata
362	The system should allow identification of personal data within unstructured files or text fields.	Comply with security – data classification and encryption
363	The system should allow identification of records relating to a given person (personal data) in the whole data storage.	Comply with security – data classification and encryption

7.3.4 Data Access for Data Management – SQL Data Access

Ability to perform SQL queries directly on BI & Analytics storage.

Table 20: Data Access for Data Management – SQL Data Access

ID	Description	Objective
380	The system should expose ingested raw data in a structured query-executionable format (create table, add row). Rationale: Data must be query-executionable to be processed.	Store data as an asset, including metadata

7.4 Data Engineering and Data Science Development

7.4.1 Data Engineering and Data Science Development – Data Wrangling and Data Discovery Tools

Data preparation tooling for the data engineer or data scientist.

Table 21: Data Engineering and Data Science Development – Data Wrangling and Data Discovery Tools

ID	Description	Objective
1120	The system should provide data selection functionalities such as random sampling, systematic sampling, and stratified sampling.	Support digital and data innovation
1121	The system should provide a UI to load data, infer schema, profile data, and test data transformation in order to design the prepared data in BI & Analytics according to access rights. Rationale: Accelerate data discovery.	Improve operational efficiency
1122	The system should provide library management for data wrangling and data discovery tools.	Improve operational efficiency

7.4.2 Data Engineering and Data Science Development – Integrated Development Environment

IDE for the data engineer or data scientist.

Table 22: Data Engineering and Data Science Development – Integrated Development Environment

ID	Description	Objective
1140	The system should prevent the transfer of data on a collaborator's laptop.	Comply with security – data leakage
1141	The system should give access from the IDE to the data available in BI & Analytics (development, exploration, etc.) with appropriate access control to data scientists and data engineers.	Reduce use-case time-to-market
1142	The system should provide out-of-the-box IDE for code development (e.g., Python, R, Scala) for data engineers and data scientists with appropriate autonomy to install a library and specific tools.	Reduce use-case time-to-market
1143	The system should generate code documentation from source code.	Improve operational efficiency
1144	The system should provide Scala, Python library, and project dependency isolation for multiple library versions.	Reduce use-case time-to-market

ID	Description	Objective
1145	The system should provide a library and up-to-date development tools every six (6) months. Rationale: Use the best tools available on the market.	Support digital and data innovation
1146	The system should allow presentation of data in a dashboard to allow prototypes and demos to the business for data scientists and data engineers. Rationale: Embark on business in data science.	Support digital and data innovation
1147	The system should be able to process distributable computing during development phases. Rationale: Needed for non-distributable computing when the algorithms are independent from each other; for example, a neural network).	Guarantee quality of service
1148	The system should provide UI for collaboration on model development between data scientists and actuaries or statisticians with heterogeneous development skills. Rationale: Collaboration between data scientists and actuaries that are not trained to launch Apache Spark® command lines on distributed environments.	Generate business value leveraging new data capabilities (e.g., data science, etc.)
1149	The system should allow data scientists to independently generate data products; i.e., applications for end users from data science code, such as R. Rationale: Reduce the gap between the data analysis process and building a relevant product for the business.	Generate business value leveraging new data capabilities (e.g., data science, etc.)
1150	The system should allow the use of anonymization libraries from development and data scientist environments.	Comply with security – data leakage

7.4.3 Data Engineering and Data Science Development – Interactive Notebook-Like Development

Notebook environment (for data analytics in early exploration phase, algorithm comparison) for data scientists, data engineers, and graphical presentation of data.

Table 23: Data Engineering and Data Science Development – Interactive Notebook-Like Development

ID	Description	Objective
1160	The system should provide interactive access – execute code and look at result – to Apache Spark using Scala code. Rationale: Increase developer efficiency.	Reduce use-case time-to-market
1161	The system should provide interactive access to Apache Spark using Python or R code. Rationale: Increase developer efficiency.	Reduce use-case time-to-market

ID	Description	Objective
1162	The system should allow the visualization of data in charts and diagrams. Rationale: Increase developer efficiency.	Democratize use of data
1163	The system should allow query execution of different Not Only SQL (NOSQL) storage (Hive®/Mapreduce, MongoDB®, Hbase, Elastic search). Rationale: Increase developer efficiency.	Reduce use-case time-to-market

7.4.4 Data Engineering and Data Science Development – Statistic and Machine Learning Tools

Statistic and machine learning tools for data scientists and data engineers.

Table 24: Data Engineering and Data Science Development – Statistic and Machine Learning Tools

ID	Description	Objective
1180	The system should provide library management for data wrangling and data discovery tools.	Improve operational efficiency

7.5 Data Ingestion and Distribution

7.5.1 Data Ingestion and Distribution – Batch Exchange

Scheduled asynchronous distribution of files.

Table 25: Data Ingestion and Distribution – Batch Exchange

ID	Description	Objective
120	The system should be able to ingest executable files on the platform to execute business rules. Rationale: Autonomy of projects.	Provide interoperability between systems
121	The system should enable file transfers from 10GB up to 1TB per file in less than 18 hours. Rationale: Enable input and output data stream with other systems (core IT, data warehouses).	Guarantee quality of service
122	The system should enable files transfer up to 10GB per file in less than six (6) hours. Rationale: Enable input and output data stream with other systems (core IT, data warehouses).	Guarantee quality of service
123	The system should automate – i.e., with no manual steps – data ingestion through a permanent connection.	Provide interoperability between systems

ID	Description	Objective
124	The system should be fault-tolerant.	Guarantee quality of service

7.5.2 Data Ingestion and Distribution – Data Interface

Web service integration, APIs, etc.

Table 26: Data Ingestion and Distribution – Data Interface

ID	Description	Objective
140	The system should be able to connect to a web service provided by another platform to retrieve data. Rationale: Operational processing.	Provide interoperability between systems
141	The system should be able to connect to an API provided by another platform to retrieve data. Rationale: Operational processing.	Provide interoperability between systems
142	The system should be able to ingest website scrapping data.	Provide interoperability between systems
143	The system should enable fetching of external data files (weather, traffic, web scrapping, twitter). Rationale: Need for use-cases, exploration, and industrial usage.	Provide interoperability between systems
144	The system should be able to download page content from an external website network. Rationale: Use data for enrichment.	Provide interoperability between systems
511	The system exposes an API. Rationale: Target integration with Metadata Management tool.	Provide interoperability between systems

7.5.3 Data Ingestion and Distribution – External Device Connection

Integration with an external device connection.

Table 27: Data Ingestion and Distribution – External Device Connection

ID	Description	Objective
160	The system should allow integration from an external device connection in a secured manner.	Provide interoperability between systems

7.5.4 Data Ingestion and Distribution – Interface Management

Capabilities applicable to full "Interface Management" scope to provide a central point to monitor and configure interfaces in a self-service mode for entities.

Table 28: Data Ingestion and Distribution – Interface Management

ID	Description	Objective
280	The system should provide error details of failed file integration. Rationale: Debug facility.	Guarantee quality of service
281	The system should provide administrators of the data platform with configuration files/UIs to grant privilege for manual upload in the tenant/container defining upload authorization to users, expected file type, expected schema, and naming convention. Rationale: Control rights to upload file.	Comply with security – identity and access management
282	The system should provide super-administrators and entity administrators with a report in the UI, listing the file's manual import authorizations granted.	Comply with security – identity and access management
283	The system should allow the configuration of a new file transfer for ingestion or distribution in a self-service mode. Rationale: Autonomy of projects.	Provide interoperability between systems
284	The system should be able to schedule data ingestion in a self-service mode. Rationale: Autonomy of projects.	Provide interoperability between systems
285	The system should allow subscription to a data service in a self-service mode. Rationale: Autonomy of projects.	Provide interoperability between systems
286	The system should be able to transport and leverage technical metadata (schema) with the data.	Provide interoperability between systems
287	The system should be able to configure new data transfer from or to another integration self-service instance. Rationale: Interoperability with other data hubs.	Provide interoperability between systems
288	The system should allow management of the workflow in multi-tenancy: each entity can administrate its workflows, and only its workflows. Rationale: Allow management of several entities' space.	Reduce the TCO
289	The system should enable configuration of event storage time up to seven (7) days. Rationale: Resilience to guarantee data transmission even in the case of a shortage of data producer or consumer. Provide error recovery capability.	Guarantee quality of service

ID	Description	Objective
290	The system should guarantee the delivery of events only once. Rationale: Guarantee transactions.	Guarantee quality of service
291	The system should allow the authentication of data producers and consumers. Rationale: Enable multi-tenancy.	Comply with security – identity and access management

7.5.5 Data Ingestion and Distribution – In-Transit Encryption

Protection of the data transport through encryption of the data in motion (transit) on the network.

Table 29: Data Ingestion and Distribution – In-Transit Encryption

ID	Description	Objective
180	The system should securely manage encryption keys and access policies. Rationale: Secure data at the transport layer.	Comply with security – data classification and encryption
181	The system should encrypt data in motion with a security-approved encryption algorithm. Rationale: Secure data at the transport layer.	Comply with security – data classification and encryption
182	The system should allow data encryption at rest if, during transportation, data is stored on a temporary server. Rationale: Secure data at the transport layer.	Comply with security – data classification and encryption
183	The system should allow data decryption only for authorized users and processes. Rationale: Secure data at the transport layer.	Comply with security – data classification and encryption

7.5.6 Data Ingestion and Distribution – Manual Upload or Download

Integration of a file provided by an end user or download of a set of data.

Table 30: Data Ingestion and Distribution – Manual Upload or Download

ID	Description	Objective
200	The system should provide a UI allowing upload or download of file and related metadata (schema) by authorized business users (data analysts, etc.) in an authorized repository. Rationale: Upload of reference data not managed in an information system or apply corrections to the facts.	Provide interoperability between systems

ID	Description	Objective
201	The system should perform control on the file format with a schema on the ten (10) first lines (column and types for CSV files). Rationale: Controlling the full file will be too costly in terms of time.	Data quality is under control
202	The system should allow the configuration of the uploaded file name, systematic rename, or not; and adding of a time stamp (hours, minutes, seconds), or not.	Data quality is under control
203	The system should not change the encoding of the file format that should remain one of the files selected by the business user. Rationale: Under the responsibility of the business user.	Data quality is under control
204	The system should control synchronously the uploaded file extension.	Data quality is under control
205	The system should be compatible with identity access management policies SSO authentication). Rationale: The tool is designed for business users; simplicity is key.	Comply with security – identity and access management
206	The system should keep track of all attempted uploads, successful or not, including original file name, original file time stamp, target container, and user performing the upload for a minimal duration of six (6) months and make it available to users through the UI. Rationale: Auditability of the service.	Comply with security – identity and access management
207	The system should make the file available on the target repository in less than 15 minutes from the time the business user submitted it.	Provide interoperability between systems
208	The system should allow the launch of job processing once the file is uploaded and when requested by entity administrators configuring the upload.	Provide interoperability between systems
209	The system should identify conflict when the same file is uploaded twice in the same place and allow configuration of system behavior (overwrite or refuse upload). Rationale: Avoid file operation workload for the competency center.	Data quality is under control

7.5.7 Data Ingestion and Distribution – Metadata Exchange

Metadata ingestion and distribution in parallel from data.

Table 31: Data Ingestion and Distribution – Metadata Exchange

ID	Description	Objective
220	The system should ingest and distribute technical metadata.	Provide interoperability between systems
221	The system should ingest and distribute business metadata (e.g., personal data identification).	Provide interoperability between systems
222	The system should integrate with the metadata platform in master or slave mode. Rationale: Metadata reference is managed in a metadata repository.	Provide interoperability between systems
223	The system should propose a pivot format for metadata integration.	Provide interoperability between systems
224	The system should ingest and distribute operational metadata (lineage, audit event, etc.). Rationale: Interoperability/ability to integrate in an existing architecture.	Provide interoperability between systems
225	The system should expose lineage operational metadata. Rationale: Interoperability to enable access to metadata stored in the repository for systems/persons.	Comply with security – infrastructure baseline

7.5.8 Data Ingestion and Distribution – Query Execution

Synchronous JDBC or ODBC database connection.

Table 32: Data Ingestion and Distribution – Query Execution

ID	Description	Objective
240	The system should allow data ingestion through direct synchronous connection to the database of other internal platforms.	Provide interoperability between systems
241	The system should support synchronous ANSI® – SQL query execution through standard connection on BI & Analytics tables (data, information, or storage) by external consumer application. Rationale: Enable access to the data stored in the data lake.	Store data as an asset, including metadata
242	The system should allow insert in a table within BI & Analytics. Rationale: Enable access to the data stored in the data lake.	Store data as an asset, including metadata

ID	Description	Objective
243	The system should allow updates in a table within BI & Analytics. Rationale: Enable access to the data stored in the data lake.	Store data as an asset, including metadata
244	The system should allow deletions in a table within BI & Analytics. Rationale: Enable access to the data stored in the data lake.	Store data as an asset, including metadata
245	The system should allow query execution leveraging the distributed processing capability. Rationale: Enable access to the data stored in the data lake.	Guarantee quality of service
246	The system should execute a query execution on a small dataset (<1GB) in less than a few seconds from an external consumer application. Rationale: Ensure performance for user efficiency.	Guarantee quality of service
247	The system should execute a query execution on a medium dataset (1-100GB) in less than 100 seconds from an external consumer application (i.e., including BI & Analytics execution time and network transfer). Rationale: Ensure performance for user efficiency.	Guarantee quality of service
248	The system should execute a query execution on a large dataset (100GB-1TB) in less than 500 seconds from an external consumer application (i.e., including BI & Analytics execution time and network transfer). Rationale: Ensure performance for user efficiency.	Guarantee quality of service
249	The system should log each nominative access through queries. Rationale: Track data access.	Comply with security – infrastructure baseline

7.5.9 Data Ingestion and Distribution – Streaming

Streaming through a messaging system that transports and secures data flow back pressure.

Table 33: Data Ingestion and Distribution – Streaming

ID	Description	Objective
260	The system should enable a very short timeframe between producing (from a source system) and consuming (availability in the data lake) an event. Rationale: Support business use-cases requiring near-real time.	Guarantee quality of service
261	The system should enable streaming ingestion of high volumes daily (above 100GB/days).	Store data as an asset, including metadata

7.6 Data Processing

7.6.1 Data Processing – Batch Processing

Daily, weekly, monthly, yearly data transfers performed by batch.

Table 34: Data Processing – Batch Processing

ID	Description	Objective
420	The system should guarantee computing power predictability; i.e., that business-critical workflows are executed in framed elapsed time.	Guarantee quality of service
421	The system should offer different service levels depending on the desired predictability of data processing time execution. Rationale: Cost adapted to use-cases requirement – ROI optimization.	Reduce the TCO
422	The system should provide basic data processing for structured data (query execution, joining, filtering, sorting, and aggregating). Rationale: Enable data processing.	Store data as an asset, including metadata
423	The system should provide basic data processing for unstructured data (search/select by regex, parsing, stemming, word distribution, and filtering). Rationale: Enable data processing.	Store data as an asset, including metadata
424	The system should scale to enable growth of processing within an agreed delay. Rationale: Anticipate the future load/cost efficiency.	Guarantee quality of service
425	The system should provide an up-to-date library at least every six (6) months.	Improve operational efficiency
426	The system should allow distributed processing. Rationale: Deal with big datasets.	Guarantee quality of service
427	The system should allow non-distributed processing. Rationale: Cost adapted to use-cases requirement – ROI optimization.	Reduce the TCO
428	The system should provide basic data processing based on a public library of statistically oriented languages (Python, R, Spark-ML). Rationale: Use-cases need sometime specific packages for some data processing algorithms.	Reduce use-case time-to-market
429	The system should guarantee that the elapsed time of standard workflow does not exceed a determined threshold.	Guarantee quality of service

ID	Description	Objective
430	The system should guarantee completeness and integrity of processed data after the processing execution (job).	Guarantee quality of service
431	The system should be able to retry an automatically failed processing execution (job).	Guarantee quality of service
432	The system should scale to manage occasional peaks of processing resource use while ensuring service performance. Rationale: Anticipate the future load/cost efficiency.	Guarantee quality of service
433	The system should ensure a maximum loss of only one transaction of data processing during a fallout or error.	Improve operational efficiency
434	The system should ensure data subjects consent before any data processing.	Comply with regulations (e.g., privacy, Solvency II, etc.)
435	The system should only process personal data when an explicit purpose or consent is mentioned.	Comply with regulations (e.g., privacy, Solvency II, etc.)
436	The system should ensure the integrity of deleted, anonymized, or pseudonymized tables through technical metadata.	Store data as an asset, including metadata

7.6.2 Data Processing – Business Transformation

Business processing (filtering, aggregation, joining, and de-duplication).

Table 35: Data Processing – Business Transformation

ID	Description	Objective
440	The system should allow picture file merging.	Improve operational efficiency
441	The system should allow the processing of voice data: transcript → speech2text → semantic → emotional analysis.	Improve operational efficiency
442	The system should allow emotional analysis on text.	Improve operational efficiency
443	The system should enable deep learning modeling. Rationale: Use-case: estimate the level of damage based on a picture of a crash.	Improve operational efficiency
444	The system should allow the extraction of written information from a picture; Optical Character Recognition (OPR). Rationale: Exploitation of text within a picture.	Improve operational efficiency
445	The system should allow file conversion.	Improve operational efficiency
446	The system should allow image feature extraction functionality on binary files.	Improve operational efficiency

ID	Description	Objective
447	The system should provide geocoding services.	Improve operational efficiency

7.6.3 Data Processing – Change Data Capture

Ingestion or distribution of change data capture flows.

Table 36: Data Processing – Change Data Capture

ID	Description	Objective
460	The system should expose the accumulated deltas and transformation that occurred to data in a change data capture mode. Rationale: Change data capture for consumers.	Improve operational efficiency

7.6.4 Data Processing – Deletion

Deletion of data from BI & Analytics (i.e., GDPR).

Table 37: Data Processing – Deletion

ID	Description	Objective
480	The system should allow file deletion within the “raw storage”.	Comply with security – data classification and encryption
481	The system should allow a full “container” or “tenant” deletion.	Comply with security – data classification and encryption
482	The system should allow record deletion of a structured file within “raw”, “prepared”, or “use-cases” storage.	Comply with security – data classification and encryption
483	The system should definitively delete a record when the data retention period allowed is over.	Comply with security – data classification and encryption
484	The system should log any file deletion event.	Comply with security – infrastructure baseline
485	The system should keep deleted files for a defined amount of time. Rationale: Prevent the deletion of mistakes made with the right to be forgotten.	Comply with security – data classification and encryption
486	The system should keep deleted records for a defined amount of time. Rationale: Prevent the deletion mistakes made with the right to be forgotten.	Comply with security – data classification and encryption

ID	Description	Objective
487	The system should log any record deletion events.	Comply with security – infrastructure baseline
488	The system should provide file logs of deletions on request.	Comply with security – infrastructure baseline
489	The system should provide records of deleted logs on request.	Comply with security – infrastructure baseline

7.6.5 Data Processing – Masking

Anonymizing or pseudonymizing BI & Analytics data.

Table 38: Data Processing – Masking

ID	Description	Objective
500	The system should be able to generate an anonymized copy of a full dataset – i.e., perform a definitive masking – choosing basic anonymization techniques operated at record/line level (masking, generalization), columns to anonymize, and decide on schema alteration (change type, break foreign keys). Rationale: Store data for analytic use-cases; provide data for development or test environments.	Comply with security – data classification and encryption
501	The system is able to anonymize a record in a dataset without altering column types, alteration of foreign key should be a decision – i.e., perform a definitive masking – choosing basic anonymization techniques (masking, generalization, etc.), columns to anonymize, and using record identification keys. Rationale: Deal with data "to be forgotten" and avoid any impact on downstream data workflow; support use-cases mixing use of personal data and analytics processable to work on anonymized data.	Comply with security – data classification and encryption
502	The system should be able to generate an anonymized copy of a full dataset that provides basic anonymization on datasets – i.e., perform a definitive masking – choosing advanced anonymization techniques operated at record/line/table level (randomization, advanced algorithm), columns to anonymize, decide on schema alteration (change type, break foreign keys). Rationale: Store data for analytic use-caess; provide data for development or test environments.	Comply with security – data classification and encryption
503	The system enables control point and rollback during the anonymization process of a dataset. Rationale: Avoid data loss.	Comply with security – data classification and encryption
504	The system should log the anonymization processes executed. Rationale: Prove the execution of anonymization.	Comply with security – infrastructure baseline

ID	Description	Objective
505	The system should provide capabilities that allow estimation of anonymization processing time upfront. Rationale: Plan anonymization processing.	Comply with security – infrastructure baseline
506	The system should provide capabilities to perform a definitive masking/anonymization on a semi-structured/hierarchical file (JSON, XML).	Comply with security – data classification and encryption
507	The system should provide capabilities to perform a definitive masking/anonymization on a structured file/delimited (e.g., table export).	Comply with security – data classification and encryption
508	The system should provide capabilities to perform a definitive masking/anonymization on a table/partition (Hive, Impala®).	Comply with security – data classification and encryption
509	The system should control the user (final user or application) authorization to anonymize data, taking into account access control policies.	Comply with security – identity and access management
510	The system preserves the data consistency of a set of datasets (table/files) when anonymizing using a deterministic approach that allows keeping the link with the key. Rationale: Keep the link between sets of data when the key of the link has to be anonymized.	Comply with security – data classification and encryption
512	The system exposes a GUI. Rationale: Target integration with a Metadata Management tool.	Comply with security – data classification and encryption
513	The system processes anonymization in an asynchronous way. Rationale: Avoid locking by other process/performance optimization.	Comply with security – data leakage
514	The system stacks the request and piles similar anonymization/pseudonymizing requests (by similar datasets, at container level, etc.). Rationale: Optimize batches performances to take into account the physical storage (file or table partition).	Comply with security – data leakage
515	The system factorizes requests on files with the same structure.	Comply with security – data classification and encryption
516	The system should split batch processing to fit in a chosen timeframe (night, weekend). Rationale: Avoid write conflict.	Comply with security – data classification and encryption

ID	Description	Objective
517	The system should require a single code deployment in production with the required parameters: dataset to anonymize, method to apply: • A multiple source code generator (one code by dataset to be compiled and deployed) • A dataset-level metadata generator as input parameters to a unique compiled code (used for all datasets) Rationale: Limit deployment and maintenance effort.	Improve operational efficiency
518	The system should allow pseudonymizing of a dataset – i.e., perform a reversible masking – choosing pseudonymizing techniques (generalization or randomization), and columns to pseudonymize. Rationale: Data concentration.	Comply with security – data leakage
519	The system should provide the capabilities to come back to the initial data before the pseudonymizing process. Rationale: Data concentration.	Comply with security – data leakage
520	The system should allow the exposition of an anonymized records identifier to allow the cascade anonymization of downstream applications. Rationale: Cascade anonymization order.	Comply with security – data classification and encryption
521	The system should be able to anonymize a given dataset for two different businesses that could have the same format or use the same rule but cannot be joined. Rationale: Protect data domains.	Comply with security – data leakage

7.6.6 Data Processing – Metadata Inferring

Process of deducing properties of data such as type, classification, metadata, etc. through the analysis of the data itself.

Table 39: Data Processing – Metadata Inferring

ID	Description	Objective
540	The system should infer – discover automatically – relationships between tables, files, etc. (interpreting same values, repartition, etc.). Rationale: Accelerate the initialization of the metadata repository and ensure its maintenance; reduce manual work.	Improve operational efficiency
541	The system should identify metadata automatically from patterns (email, phone number, credit cards, etc.).	Improve operational efficiency

ID	Description	Objective
542	The system should identify metadata automatically from available dictionaries (e.g., company name, city, first name, last name, synonyms, etc.). Rationale: Accelerate the initialization of the metadata repository and ensure its maintenance; reduce manual work.	Improve operational efficiency
543	The system should infer personal data information.	Comply with security – data classification and encryption
544	The system should be able to infer sensitive (e.g., personal) data in unstructured files.	Comply with security – data classification and encryption

7.6.7 Data Processing – Near-Real-Time Processing

Low-latency data processing – less than 24 hours/day (includes micro batch and near-real time with minute or second latency).

Table 40: Data Processing – Near-Real-Time Processing

ID	Description	Objective
560	The system should be able to persist streamed data.	Store data as an asset, including metadata
561	The system should be able to stream back data.	Improve customer or partner satisfaction
562	The system should provide basic data stream processing for structured data (query execution, joining, filtering, sorting).	Improve operational efficiency
563	The system should provide basic data stream processing for unstructured data (search/select by regex, parsing, stemming, word distribution, filtering).	Improve operational efficiency
564	The system should allow the real-time use of a trained model to get a prediction related to the streamed data. Rationale: Some use-cases require real-time access.	Improve operational efficiency
565	The system should stream process data in near-real time (down to three (3) seconds). Rationale: Enable real-time use-cases.	Guarantee quality of service
566	The system should anonymize or pseudonymize personal data coming from data stream ingestion (provided by an entity self-service configuration).	Comply with security – data leakage
567	The system should present the data leveraging metadata (schema) to make them available in a format that allows queries in an SQL-like way.	Store data as an asset, including metadata

ID	Description	Objective
568	The system should be able to adapt on-the-fly to a new version of the metadata (schema) sent in the stream with the data. Rationale: Adapt to source systems evolutions.	Reduce use-case time-to-market
569	The system should process data in micro batches with an hour latency (several times a day).	Improve operational efficiency
570	The system should process data in near-real time with a minute latency.	Improve operational efficiency
571	The system should process data in near-real time with a second latency.	Improve operational efficiency

7.6.8 Data Processing – Quality Profiling

Profiling of data quality through statistical techniques.

Table 41: Data Processing – Quality Profiling

ID	Description	Objective
580	The system should enable scheduling of a data quality check based on quality rules (business or statistical).	Data quality is under control
581	The system should provide UIs to navigate the results of quality profiling.	Data quality is under control
582	The system should expose the result of profiling in a readable format. Rationale: Profiling that is not exposed through tables and kept in files cannot be exploited.	Data quality is under control
583	The system should provide statistical data quality profiles (checking null frequency, duplicate keys, foreign keys, data distribution, belongs to a list of value, min, max, etc.) at different stages of BI & Analytics (from inbound to outbound). Rationale: The quality profiling allows resolution of a quality issue by identifying inconsistency.	Data quality is under control

7.6.9 Data Processing – Statistical Computing

Data science and statistical processing (classification, clustering of data using statistics, data science modeling, including machine learning).

Table 42: Data Processing – Statistical Computing

ID	Description	Objective
600	The system should allow the execution of a statistical analysis algorithm. Rationale: Enable statistical use-cases leveraging BI & Analytics data.	Generate business value leveraging new data capabilities (e.g., data science, etc.)
601	The system should allow the execution of a machine learning algorithm. Rationale: Enable machine learning use-cases leveraging BI & Analytics data.	Generate business value leveraging new data capabilities (e.g., data science, etc.)

7.6.10 Data Processing – Technical Preparation

Business-agnostic technical processing common to all use-cases (i.e., type format, cleansing, integration).

Table 43: Data Processing – Technical Preparation

ID	Description	Objective
620	The system should provide a data ingestion service to aggregate files to reduce the final number of files in tenants. Rationale: Optimization of the storage.	Reduce the TCO
621	The system should allow data cleansing based on statistical/relational rules based on row, column, table, and sliding windows-based attributes.	Data quality is under control
622	The system should allow data cleansing based on business rules.	Data quality is under control
623	The system should allow the harmonization of dates (ISO 8601[1]) before ingestion.	Store data as an asset, including metadata
624	The system should allow the creation of table structures for structured data.	Store data as an asset, including metadata
625	The system should uniformly input data fields and columns (date unification, decimal and real unification, separator unification).	Store data as an asset, including metadata

[1] ISO 8601: Date and Time Format; refer to: https://www.iso.org/iso-8601-date-and-time-format.html.

ID	Description	Objective
626	The system should isolate delta between two snapshots received at two different times and return a consistent copy of the data source, including historical changes. Rationale: Keep the memory of previous values when data is updated.	Store data as an asset, including metadata
627	The system should allow parsing of an XML file and publication through a table to perform queries. Rationale: Preparation of data for analysis.	Store data as an asset, including metadata
628	The system should uniformize ingested files (UTF8 transcoding, field and line terminator, header and footer removal, and Parquet format). Rationale: Preparation of data for analysis.	Store data as an asset, including metadata
629	The system should maximize the automatization of data preparation. Rationale: Preparation of data for analysis.	Improve operational efficiency

7.7 Data Storage

7.7.1 Data Storage – Information Data Storage

Storage of data that has been prepared for specific business use-cases: insights, features calculated by data science, aggregations. Includes the capability to execute low-latency queries for streaming jobs.

Table 44: Data Storage – Information Data Storage

ID	Description	Objective
720	The system should allow the preparation of data in a use-case table format with schema, column, and defined types to enable SQL-like queries on the data. Rationale: Prepare data for a specific use-case.	Store data as an asset, including metadata
721	The system should create business use-cases in preparation of data. Rationale: Prepare data for a specific use-case.	Store data as an asset, including metadata
722	The system should allow low-latency queries on the data for jobs working in a stream mode. Rationale: Prepare data for a specific use-case using batch streaming.	Guarantee quality of service

7.7.2 Data Storage – Prepared Data Storage

Storage of data that has been processed applies standard technical rules that are business-agnostic: types formats, management of null value, generation of unique keys, etc.

Table 45:Data Storage – Prepared Data Storage

ID	Description	Objective
740	The system should carry out business-agnostic data preparation that can support several use-cases. Rationale: Prepare data once to serve several use-cases.	Reduce use-case time-to-market
741	The system should allow preparation of data in a raw table format with schema, column, and defined types to enable SQL-like queries on the data. Rationale: Enable SQL-like queries on the data.	Store data as an asset, including metadata

7.7.3 Data Storage – Raw Data Storage

Storage of source data without any change.

Table 46: Data Storage – Raw Data Storage

ID	Description	Objective
760	The system should store raw data systematically before the transformation of any data. Rationale: Big data principle and big practice provide a single baseline for all usage.	Data quality is under control
761	The system should associate technical metadata (schema) with the raw data to allow table exposition (e.g., Avro, Parquet). Rationale: Required for data transformation into a query-executionable format.	Store data as an asset, including metadata
762	The system should provide a standard tree-folder organization to enable different access rights for different actors and allow collaboration.	Comply with security – identity and access management
763	The system should store all extracted datasets in a readable and open format (e.g., UTF-8 XML, CSV, etc.). Rationale: Software engineering efficiency.	Store data as an asset, including metadata
764	The system should store flat file-based data (e.g., CSV, TXT, JSON).	Store data as an asset, including metadata
765	The system should store binary file-based data (e.g., SAS datasets, Portable Document Format (PDF) files).	Store data as an asset, including metadata

ID	Description	Objective
766	The system should store small file-based data (e.g., Microsoft Office 365® documents).	Store data as an asset, including metadata
768	The system should collect all raw data when configuring a permanent connector for an operational system, even if no specific use-case is planned. Rationale: Support data discovery and improve use-cases time-to-market.	Store data as an asset, including metadata

7.7.4 Data Storage – Archiving and Compression

Archiving and compression functions to manage the data lifecycle and optimize storage resources.

Table 47: Data Storage – Archiving and Compression

ID	Description	Objective
780	The system should archive data according to defined rules. Rationale: Limit storage.	Reduce the TCO
781	The system should be able to delete archived data at dataset level. Rationale: Limit storage.	Reduce the TCO

7.7.5 Data Storage – Sensitive Data Storage

Storage of sensitive data; e.g., personal, financial, strategy, business, etc.

Table 48: Data Storage – Sensitive Data Storage

ID	Description	Objective
840	The system should isolate storage of personal data from non-personal data.	Comply with security – data classification and encryption
841	The system should tag personal data with privacy classification metadata (initialization by a data steward and running automatically).	Comply with security – data classification and encryption
842	The system should avoid multiple copies of individual personal data.	Comply with security – data classification and encryption
843	The system should provide protection of personal data according to security requirements.	Comply with security – data leakage
844	The system should provide a view to aggregate transparently personal data and non-personal data.	Comply with security – data classification and encryption

7.7.6 Data Storage – Stored Data Encryption

Protection of data within a node through encryption of the data at rest.

Table 49: Data Storage – Stored Data Encryption

ID	Description	Objective
860	The system should allow file/volume encryption with a security-approved encryption algorithm with acceptable impact on treatments performance. Rationale: Secure data at storage level.	Comply with security – data classification and encryption
861	The system should allow field/column encryption with an AXA security approved encryption algorithm with acceptable impact on treatments performance. Rationale: Secure data at storage level.	Comply with security – data classification and encryption
862	The system should securely manage access policies. Rationale: Secure data at storage level.	Comply with security – identity and access management
863	The system should allow data decryption only for authorized users and processes. Rationale: Secure data at storage level.	Comply with security – data classification and encryption
864	The system should be able to decrypt file/volume.	Comply with security – data classification and encryption
865	The system should allow at rest encryption with dedicated Hardware Security Module (HSM). Rationale: Topics: encryption.	Comply with security – data classification and encryption
866	The system should securely manage encryption keys. Rationale: secure data at storage level	Comply with security – data classification and encryption

7.8 Decision Support Analysis and Visualization

7.8.1 Decision Support Analysis and Visualization – Statistical Analysis

Apply dynamic statistical "on-the-shelf" models to information available online.

Table 50: Decision Support Analysis and Visualization – Statistical Analysis

ID	Description	Objective
1640	The system should allow applying statistics to graphical visualizations (e.g., time series projections). Rationale: Avoid poor user experience because of bad performance (data transfer delay, etc.).	Improve operational efficiency

7.8.2 Decision Support Analysis and Visualization – Visual Data Discovery

Dynamic drill down to navigate the data and find the right chart, graphic, etc.

Table 51: Decision Support Analysis and Visualization – Visual Data Discovery

ID	Description	Objective
1660	The system should allow interactive query execution on large datasets with drill-down capability. Rationale: Avoid poor user experience because of bad performance (data transfer delay, etc.).	Democratize use of data
1661	The system should allow a shift from synthetic data presentations in graphics to more detailed tabular ones.	Democratize use of data

7.9 Information Reporting and Dashboarding

7.9.1 Information Reporting and Dashboarding – Formatted Online Reporting

Ready-to-use reports and charts that can be refreshed interactively to use data with no specific data management skills.

Table 52: Information Reporting and Dashboarding – Formatted Online Reporting

ID	Description	Objective
1520	The system should be able to export the data of a given person as requested by an external party.	Comply with regulations (e.g., privacy, Solvency II, etc.)
1521	The system should allow diffusion scheduling of paginated printable static reports; e.g., PDF, Microsoft PowerPoint® (PPT), and Microsoft Excel® (XLS).	Democratize use of data
1522	The system should provide reporting for regulators (e.g., Solvency II QRT XBRL format).	Comply with regulations (e.g., privacy, Solvency II, etc.)
1523	The system should apply information context/customization before diffusion.	Democratize use of data
1524	Online (to be refreshed with filters) report with advanced layout that allows printing, with pagination, democratized access. Rationale: Easy data access.	Democratize use of data

7.9.2 Information Reporting and Dashboarding – Information Servicing

Storage of information data in a format adapted to queries (cube, in memory, view, etc.).

Table 53: Information Reporting and Dashboarding – Information Servicing

ID	Description	Objective
1540	The system should provide a business semantic layer.	Democratize use of data
1541	The system should apply security rules that allow access to the information to a given identified user only when authorized.	Comply with security – identity and access management

7.9.3 Information Reporting and Dashboarding – Report Diffusion Scheduling

Ready-to-use static reports, charts generated periodically, usually paginated and in a printable format.

Table 54: Information Reporting and Dashboarding – Report Diffusion Scheduling

ID	Description	Objective
1560	The system should allow massive printing. Rationale: Allow access of the report through a standard workstation toolset.	Democratize use of data
1561	The system should generate reports and charts generated periodically in batch mode. Rationale: Provide ready-to-use updated reports.	Improve operational efficiency
1562	The system should allow report pagination including tabs, with graphics presented in reports that can have several pages. Rationale: Provide ready-to-use updated reports.	Improve operational efficiency

7.9.4 Information Reporting and Dashboarding – Self-Service Reporting

Build reports and refresh them with data. Ability for a business user – with no SQL skills, and a drag-and-drop UI – to create a report from available information and share it with other users.

Table 55: Information Reporting and Dashboarding – Self-Service Reporting

ID	Description	Objective
1580	The system should enable users to download reports in human-readable format.	Democratize use of data
1581	The system should allow the use of Microsoft Excel integrated in Office 365 to query execution the data.	Democratize use of data

7.10 Integration and Deployment

7.10.1 Integration and Deployment – Build Management

Prepare the package to be deployed: run the test and generate binaries.

Table 56: Integration and Deployment – Build Management

ID	Description	Objective
1220	The system should allow automated (i.e., without manual access to the platform) packaging of code releases (Python®, R®, Scala) delivered in a code repository on request of authorized developers on BI & Analytics. Rationale: Continuous integration.	Reduce use-case time-to-market
1221	The system should allow the execution of an automated integration test at deployment time, preventing deployment of code not passing the test. Rationale: The model must be tested before its deployment in the production environment.	Reduce use-case time-to-market
1222	The system should allow the execution of an automated unitary test at deployment time, preventing deployment of code not passing the test. Rationale: Secure SLA.	Reduce use-case time-to-market
1223	The system should be able to access public artifacts and packages on the web to perform the build. Rationale: Leverage existing available free packages on the Internet.	Reduce use-case time-to-market
1224	The system should allow the storage of program binaries delivered. Rationale: Continuous integration.	Reduce use-case time-to-market
1225	The system should allow the parametrization of job scheduling when configuring the deployment workflow.	Reduce use-case time-to-market

7.10.2 Integration and Deployment – Deployment

Deploy a package – new software release – on a given environment (development/integration/pre-production/production).

Table 57: Integration and Deployment – Deployment

ID	Description	Objective
1240	The system should allow automated (i.e., without manual access to the platform) deployment of code releases (Python, R, Scala) upon the request of authorized users on the shared data platform cluster. Rationale: Continuous integration.	Reduce use-case time-to-market
1241	The system should log all deployment operations on the platform. Rationale: Automation allows tracking of operations on production, which is not possible when deployment is done manually.	Comply with security – infrastructure baseline
1242	The system should provide a token identifying the tenant/container and providing authorization for code to be deployed at tenant/container creation time for projects teams. Rationale: Increase the pace of release delivery to deliver better service to the customer.	Comply with security – identity and access management
1243	The system should provide capability to recover the token identifying the tenant/container and providing authorization to project teams for code to be deployed when lost. Rationale: Increase the pace of release delivery to deliver better service to the customer.	Comply with security – identity and access management
1244	The system should allow the creation of a tenant/container and the automation of binary code release updates. Rationale: Increase the pace of release delivery to deliver better service to the customer.	Reduce use-case time-to-market
1245	The system should allow a user to deploy only in the tenants/containers of the entity where they are authorized. Rationale: Multi-tenancy.	Comply with security – identity and access management
1246	The system should expose services to allow the scheduling of jobs on BI & Analytics. Rationale: Fully automated delivery.	Reduce use-case time-to-market
1247	The system should provide a capability to return test (non-regression, etc.) result when executed. Rationale: Test automation.	Improve operational efficiency
1248	The system should allow the parameterization of job scheduling to allow planned execution in production.	Reduce use-case time-to-market

7.10.3 Integration and Deployment – Source Code Management

Versioning of development code (trunk) and release (branch) for developers or data scientists.

Table 58: Integration and Deployment – Source Code Management

ID	Description	Objective
1260	The system should version source code of the different applications running on the platform developed either by developers or data scientists. Rationale: Software engineering industrialization.	Reduce use-case time-to-market
1261	The system should allow version and archive documentation generated from source code. Rationale: Documentation traceability.	Improve operational efficiency
1262	The system should allow code sharing (e.g., Python, R, Scala) among developers and data scientist communities. Rationale: Maximize code reuse.	Improve operational efficiency
1263	The system should allow sharing features developed (e.g., Python or R) among a group of data scientists or development community.	Improve operational efficiency
1264	The system should support the transformation of code produced during data science exploratory phases to code for industrialized processing.	Reduce use-case time-to-market

7.10.4 Integration and Deployment – Testing

Execution of tests on the new code release (non-regression).

Table 59: Integration and Deployment – Testing

ID	Description	Objective
1280	The system should execute tests (unitary, non-regression, etc.) to check the validity of code deployed. Rationale: Test automation.	Improve operational efficiency

7.11 Metadata Management

Capabilities applicable to full "Metadata Management" scope such as storage, stewardship, reporting, and administration of metadata.

Table 60: Metadata Management

ID	Description	Objective
1700	The system should allow management of metadata in a multi-tenant architecture. Rationale: Enable sharing of language capabilities between entities.	Store data as an asset, including metadata
1701	The system should provide full web zero-footprint UIs for end-users. Rationale: Usability.	Improve operational efficiency
1702	The system should be translated into major languages (English, French, German, Japanese, etc.). Rationale: Local language.	Democratize use of data

7.11.1 Metadata Management – Business Metadata Storage

Storage of business metadata: non-technical information about data. Capabilities applicable to full "Business Metadata Storage" scope.

Table 61: Metadata Management – Business Metadata Storage

ID	Description	Objective
1720	The Metadata Management should allow management of the "Privacy Policy" metadata.	Comply with security – data classification and encryption

7.11.2 Metadata Management – Business Metadata Storage – Data Classification

Storage of data classification: privacy, security, valuable assets, etc.

Table 62: Metadata Management – Business Metadata Storage – Data Classification

ID	Description	Objective
1740	The security should allow management of "Security Classification" metadata. Rationale: Increase knowledge and mastership of data.	Comply with security – data classification and encryption
1741	The system should allow management of "Privacy Classification" metadata.	Comply with security – data classification and encryption

ID	Description	Objective
1742	The system should allow management of “Business Criticality, Valuable Assets“ metadata. Rationale: Increase knowledge and mastership of data.	Comply with security – data classification and encryption

7.11.3 Metadata Management – Business Metadata Storage – Data Glossary

Storage of business metadata, such as business name, definition. Uniqueness should be respected in the data glossary.

Table 63: Metadata Management – Business Metadata Storage – Data Glossary

ID	Description	Objective
1760	The system should allow management of “Data Provider” metadata. Rationale: Increase knowledge and mastership of data.	Store data as an asset, including metadata
1761	The system should allow enrichment with business metadata (tag, classification, etc.).	Store data as an asset, including metadata
1762	The system should allow management of “Business Definition, Business Glossary” metadata. Rationale: Identify personal data.	Comply with security – data classification and encryption
1763	The system should allow management of “Business Data Classification” metadata. Rationale: Increase knowledge and mastership of data.	Store data as an asset, including metadata
1764	The system should allow management of “Business Object Model/Conceptual Data Model” metadata. Rationale: Increase knowledge and mastership of data.	Store data as an asset, including metadata
1765	The system should allow management of “Acronym/Synonym” metadata. Rationale: Increase knowledge and mastership of data.	Store data as an asset, including metadata
1766	The system should allow management of unique enterprise-level “Business Name” metadata that can be an object (like a person) or a field (date of birth). Rationale: Increase knowledge and mastership of data.	Store data as an asset, including metadata

7.11.4 Metadata Management – Business Metadata Storage – Data Quality

Storage of business data quality rules and measurement.

Table 64: Metadata Management – Business Metadata Storage – Data Quality

ID	Description	Objective
1780	The system should allow management of "Deletion/Retention Date" metadata for deletion rules.	Comply with security – data leakage
1781	The system should allow management of "Data Quality Rule" metadata. Rationale: Increase knowledge and mastership of data.	Data quality is under control

7.11.5 Metadata Management – Business Metadata Storage – Data Stewardship

Storage of information, such as the stakeholders in charge of data management.

Table 65: Metadata Management – Business Metadata Storage – Data Stewardship

ID	Description	Objective
1800	The system should allow management of "Owner" metadata. Rationale: Increase knowledge and mastership of data.	Store data as an asset, including metadata
1801	The system should allow management of "Data Steward" metadata. Rationale: Increase knowledge and mastership of data.	Data quality is under control

7.11.6 Metadata Management – Business Metadata Storage – Data Usage

Storage of data usage activities, including processing identification of personal data.

Table 66: Metadata Management – Business Metadata Storage – Data Usage

ID	Description	Objective
1820	The system should allow management of "Data Usage Declared to Authorities" metadata.	Comply with security – data classification and encryption
1821	The system should allow management of data usage.	Comply with regulations (e.g., privacy, Solvency II, etc.)

7.11.7 Metadata Management – Business Metadata Storage – Reference Data

Storage of reference data required as examples for reporting.

Table 67: Metadata Management – Business Metadata Storage – Reference Data

ID	Description	Objective
1840	The system should allow management of "Code List/Field Values", including hierarchy metadata. Rationale: Increase knowledge and mastership of data.	Store data as an asset, including metadata

7.11.8 Metadata Management – Data Modeling

Capability to design data models.

Table 68: Metadata Management – Data Modeling

ID	Description	Objective
1860	The system should generate the database creation script automatically from a designed model. Rationale: Easy evolution, loose-coupling, and interoperability.	Reduce use-case time-to-market
1861	The system should provide a data modeling capability for data engineers/data scientists. Rationale: Data model preparation.	Improve operational efficiency
1862	The system should allow reverse engineering of the schema stored in BI & Analytics. Rationale: Data model preparation.	Improve operational efficiency
1863	The system should provide data modeling guidelines. Rationale: Data model preparation.	Improve operational efficiency

7.11.9 Metadata Management – Metadata Reporting

Generation of reports on Metadata Management.

Table 69: Metadata Management – Metadata Reporting

ID	Description	Objective
1880	The system should allow the identification of records that have reached the end of the retention date.	Comply with security – data classification and encryption
1881	The system should perform online queries on metadata in batch mode. Rationale: Usability.	Improve operational efficiency

ID	Description	Objective
1882	The system should provide a report on metadata updates with metadata lineage visualization on the full information value chain (be able to show lineage in front of and behind BI & Analytics as well as within it). Rationale: Control metadata process.	Store data as an asset, including metadata
1883	The system should measure and report on metadata quality (number of attributes per source system, completeness of the metadata, etc.). Rationale: Improve the quality and completeness of metadata.	Improve operational efficiency
1884	The system should allow the creation of out-of-the-box reporting (graphical or textual) with an office-readable export format (e.g., Microsoft Word, Excel). Rationale: Interoperability: enable access to metadata stored in the repository for systems/persons.	Improve operational efficiency
1885	The system should provide an audit report on metadata usage. Rationale: Improve the quality and completeness of metadata.	Improve operational efficiency
1886	The system should offer the ability to create reporting with customizable filters. Rationale: Usability.	Improve operational efficiency
1887	The system should be able to interface BI tools that allow query, reporting, etc.. Rationale: Interoperability: enable access to metadata stored in the repository for systems/persons.	Improve operational efficiency
1888	The system should provide a graphical view of relations between metadata (data model, graph). Rationale: Usability.	Improve operational efficiency
1889	The system should provide multi-criteria search on metadata. Rationale: Increase knowledge and mastership of data.	Store data as an asset, including metadata
1890	The system should display data lineage graphically. Rationale: Auditability of the platform.	Store data as an asset, including metadata

7.11.10 Metadata Management – Metadata Stewardship

Addition of business metadata related to data.

Table 70: Metadata Management – Metadata Stewardship

ID	Description	Objective
2000	The system should manage (create, update, delete) business metadata (glossary, classification – security, privacy) of the data (raw, information, use-case).	Store data as an asset, including metadata
2001	The system should enable metadata CRUD. Rationale: Increase knowledge and mastership of data.	Store data as an asset, including metadata
2002	The system should link technical metadata and operational metadata with business metadata. Rationale: Make value from metadata.	Store data as an asset, including metadata
2003	The system should propose free tagging on metadata with auto-completion that can be changed by authorized users. Rationale: Increase knowledge and mastership of data.	Store data as an asset, including metadata
2004	The system should enable reference data (list of values) CRUD. Rationale: Improve the quality and completeness of metadata.	Democratize use of data
2005	The system should show metadata change impact; for example, if the business definition is changed, how many technical definitions are based on it? Rationale: Control metadata process.	Store data as an asset, including metadata
2006	The system should enable the adding of Uniform Resource Locator (URL) links to external documentation references in the metadata. Rationale: Increase knowledge and mastership of data.	Store data as an asset, including metadata
2007	The system should enable the creation and comparison of different branches (development, User Acceptance Testing (UAT), production, etc.). Rationale: Improve the quality and completeness of metadata.	Improve operational efficiency
2008	The system should save the history, source, and author of metadata updates. Rationale: Auditability of the platform.	Comply with security – infrastructure baseline
2009	The system should be able to create mapping rules between technical metadata (e.g., convert Oracle® data types to Hive data types). Rationale: Accelerate the initialization of Metadata Management and ensure its maintenance; reduce manual work.	Improve operational efficiency

ID	Description	Objective
2010	The system should manage concurrency access.	Guarantee quality of service
2011	The system should be able to compare different branches (development, UAT, production, etc.). Rationale: Improve the quality and completeness of metadata.	Improve operational efficiency

7.11.11 Metadata Management – Repository Administration

Administration of Metadata Management (workflow, authorization, etc.).

Table 71: Metadata Management – Repository Administration

ID	Description	Objective
2020	The system should allow the customization of technical and business metadata models (attributes, relations, unstructured documents, etc.) to add new metadata types or attributes, and benefit by design of standard functionalities. Rationale: Easy evolution, loose-coupling, and interoperability.	Reduce use-case time-to-market
2021	The system should enable the definition of metadata classification by the data manager. Rationale: Create value from metadata.	Reduce use-case time-to-market
2022	The system should enable the modification of validation rules of metadata by data manager. Rationale: Easy evolution, loose-coupling, and interoperability.	Reduce use-case time-to-market
2023	The system should enable the creation of metadata validation workflows by the data manager.	Store data as an asset, including metadata
2024	The system should send an email notification to the data manager when new events appear from the validation workflow. Rationale: Control metadata process.	Store data as an asset, including metadata
2025	The system should display the status and tasks of the metadata validation workflow (e.g., candidate, in-review, accepted). Rationale: Control metadata process.	Store data as an asset, including metadata
2026	The system should display a customized view depending on the user role. Rationale: Easy evolution, loose-coupling, and interoperability.	Reduce use-case time-to-market
2027	The system should enable authorizations of Metadata Management depending on roles, including batch or user role. Rationale: Security and accessibility of metadata.	Comply with security – identity and access management

ID	Description	Objective
2028	The system should support external system authentications. Rationale: Security and accessibility of metadata.	Comply with security – identity and access management
2029	The system should enable the creation of a user or group, and give authorizations on Metadata Management depending on roles, including batch or user role. Rationale: Security and accessibility of metadata.	Comply with security – identity and access management

7.11.12 Metadata Management – Technical Metadata Storage – Data Access

Storage of data access permissions.

Table 72: Metadata Management – Technical Metadata Storage – Data Access

ID	Description	Objective
2060	The system should allow management of "Data Access and Privileges" metadata. Rationale: Increase knowledge and mastership of data.	Store data as an asset, including metadata

7.11.13 Metadata Management – Technical Metadata Storage – Data Dictionaries

Storage of data dictionaries (databases, tables, columns).

Table 73: Metadata Management – Technical Metadata Storage – Data Dictionaries

ID	Description	Objective
2080	The system should allow management of "Technology/Application" metadata. Rationale: Increase knowledge and mastership of data.	Store data as an asset, including metadata
2081	The system should allow management of "Physical Location" metadata. Rationale: Increase knowledge and mastership of data.	Store data as an asset, including metadata
2082	The system should allow management of "Data Structure (schema: table, columns, type (string, integer, date, etc.))" metadata. Rationale: Increase knowledge and mastership of data.	Store data as an asset, including metadata
2083	The system should allow management of "Primary Key" (unique key) metadata. Rationale: Increase knowledge and mastership of data.	Store data as an asset, including metadata

ID	Description	Objective
2084	The system should allow management of “Data Relation” (foreign key) metadata. Rationale: Increase knowledge and mastership of data.	Store data as an asset, including metadata

7.12 Monitoring

7.12.1 Monitoring – Business Usage Monitoring

Monitoring of the business services of the platform (supervise data science model drifting, etc.).

Table 74: Monitoring – Business Usage Monitoring

ID	Description	Objective
1320	The system should monitor that the model is still performing according to business expectations (detect model drifting, etc.). Rationale: Supervise the efficiency of data science.	Improve customer or partner satisfaction
1321	The system should monitor the performance of analytical models in different environments (UAT/production). Rationale: Monitor processing performance.	Guarantee quality of service

7.12.2 Monitoring – Data and Application Monitoring

Monitoring of the data (e.g., quality) and application layer (e.g., availability of services).

Table 75: Monitoring – Data and Application Monitoring

ID	Description	Objective
1340	The system should provide reporting on data processing execution in the different environments (UAT, production, etc.). Rationale: Monitor processing performance.	Guarantee quality of service
1341	The system should provide reporting on data load.	Data quality is under control
1342	The system should provide reporting on data leakage within 72 hours for authorities.	Comply with security – data leakage
1343	The system should provide reports on deleted data.	Comply with security – infrastructure baseline
1344	The system should provide reports on data access (connections and activity). Rationale: Provide answers to audit questions.	Comply with security – identity and access management

ID	Description	Objective
1345	The system should provide reports on usage statistics. Rationale: Know what reports are used.	Guarantee quality of service
1346	The system should allow configuration of alerts to the data manager through workflow or email, whether the data is ingested or not. Rationale: Preparation of data for analysis.	Guarantee quality of service
1347	The system should monitor the flow of data across BI & Analytics (from inbound to outbound). Rationale: Monitor Data Architecture.	Guarantee quality of service
1348	The system should monitor data usage. Rationale: Increase knowledge and mastership of data.	Democratize use of data
1349	The system should provide reporting on the content of BI & Analytics within a given scope (number of files, number of databases, or number of tables). Rationale: Knowledge of the leveraged data for an organization.	Store data as an asset, including metadata
1350	The system should execute data quality checks (consistency, type, etc.) based on data quality business rules in BI & Analytics data in the overall data lifecycle (raw, information, use-cases). Rationale: Identify quality issues to resolve/raise awareness of incorrect data to data stewards.	Data quality is under control
1351	The system should expose the results of data quality monitoring in a UI. Rationale: Monitoring that is not exposed through tables and kept in files cannot be exploited.	Data quality is under control
1352	The system should execute data quality checks based on GDPR rules on BI & Analytics data (raw, prepared, use-cases).	Comply with security – data classification and encryption
1353	The system should allow configuration of alerts to the data steward when the data quality check returns quality issues.	Data quality is under control

7.12.3 Monitoring – Lineage and Audit

Capture of the data lifecycle including the data origin and where it moves over time.

Table 76: Monitoring – Lineage and Audit

ID	Description	Objective
1360	The system should provide data lineage at table level from inbound to outbound (data lifecycle from origins to moves over time) among operational metadata. Rationale: Increase knowledge and mastership of data.	Comply with security – infrastructure baseline
1361	The system should generate "time monitoring statistics": ingestion date for raw data, update frequency, or last update for use-case or information. Rationale: Increase knowledge and mastership of data.	Comply with security – infrastructure baseline
1362	The system should capture job statistics to identify which processes transform which data among operational metadata. Rationale: Increase knowledge and mastership of data.	Comply with security – infrastructure baseline
1363	The system should save data access logs for ten (10) years among operational metadata. Rationale: Increase knowledge and mastership of data.	Comply with security – infrastructure baseline
1364	The system should capture data load statistics among operational metadata. Rationale: Increase knowledge and mastership of data.	Comply with security – infrastructure baseline
1365	The system should log record deletion among operational metadata.	Comply with security – infrastructure baseline
1366	The system should provide data lineage at file level from inbound to outbound (data lifecycle from origins to moves over time) among operational metadata. Rationale: Increase knowledge and mastership of data.	Comply with security – infrastructure baseline
1367	The system should provide data lineage at field level from inbound to outbound (data lifecycle from origins to moves over time) among operational metadata. Rationale: Increase knowledge and mastership of data.	Comply with security – infrastructure baseline
1368	The system should log the processing of personal data among operational metadata.	Comply with security – infrastructure baseline
1369	The system should allow management of the "Acquisition Data" metadata type on external data. Rationale: Increase knowledge and mastership of data.	Store data as an asset, including metadata

7.12.4 Monitoring – Security Monitoring

Monitoring of the security layer.

Table 77: Monitoring – Security Monitoring

ID	Description	Objective
1380	The system should detect data misuse, data breaches, and leakage within 72 hours into the event. Rationale: Detect data leakage.	Comply with security – data leakage

7.12.5 Monitoring – Technical Monitoring

Monitoring of the network layer and infrastructure layer.

Table 78: Monitoring – Technical Monitoring

ID	Description	Objective
1400	The system should monitor the infrastructure of data capabilities (permanent connectors, servers, etc.) to check that they are up-and-running and not overloaded. Rationale: Monitor BI & Analytics architecture level of service.	Guarantee quality of service
1401	The system should check that the network layer is up and running and available at the expected level of availability and performance. Rationale: Monitor ingestion.	Guarantee quality of service
1402	The system should allow the configuration of alerts to notify administrators when the network level is not at the expected service level. Rationale: Monitor distribution.	Guarantee quality of service
1403	The system should publish BI & Analytics availability status in real time. Rationale: Business continuity.	Guarantee quality of service
1404	The system should provide statistics through reports on BI & Analytics (number of connections, volume stored/disk usage, volume transferred, execution time/performance/CPU usage, etc.). Rationale: Optimize resource usage – invoice the service to entities.	Guarantee quality of service

Glossary

ACL	Access Control List
ADM	Architecture Development Method
API	Application Program Interface
B2B2C	Business to Business to Consumer
BI	Business Intelligence
BIM	Business Information Model
CPU	Central Processing Unit
CRM	Customer Relationship Management
CRUD	Create, Read, Update, and Delete
CSV	Comma-Separated Values
DII	Data Integration & Interoperability
ETL	Extract, Transform, Load
ESB	Enterprise Service Bus
GDPR	General Data Privacy Regulations
GUI	Graphical User Interface
HSM	Hardware Security Module
IaaS	Infrastructure as a Service
IDE	Integrated Development Environment
IT	Information Technology
JDBC	Java Database Connectivity
JSON	JavaScript Object Notation
KPI	Key Performance Indicator
MDM	Master Data Management
NOSQL	Not Only Structured Query Language
NPS	Net Promoter Score

ODBC	Open Database Connectivity
OPR	Optical Character Recognition
PaaS	Platform as a Service
RDMS	Relational Database Management System
REST	Representational State Transfer
ROI	Return on Investment
RPO	Recovery Point Objective
SaaS	Software as a Service
SI	System Information
SLA	Service-Level Agreement
SQL	Structured Query Language
SSO	Single Sign On
TB	Terabytes
TCO	Total Cost of Ownership
UAT	User Acceptance Testing
URL	Uniform Resource Locator
UI	User Interface
XML	Extensible Markup Language

Index