POLITECNICO DI TORINO

Engineering and Management course

Master of Science Course

in Engineering and Management

Master of Science Thesis

Integration of IBM Knowledge Catalog into a highly complex analytical workflow and comparison with Data Governance tools for enterprise data management

Supervisors

Prof. Alberto De Marco

Dott. Luca Bregata

Candidate

Carlo Crescenzi

Alla mia Famiglia.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

In an era increasingly driven by data, a robust Data Governance framework is paramount to ensuring data quality, regulatory compliance, and strategic alignment with business objectives. This is particularly critical in data-intensive sectors such as the fashion industry. This thesis investigates the integration of IBM Knowledge Catalog (IKC) into a sophisticated analytical workflow, providing a comparative analysis with other Data Governance tools for enterprise data management.

The study explores Business Intelligence (BI) and ETL (Extract, Transform, Load) processes, which are fundamental in converting raw data into actionable insights that support strategic decision-making. It examines established Data Governance frameworks, key performance indicators (KPIs), and industry standards such as DAMA-DMBOK and COBIT. A structured approach to enterprise data integration is proposed, focusing on metadata tiers—Bronze, Silver, and Gold—which serve as essential mechanisms for standardizing and managing data flow efficiently. Furthermore, the functionalities of IBM Knowledge Catalog are analysed, with a particular focus on its benefits, implementation challenges, and economic implications regarding cost optimization and risk mitigation.

The research documents the implementation of an advanced Data Governance framework at a fashion company, aiming to elevate data quality, regulatory adherence, and strategic cohesion. The project encompasses a SSIS flow, metadata-driven integration, the application of Data Refinery for data optimization and quality assurance, the deployment of an Enterprise Knowledge Catalog, and the utilization of Watson Studio to harness governance frameworks for strategic intelligence.

In conclusion, this thesis underscores the necessity of balancing centralization and decentralization in Data Governance, advocating for a data-driven paradigm that quantifies benefits through empirical performance indicators. It critically addresses challenges associated with initial resource constraints and proposes solutions for a sustainable and adaptable governance model.

1 BUSINESS INTELLIGENCE

In an increasingly data-driven environment, organizations employ Business Intelligence (BI) to transform raw data into meaningful information. BI includes strategies, tools, and processes that enable data analysis and report creation, supporting the decision-making process at both strategic and operational levels (IBM, s.d.). BI is not just technological implementation, but a true paradigm shifts in the way companies extract, process, and use information to improve efficiency and competitiveness.

Business Intelligence (BI), a term popularized in the 1990s by the Gartner Group, allows for the collection, analysis, and visualization of data to improve business performance. Through the analysis of historical data, companies can evaluate sales trends, identify the most profitable products, analyse customer behaviour, and improve operational efficiency.

The integration of Data Governance into BI is essential to ensure the quality, security, and compliance of data, thereby increasing the reliability of analyses and decision-making processes. A structured understanding of the fundamentals of BI, its relationship with DG, and its role within data infrastructures such as Data Warehouse, Data Mart, and Data Lake is essential to grasp its strategic importance.

Business intelligence improves value across various sectors like customer service, finance, healthcare, retail, sales, marketing, security, compliance, statistical analysis, and supply chain management. It enables agents to promptly address client questions, assess an organization's health state, provide prompt responses to urgent problems, enhance cost efficiency in retail, and develop effective promotions and campaigns. Centralized data and dashboards improve accuracy and identify security issues, while statistical analysis can identify trends and inefficiencies in the supply chain.

Data Governance is crucial for managing the interactions between these elements, ensuring data quality, consistency, and compliance. It defines policies for data lineage, metadata management, and access control, reducing fragmentation and inconsistencies that could compromise the effectiveness of BI.

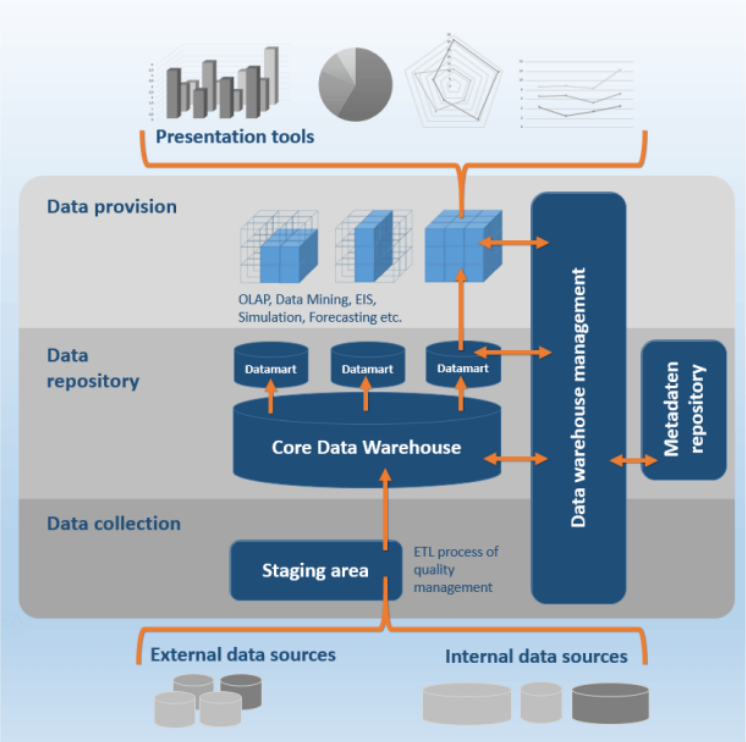


Figure 1:BI Architecture

1.1 BI Workflow

The architecture of a Business Intelligence environment has five distinct components.

1. Operational sources
2. ETL processes
3. DWH (and possible Data Marts)
4. Presentation area and business intelligence applications
5. Machine Learning and AI tools

The preliminary stages of data collecting entail acquiring data from internal and external sources, processing it in the staging area, and applying cleaning, transformation, and integration procedures to guarantee consistency and quality for analysis through ETL procedures (Extract, Transform, Load). The presentation layers comprise tools and user interfaces designed for intuitive data search and visualisation. Dashboards, reports, and analytical tools offer extensive access to corporate data, improving decision-making. The metadata repository contains information regarding data structure, relationships. Security is maintained by role-based access

controls and encryption, while performance management is achieved through intelligent indexing and query optimisation.

A robust Business Intelligence workflow is composed by the following stages:

- Data Sources: identify the data to be examined and analysed, for example from a data warehouse or data lake, cloud, Hadoop, industry statistics, supply chain, CRM, inventory, pricing, sales, marketing, or social. In the next chapters the different data sources will be analysed deeply.
- Data collection: At this stage, the data is subjected to cleaning, transformation, and integration processes to guarantee the consistency and quality required for analysis. This pertains to ETL operations (Extract, Transform, Load), which facilitate the extraction of data, its transformation, and subsequent loading into Data Marts, if applicable, and into the Data Warehouse, the core of the system. The ETL operations are explained in the next chapters.
- Data Analysis: look for trends or unexpected results in the data. This may require the use of data mining, data discovery, or data modelling tools. The integration of analytical models, which leverage Machine Learning and Artificial Intelligence, enhances forecasting capabilities and automates decision support (Compendium, 2024).

  The integration of Machine Learning (ML) and Artificial Intelligence (AI) significantly enhances data analysis capabilities, automating decision-making processes and improving forecasts. Data warehouses, which serve as central repositories of data from various sources, organize information into predefined schemas and lend themselves to joint use with AI and ML systems to identify useful business insights.

  Data analysis, supported by techniques such as data mining, data discovery, and modelling, allows for the identification of trends and unexpected outcomes, enriching data warehouses with valuable information. Artificial Intelligence and Machine Learning are particularly effective in managing large volumes of data, accelerating processing and the identification of complex correlations. ML tools, by monitoring queries, can identify areas for improvement in terms of speed and accuracy in data processing. Moreover, AI automates repetitive tasks such as data integration, monitoring, and cleaning, freeing up IT resources for more

strategic activities. The use of Machine Learning algorithms allows for the simplification of data schemas, reducing the operational costs of the data warehouse.

Analytical models, enhanced by ML and AI, improve predictions and the automation of decision support. Artificial Intelligence can analyse data sources and automatically generating models, saving time and resources for data scientists and improving data accuracy (Anwar, 2024).

In this context, platforms like IBM Cloud Pak for Data, with reference to Watson, provide advanced tools for data analysis and the creation of predictive models. Watson offers Machine Learning capabilities that integrate with IBM Knowledge Catalog, facilitating the discovery of insights and data management (Microsoft, s.d.). In summary, AI and ML allow for deeper and more valuable insights to be obtained from stored data, improving both reporting and forecasting (Ferrari, 2024).

Data Visualization: Visualization is an essential element of an effective Business Intelligence (BI) system, encompassing the development of graphical representations of data, including graphs, tables, and dashboards, with business intelligence tools such as Tableau, Cognos Analytics, Microsoft Excel or Power BI. Essential components of visualization encompass certain tools for generating visual data representations, which may range in complexity from basic spreadsheets such as Microsoft Excel to sophisticated analytical systems like Tableau.

The primary aim of visualization is to render data comprehensible and accessible for people, so aiding in the recognition of trends, anomalies, and notable patterns. Enhanced visualizations from BI systems facilitate the strategic decision-making process, enabling executives and analysts to comprehend data and discern opportunities or issues more effectively. Visualization converts skewed data into significant and usable information, enabling firms to make informed decisions based on empirical facts.

Visualizations can manifest in multiple formats, including bar charts, table charts, line graphs, maps, and interactive dashboards. The selection of visualization type is contingent upon the nature of the data and the message

conveyed. An effective visualization should enable users to examine data at several levels of granularity through functionalities such as drill-down, drill-through, and drill-up. These functions facilitate the study of data from a broad perspective to precise details or to illustrate overarching trends.

- Action plan: develop useful insights based on the analysis of historical data in relation to key performance indicators (KPIs). The actions could include more efficient processes, changes in marketing, resolving supply chain issues, or addressing problems in the customer experience.

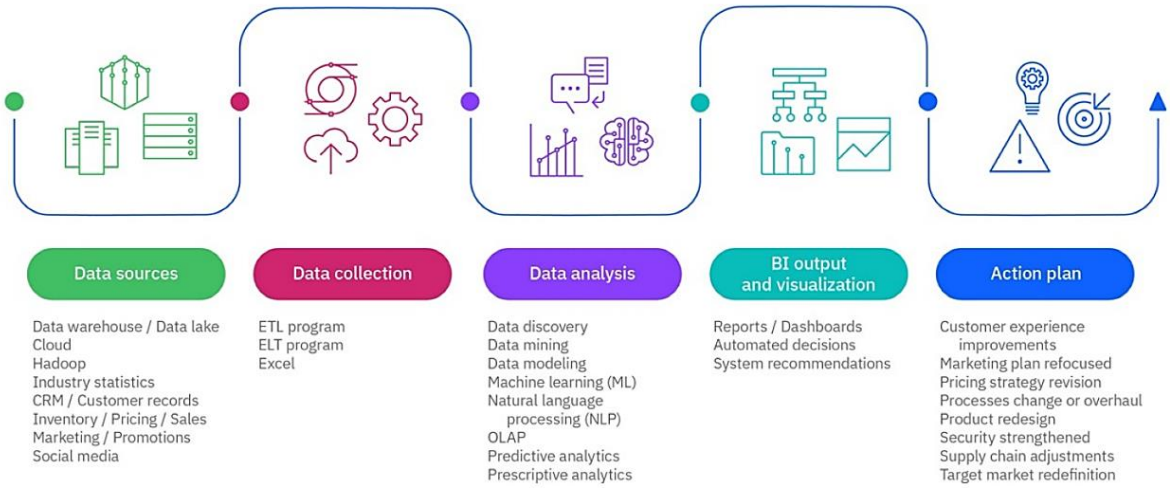


Figure 2: BI Workflow

The effective integration of these architectures allows companies to leverage BI information while maintaining a structured, scalable, and well-governed data ecosystem. In the absence of governance, BI results could present discrepancies, leading to erroneous insights. The integration of all these aspects ensures that Business Intelligence remains a fundamental resource for companies, promoting data-driven strategies and ensuring efficiency, compliance, and governance.

1.2 ETL Process

ETL tools serve to provide a singular, complete, comprehensive, and high-quality data source that subsequently populates the Data Warehouse. The activities they undertake are commonly termed reconciliation, which transpires twice during the Data Warehouse feeding process: initially when the DW is populated and again at regular intervals during updates. Reconciliation comprises four distinct steps referred to as:

- Extraction or Capture
- Cleaning or Scrubbing
- Transformation
- Loading

The distinction between cleaning and transformation is often ambiguous; thus, for clarity, cleaning is defined as the process of rectifying data values, whereas transformation pertains mostly to their format (Bregata, 2019). They are essential operations in a Data Warehouse/Business Intelligence system that facilitate communication between operational systems and the display layer. They comprise the subsequent phases.

1.2.1 Extraction

Data Integration consists of two sub-phases called extraction and cleaning. In the initial subphase, relevant data is extracted from the sources, and this procedure may encompass the following types:

- Static: it is executed when the Data Warehouse requires initial population and represents a snapshot of the operational data.
- Incremental: utilised for the periodic updating of the Data Warehouse, it captures just the alterations that have occurred in the sources since the last extraction. The concept is to use changes documented at the data level to refresh the Data Warehouse. The advantages obtained include the minimal data volume engaged in each operation relative to static extraction, and the fact that most of the data in the Data Warehouse remains constant, with only modified data being analysed. Change Data Capture (CDC) approaches are employed to observe data sources to detect alterations at the data level. These strategies are crucial for Data

Warehouse maintenance because of the transmission of changes identified at the source level.

Cleaning is the sub-phase dedicated to enhancing data quality by removing "dirty" data resulting from duplications, inconsistencies, missing information, erroneous values, and similar issues. The primary data cleansing features of ETL solutions are correction and homogenisation, use specialised dictionaries to rectify spelling mistakes and identify synonyms, as well as rule-based cleaning, which employs domain-specific regulations to ensure accurate value correspondences.

1.2.2 Transformation

This is the pivotal stage of the reconciliation process, intended to transform data from the source operational format to that of the Data Warehouse (DW). This level includes functions for supplying the reconciled data level, such as:

- Conversion and normalisation: they function at both the storage format level and the unit of measurement level to standardise the data.
- Matching: this process establishes correspondences between equivalent fields in disparate sources, verifying the referential integrity of the data (the Foreign Keys in the tables correspond to Primary Keys in the referenced tables).
- Selection: which, if required, diminishes the quantity of fields and records according to the sources. During the selection process, duplicates and merely operational data lacking value for decision support are discarded. The unpopulated "not null" fields are also addressed.

The transformation of data is essential for mapping, which refers to the correspondence between the fields of the source system and those of the destination. During the loading phase, there are two significant distinctions: Normalisation is replaced by denormalization, and aggregation is implemented to facilitate the requisite data synthesis. Normalisation is a crucial procedure in relational database design, intended to structure data to minimise redundancy and enhance integrity. This procedure entails deconstructing the data into smaller, interconnected tables. The primary aim of normalisation is to guarantee data integrity, minimise errors, and enhance the efficiency of insertion, update, and deletion processes. A normalised database facilitates management and enables expedited data access, hence reducing the likelihood of anomalies (Lavecchia, 2023). Denormalization is a procedure that amalgamates data from

various tables into one table. This method is employed to enhance query performance, particularly in situations when read operations greatly exceed write operations. Denormalization increases data redundancy to enhance information retrieval time (PureStorage, s.d.).

Aspect	Normalization	Denormalization
Description	Organizing data to reduce redundancy	Combining data to speed up retrieval
Objective	Improve data integrity and consistency	Increase query speed
Application	Used in OLTP (Online Transaction Processing) systems	Used in OLAP (Online Analytical Processing) systems
Effects	Reduces errors and increases efficiency	Faster read operations at the cost of some writes

Table 1: Normalization vs Denormalization

1.2.3 Loading

The last step of the ETL process involves physical structuring and data collection in destination models. This is essential for obtaining a "certificate" and ensuring that data is ready for analysis and reporting. Data collection can occur in two modes:

- **Refresh**, which involves complete data writing in the Data Warehouse (DWH) and replacement of existing data, ensuring a recent, complete, and intensive collection of resources and time.
- **Update** is the primary mode for incremental data collection, keeping the Data Warehouse aligned with the source, resulting in time and resource savings compared to the first method.

In the presence of Data Marts, the latter are populated either from the centralized DWH or directly from the source. Once the dimension and fact tables of the dimensional model have been updated, indexed, and enriched with appropriate aggregates, the project manager is notified of the publication (Eliasy, 2024).

To reduce the time spent on drawing, parallelizing the ETL process can be done in two ways: by executing more parallel steps and executing a single parallel step. This involves multiplying the ETL flow into more independent tasks, ensuring more attention to each task and better error management. Additionally, the database can identify certain tasks that can be executed in parallel, such as creating an index through all available processors on the machine (Bregata, 2019).

1.2.4 ETL vs ELT

The ELT technique has recently evolved alongside the long-established ETL procedure from the 1970s, featuring identical activities but with the transformation and loading processes executed in reverse order. By transformation, we refer not to modifications immediately related to reporting, but rather to those characteristics of preprocessing, including altering types or formats, addressing inconsistent or erroneous data, and eliminating duplicates. The ELT will thus be delineated by the subsequent three phases:

- Extraction of data from several sources in its original form.
- Loading them into a Data Warehouse or a Data Lake.
- Useful transformations in the target system when they are required. The data cleaning, transformation, and enrichment occur entirely within the Data Warehouse.

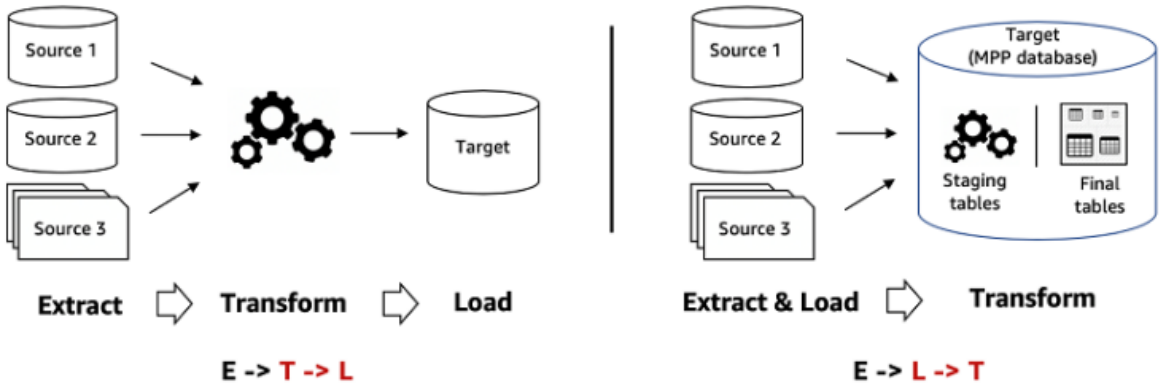


Figure 3: ETL and ELT

ETL is today the most widely used procedure due to the evolution of cloud technologies, allowing companies to archive unlimited non-structured data and analyse them quickly. Big data, such as images, videos, and audio, is also no longer archivable in tabular formats. ETL can transform structured data into another format, while ELT manages several types of data by storing them in the destination Data Warehouse and then converting them into the desired

format. ELT is also faster than ETL because it can use the internal resources of Data Warehouse (Eliasy, 2024).

Aspect	ETL (Extract, Transform, Load)	ELT (Extract, Load, Transform)
Data Type Compatibility	Structured data only	Structured, semi-structured, and unstructured data
Processing Procedure	Extract raw data → Transform into a predetermined format → Load into the destination DWH	Extract raw data → Load into the destination DWH → Transform before analysis
Transformation Location	Transformation occurs on a secondary processing server	Transformation occurs directly within the destination DWH
Speed	Moderate	Faster than ETL
Cost	Configuration costs vary depending on the tools used	More cost-efficient depending on the infrastructure used
Security	May require custom applications to meet data protection requirements	Leverages built-in security features of the destination database

Table 2: ETL vs ELT

Among the most used ETL tools we can find:

- Integrate.io that provides a unified stack for modern data teams. The ETL process transfers data from the source to staging and then to the data warehouse, facilitating complex data transformations and cost efficiency (Smallcombe, 2023).
- IBM DataStage is engineered for complex data transformation and integration processes.
- Talend is an open-source ETL solution featuring an extensive array of connectors and transformation capabilities.

On the other side, these are the main vendors of ELT tools:

- AWS (Amazon Web Services): ELT entails the extraction of raw data, its loading into a data warehouse or data lake, and subsequent transformation within the target system as required (AWS, s.d.).
- SnapLogic relies on cloud storage with data lakes capable of managing substantial volumes of raw, unstructured, and semi-structured data.
- Rivery is an ELT platform known for its scalability and rapidity, especially with extensive datasets and real-time processing.

1.2.5 SQL Server Integration Services (SSIS)

SQL Server Integration Services (SSIS) represents a sophisticated ETL (Extract, Transform, Load) framework developed by Microsoft, offering a robust solution for complex data integration and transformation workflows. As an integral component of Microsoft SQL Server, SSIS facilitates high-performance data ingestion, transformation, and dissemination across heterogeneous systems, ensuring scalability and automation in enterprise environments. Its capabilities extend beyond traditional ETL workflows, incorporating advanced functionalities such as data profiling, quality assessment, and extensive error-handling mechanisms to maintain data integrity throughout the pipeline.

Leveraging a modular architecture, SSIS enables organizations to efficiently orchestrate data pipelines spanning structured and semi-structured sources, including relational databases, flat files, XML, RESTful APIs, and cloud-based repositories.

SSIS operationalizes ETL processes through a series of specialized components that govern the data lifecycle, ensuring efficiency, maintainability, and robustness. These components facilitate process automation, enabling seamless integration between disparate systems while adhering to enterprise governance and compliance standards.

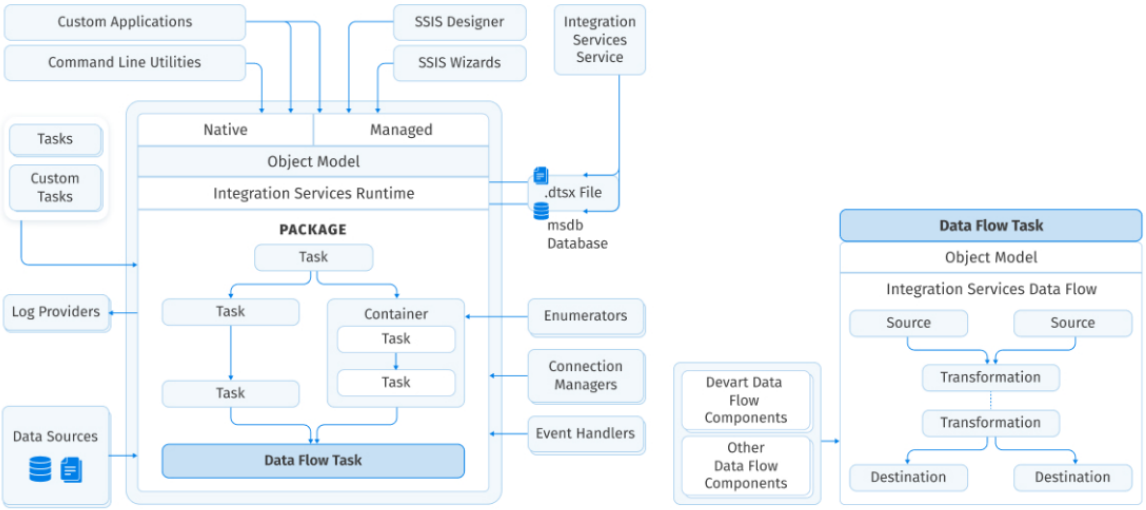


Figure 4: SSIS Architecture

- Data Extraction: the system supports data integration with various systems like SQL Server, Oracle, MySQL, PostgreSQL, Excel, and NoSQL. It uses Change Data Capture methodologies to track incremental modifications and optimizes data synchronization. Connection Managers standardize access to data repositories, ensuring seamless interoperability and dynamic parameterization. Data profiling tools assess source data quality, identifying anomalies before integration.

- Data Transformation: the software uses advanced data cleansing, enrichment, and normalization techniques to standardize diverse datasets. It employs lookup transformations to maintain consistency across normalized schema architectures. It optimizes computational efficiency through parallelized processing strategies and resource management. It supports custom scripting and advanced expressions for complex business rule implementations. It also provides real-time data summarization and calculation of KPIs for analytical insights.

- Data Loading and Optimization: SSIS facilitates structured data ingestion into various data storage systems, ensuring compatibility with BI and analytical workloads. It also provides bulk-insert strategies, partitioning techniques, and data synchronization paradigms. SSIS integrates with Azure Data Factory to extend ETL operations across hybrid and multi-cloud infrastructures and supports data validation checks post-loading for auditing purposes.

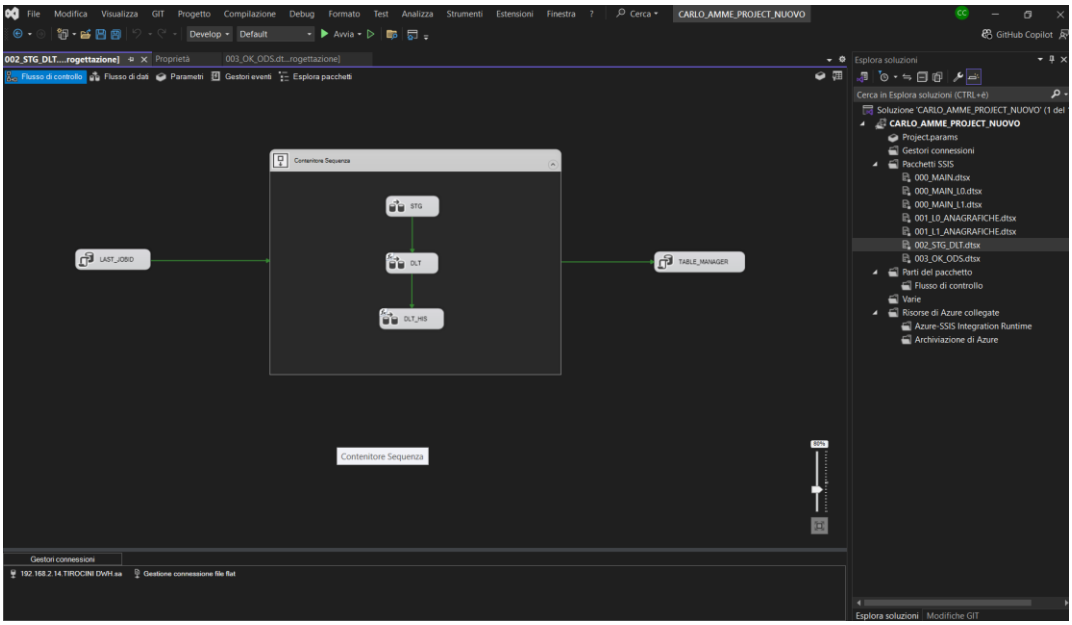


Figure 5: SSIS Interface

Microsoft SQL Server Integration Services (SSIS) provides a declarative workflow design, allowing data engineers to build ETL pipelines using a visual design interface in SQL Server Data Tools (SSDT), which simplifies development while still allowing for script-based customizations (Microsoft, 2024). SSIS incorporates pipeline parallelization, in-memory transformations, and batch processing to optimize execution speed and minimize memory usage, ensuring high-performance data processing (DataCat, 2013).

SSIS also offers operational resilience and fault tolerance through robust error handling, logging, and retry mechanisms, which maintain data consistency and recoverability in critical workloads. The platform is extensible, supporting .NET scripting and third-party libraries for custom data manipulation and enhanced connectivity (Devart, s.d.). ETL workflows can be automated, scheduled, and parameterized via SQL Server Agent and the SSISDB catalog, providing precise control over batch processing and dependency management.

Security and compliance are addressed through role-based access control (RBAC), encryption, and auditing features, ensuring adherence to data protection regulations like GDPR, HIPAA, and SOC 2 (BuzzyBrains, 2024). SSIS scales effectively in big data environments by integrating with Hadoop Distributed File System (HDFS), Spark, and other distributed computing frameworks. It also supports data preparation for AI/ML models, leveraging SQL Server Machine Learning Services for predictive analytics during data transformation (Microsoft, 2024).

To fully utilize SSIS, it is recommended to explore metadata-driven design patterns, performance tuning strategies, and cloud-based orchestration, which allows for leveraging SSIS's full potential in modern data architectures (Rubocki, 2018).With its efficient handling of large-scale data processing, SSIS remains a key technology in enterprise data management, providing the agility needed to adapt to changing business needs and emerging technologies.

1.3 Data Warehouse

Data Warehouses (DWH) serve as the primary instruments facilitating Business Intelligence. They facilitate the acquisition of integrated, consistent, and validated data pertaining to all business processes of a corporation, generated from operational systems. The data are subsequently processed through ETL processes and validated using data quality systems. Data quality is an essential prerequisite for the overall information system. Inaccurate data can impair firm performance and result in suboptimal decision-making, leading to increased expenses and missed opportunities.

The objective of a Data Warehouse (DWH) is to assist knowledge workers—such as executives, managers, administrators, and analysts—in performing studies that facilitate decision-making processes and enhance the organization's information assets. It offers a unified access point to all company data, guaranteeing consistency and dependability via ETL procedures. Moreover, the Data Warehouse ensures extensive historical depth by preserving past information states, facilitating temporal analysis. A primary purpose of the Data Warehouse is the continual lowering of information dissemination costs throughout the software lifecycle. Data Warehouses serve as the fundamental basis for Decision Support Systems (DSS), which are designed to furnish users with clear information, enabling them to examine situations comprehensively and make timely, informed decisions (Vidette Poe, 1998).The DSS depends on data from one or several databases, frequently arranged in various configurations with heterogeneous data. A system must provide regular managerial analysis and control functions, targeted investigation of problem causes, and intricate managerial decision-making. It must also provide user-friendliness for individuals with constrained time availability or hesitance to embrace new technology, particularly when the advantages are not readily evident.

The design of a Data Warehouse may adhere to either a top-down or bottom-up approach:

- **Top-down**. The preliminary stage of a top-down design process is strategizing the Data Warehouse model to incorporate all business data, thereby providing a comprehensive perspective on company data. Consequently, the design entails the establishment of Data Marts tailored to certain business areas. This approach adheres to the backward methodology for ETL and facilitates the representation of many views at a subsequent step.
- **Bottom-up**. The bottom-up method involves planning multiple Data Marts to be consolidated into the primary Data Warehouse. The bottom-up technique facilitates rapid access to targeted business area information (Borzi, 2023).

The establishment of the Data Warehouse facilitates the commencement of a data collecting and availability procedure that incurs minimal expenses over time for each additional user request. The principal aim of creating a transactional database is to enhance the management of a defined set of transactions, whereas the main goal of a Data Warehouse is to improve the accessibility of cross-functional data to meet present and future requirements.

The first step in building a Data Warehouse is to identify the type of data that will be included and subsequently the most appropriate use of that data for decision support purposes. The objective of a data warehousing project is to enable users to manage the company's business more effectively and efficiently. The Data Warehouse should therefore consistently cover the business management needs, both in its constituent parts and in its entirety, rely on business data (and the metadata that supports it) (Filippo La Noce, 2008).

The data contained in the warehouse can be synthesized on multiple levels compared to production data to provide a sequence of images that starts from a high-level business view, characterized by aggregated information, which becomes detailed to explain aspects that appear particularly critical during the analysis. Maintaining both detailed and summarized data allows for handling both unforeseen requests for new aggregations and making a single extraction of data from the operational environment necessary, thus safeguarding performance. The aggregation of operational data allows for the provision of a periodic overview of the business situation, useful for understanding its future trends. This idea is the basis of data historization, a cornerstone concept of data warehousing. This leads to a data architecture organized on multiple levels: at the highest layer, we find information at a very high level of synthesis, enough to justify their allocation in thematic Data Warehouses, the Data Marts, and at the lower

levels, we find information that is detailed as we move down the structure, until it is historicized at the lowest level (Filippo La Noce, 2008).

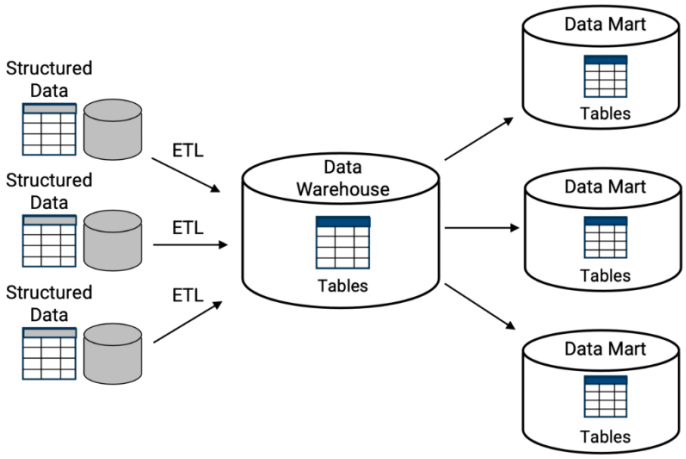


Figure 6: Data Warehouse levels

The data in a Data Warehouse system are reshaped by placing them into new structures that are:

- **Focused on the subject**: in the Data Warehouse, data is structured around pertinent issues, such products, customers, suppliers, and timeframes, to furnish comprehensive information pertaining to a certain domain.
- **Integrated**: The Data Warehouse must interact effortlessly with the many standards employed across different applications. Data must be recoded to provide semantic homogeneity and uniformity in units of measurement.
- **Time-variant**: in contrast to operational data, data within a Data Warehouse possesses an extended temporal scope (5–10 years) and can be repurposed at various intervals.
- **Non-volatile**: operational data are continuously updated, whereas data in a Data Warehouse are initially imported through comprehensive processes and thereafter updated with incremental loads. Once loaded, the data remain unaltered and preserve their integrity throughout time (Bregata, 2019) (Ladley, 2012).

Data warehouses and data marts serve different purposes within an organization's data management strategy. Data warehouses are designed to be application-neutral, centralized, and shared resources, encompassing the entire enterprise. They are characterized by low denormalization, drawing data from many sources, including external operational data, and are

intended for a broad range of users across various areas. Data warehouses are built to be flexible, extensible, and have a long lifespan, focusing on the data itself. The initial implementation of a data warehouse typically takes 9 to 18 months.

In contrast, data marts cater to specific applications, departments, or areas. They have a narrower scope, are specific to a single area, and are highly denormalized. Data marts derive their information from fewer sources, often including external operational data. They are designed to be project-oriented with a shorter lifespan and are less flexible and extensible compared to data warehouses. The implementation timeframe for a data mart is shorter, ranging from 4 to 12 months.

Aspect	Data Warehouses	Data Marts
Purpose	Application-neutral, Centralized and shared	Specific applications, Departments, or areas
Scope	Entire enterprise	Specific departments or areas
Data	Low denormalization	High denormalization
Users	Users from many areas	Users from a single area
Data Sources	Many; External operational data	Few; External operational data
Characteristics	Flexible, extensible, long lifespan, Data-oriented	Narrow, non-extensible, short lifespan, Project-oriented
Implementation Time	9-18 months for the first stage	4-12 months

Table 3: Comparison Data Warehouse and Data Marts

1.3.1 Data Lakes and Data Lakehouses

Data Lakes are modern storage systems that manage large volumes of data in its native format. They offer flexibility in collecting, storing, and analysing heterogeneous data without predefined schemas, making them crucial for analytics, Machine Learning, and real-time processing. Unlike traditional data warehouses, they employ schema-on-read, enabling dynamic data integration and querying (Pwint Phyu Khine, 2018).

Data Lakes evolve through four stages:

- **Data Puddle**: Small, unstructured repositories for specific projects.
- **Data Pond**: Aggregated Data Puddles with minimal structuring for better accessibility.
- **Data Lake**: A distributed, metadata-enriched system using technologies like HDFS or NoSQL, supporting SQL queries and BI tools.
- **Data Ocean**: A large-scale expansion integrating vast data sources, with risks of becoming a Data Swamp if poorly governed.

The Data Lakehouse merges Data Lake scalability with Data Warehouse governance, ensuring structured data management and improved reliability.

Key features of the Data Lakehouse include:

- **Open Data Formats**: Supports Parquet and ORC for interoperability and reduced vendor dependence.
- **Schema Enforcement**: Ensures data quality, compliance, and schema evolution.
- **Diverse Workload Support**: Handles SQL analytics, real-time processing, and Machine Learning with Python and R compatibility.
- **ACID Transactions**: Maintains data integrity and supports concurrent updates.

The Data Lakehouse model optimizes data management, streamlining workflows, enhancing analytics, and ensuring scalability for complex data environments.

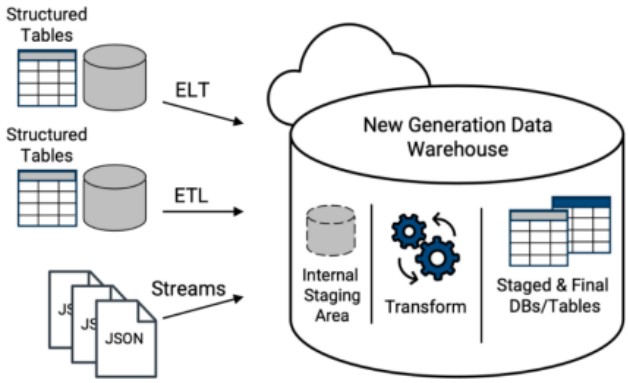


Figure 7: Data Lakehouse

1.3.2 Data Marts

Data Marts, subsets of Data Warehouses, are designed to support specific departments, improving query efficiency and speeding up access to relevant data. A Data Warehouse may be segmented into various Data Marts. In a Data Warehouse context, a Data Mart represents a model that focusses on a certain facet of the overall business architecture. A supermarket Enterprise Data Warehouse may be constructed from smaller Data Marts according to each area of expertise, such as logistics or inventory snapshots (Borzi, 2023). Specifically, a Data Mart is an analytical database designed to meet the specific needs of a business. Being a logical or physical subset of a larger Data Warehouse, it follows the same design rules with data aggregated at various levels of detail, although it can sometimes be created even in the absence of an integrated data system (Rezzani, 2012).

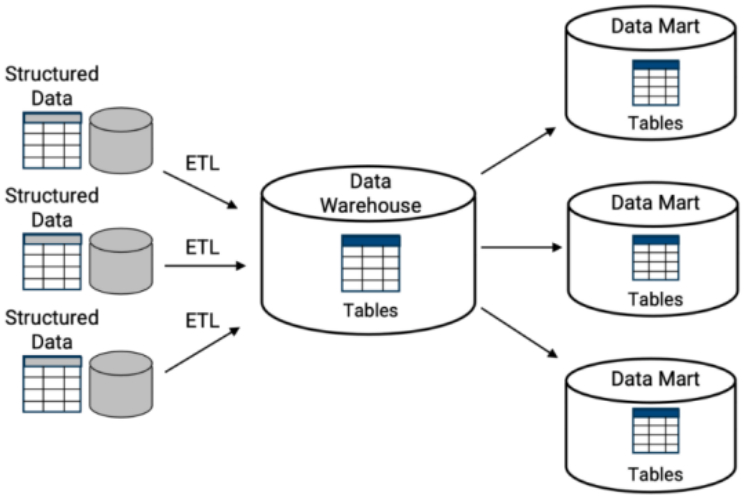


Figure 8: Data Warehouse and Data Marts

1.3.3 Star Schema and Snowflake Schema

The Data Warehouse consists of fact tables which store measurements of a business application, and dimension tables, which describe the dimensions of interest for analysis. Two architectural patterns can be adopted in a Data Warehouse model planning: the star schema and the snowflake schema.

The main difference between Star Schema and Snowflake Schema is represented by the concepts of normalization and denormalization. Let's see in detail:

- **Normalization**: It refers to the process of organizing data in a database to reduce redundancy and data ambiguity. Normalization divides tables into smaller, related tables with the aim of reducing data redundancy and making the database more efficient and easier to manage. A normalized database reduces the need to update data in multiple tables when the data is changed. Normalization is adopted in the Snowflake Schema, which is characterized by an absent redundancy of data. The snowflake schema adopts foreign keys to improve performance efficiency: the primary keys of the dimension tables are used as foreign keys in the fact table. These keys connect the fact table to the dimension tables, establishing the relationships between the numerical measures and the different dimensions. A further step of Normalization is the Third Normal Form (3NF): A relation is in Third Normal Form (3NF) if it satisfies the criteria of Second Normal Form (2NF) and all non-key attributes are directly reliant on the primary key, without any transitive dependencies:
    - Second Normal Form (2NF): A relation must be in 1NF, and all non-key attributes must be functionally dependent on the entire primary key, not just a part of it.
    - Transitive Dependency: This occurs when a non-key attribute is dependent on another non-key attribute, which then relies on the primary key. Third Normal Form (3NF) eliminates this type of dependency. For instance, in an "Employees" database, if the value "DepartmentPhone" is dependent on the attribute "Department," which is thereafter reliant on "EmployeeID," a transitive dependency exists.

In essence, 3NF guarantees that each non-key attribute is directly dependent on the primary key, hence minimising transitive dependencies and minimising data redundancy. This results in a more efficient database that is easier to manage and has a reduced chance of data inconsistency.

- **De-normalization**: Dimension tables can be de-normalized to improve query performance by including some information directly in the fact table. This is particularly useful when you want to avoid performing joins between many tables to obtain complete information. Denormalization is adopted in Star Schemas resulting in a negligible increase in stored data but a significant improvement in query time performance. On the other hand, denormalization cause difficulties in the management and traceability of data changes.

Aspect	Star Schema	Snowflake Schema
Dimension Tables	One table for each dimension containing all the information	Each dimension has a hierarchical structure with many tables
Normalization	Denormalized	Highly normalized
Redundancy	High	Absent
Storage Performance	High due to redundancy	Low due to normalization
Query Performance	High (few join operations)	Low (many joins required)

Table 4: Star schema vs Snowflake schema

1.3.4 Dimensional Fact Model

A conceptual model serves as an early phase that prioritises the most significant aspects of the data, without immediate concern for the database structure. Transitioning immediately from requirements analysis to the logical schema in database design poses challenges; specifically, there is a risk of getting lost in implementation specifics without first establishing a comprehensive model of the data to be stored.

Conceptual design facilitates the representation of real-world data through objects and relationships, concepts that support the progression from an informal business description to a precise and detailed logical model, which is executed by the selected Database Management System (DBMS). The primary distinction between operational systems and data warehousing systems is found in the data model, which refers to the conceptual organisation of data (Eliasy, 2024).

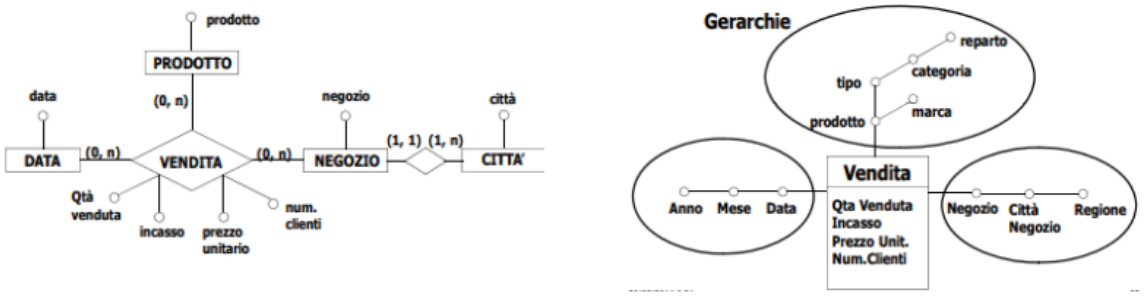


Figure 9: E/R Schema and Dimensional model

Both the star schema and the snowflake consist of:

- **Fact table** (central): contains the numerical measures (facts) that you want to analyse, such as sales or profits. They store measurements of interest in relation to the dimensions of analysis, with each record uniquely identified by a compound primary key and a time reference. Measurements, or metrics, are quantitative descriptions that address specific business information, empowering dimensional analysis on a variable level of granularity.
- **Dimension tables**: these tables collect descriptive information of each dimension of the analysis, identified by a simple primary key and containing textual fields that characterize an aspect of the analysis such as time, product, customer, or geographical location, etc. each row in the dimension table represents a member of that dimension. The Star Schema is characterized by one

table for each dimension that contains all the information. In Snowflake schema each dimension has a hierarchical structure with many tables.

Normalization techniques can be advantageous for operational databases, as they allow generic transactions to update information by touching a single point in the database. The entity-relationship model is the best data model for operational databases, as it allows for quick updates. The entity-relationship model is more suitable for operational systems, while the dimensional model is more commonly used in data warehousing systems. Dimensional models implemented in relational database management systems are called "*star schemas*" due to their "star" structure, while dimensional models implemented in a multidimensional database environment use the term "*OLAP cube*" (OnLine Analytical Processing). However, normalization can be counterproductive for data warehousing databases, as it requires heavy joins between tables, reducing overall performance. The dimensional model is the best choice, regardless of the database management system used (Eliasy, 2024).

1.3.5 OLTP vs OLAP

At the database level, On-Line Transaction Processing (OLTP) relies on fast and efficient multi-access queries. The main operations performed are INSERT, DELETE, and UPDATE as they directly modify the data. The latter are constantly updated and, consequently, require efficient support for rewrite operations. A fundamental feature of these systems is normalization, which provides a quick and effective way to perform writing in the database.

On-Line Analytical Processing (OLAP) is a set of software techniques for the accelerated and interactive analysis of substantial amounts of data, with the possibility of doing so from different perspectives. These systems will prove particularly useful for obtaining summary information, which will aim to support and improve business decision-making processes. Examples of OLAP tools are Data Warehouses and multidimensional cubes.

The main differences between the two systems are listed in the table:

Aspect	OLTP (Online Transaction Processing)	OLAP (Online Analytical Processing)
Purpose	Supports operational activities	Supports decision-making processes
Usage Mode	Process-driven, guided through successive states	Ad hoc queries
Data Volume per Operation	Low: Hundreds of records per transaction	High: Millions of records per query
Quality Focus	Integrity	Consistency
Orientation	Process/Application	Subject
Update Frequency	Continuous, through actions	Sporadic, through explicit functions
Time Coverage	Current data	Historical data
Optimization	For read and write access on a subset of data	For read-only access across the entire database

Table 5:OLTP vs OLAP

Different OLAP designs will exist depending on data storage, each with its own set of advantages and disadvantages (Adamson, 2010):

- **Relational OLAP (ROLAP)**: The data is stored in a relational database to support the OLAP engine. Multidimensional analyses are transformed into queries, which return multidimensional results.
- **Multidimensional OLAP (MOLAP)**: both the database and the multidimensional engine are included. For Drill-Down activities, it is not the best system because it can generate errors.
- **Hybrid OLAP (HOLAP)** combines the benefits of the preceding two technologies. Its pre-aggregates data in multidimensional systems for efficient and rapid analysis, while they are searched in a relational database in the case of drill-down.

- **Desktop OLAP (DOLAP)**: data is fed into a client PC, and the engine processes it locally (Bregata, 2019).

The OLAP Hypercube system was designed to achieve various purposes, such as providing support for conceptual design, creating an intuitive and formal environment for users to make queries, fostering communication between designers and users, building a stable logical design platform, and creating and publishing clear and effective documentation (Bregata, 2019).

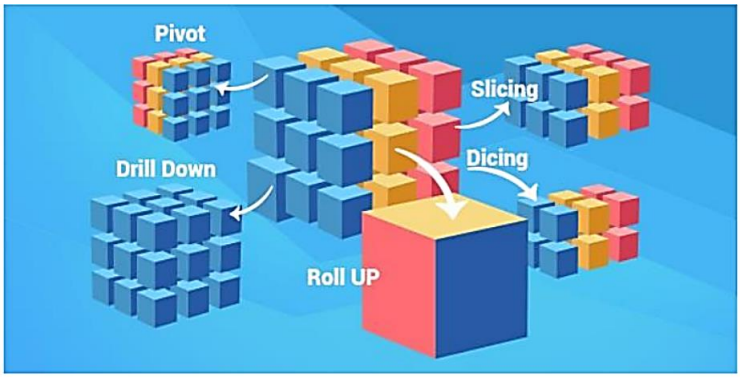


Figure 10: Hypercube OLAP

To navigate within the multidimensional cube, there are operations such as Pivoting, Slice & Dice, and Roll-Up & Drill-Down. Pivoting allows for rapid modification of data visualization by rotating the axes of the cube, while Slice & Dice selects and projects data by filtering on one or more dimensions. Roll-Up & Drill-Down allows users to move within a hierarchy, choosing the level of aggregation according to which they wish to analyse the data (Adamson, 2010).

1.3.6 The Stages of a Data Warehouse Project

The development of a Data Warehouse (DWH) starts with the collection of end-user needs. The project manager must thoroughly comprehend the client's requirements, necessitating a comprehensive analysis of the project's objectives. Initially, the client's needs are often unclear, leading to the introduction of additional or altered requirements throughout development or after project completion. Another scenario that may occur is the demand for all current functionalities to produce a product that, at least in the client's view, seems as innovative and comprehensive as feasible.

This, however, neglects the presence of specific trade-offs, wherein optimal results cannot be achieved concurrently across many dimensions. Moreover, the substantial costs and extended durations necessary for such deployments are frequently neglected. Consequently, it

is essential to assist the customer in delineating criteria and selecting cost-time-quality trade-offs to ascertain the optimal solution for the particular use case.

Upon the explicit and deliberate definition of the needs, the subsequent phase is planning. This phase varies based on the level of creativity of the solution and, accordingly, the project management methodology employed—be it *waterfall*, *hybrid*, or *agile*. Engagement of stakeholders is essential for project success, and ongoing communication with the customer ensures that deliverables and product features correspond with anticipated requirements. The swift advancement of requirements, data types, sources, and technology render the adaptability of hybrid methodologies a crucial element of contemporary undertakings.

Nonetheless, it remains essential to formulate a fundamental plan for the several phases (analysis, design, development, testing, implementation), partially organised and standardised, to enhance efficiency and competitiveness (reduced time and budget). Upon defining the *hard goals* (costs, timetables, quality) and *soft goals* (e.g., client communication techniques, internal communication strategies) of the project, implementation can start. The project entails not only implementation but also maintenance operations (Eliasy, 2024)

Since their beginnings, the models have evolved, necessitating the construction of a DWH in accordance with contemporary ideas (Adamson, 2010). The patterns described in this paragraph serve as a foundational starting point:

- **Inmon Model** - Corporate Information Factory: Data Warehouses are constructed entirely from the beginning as a singular monolithic entity; they should not be considered a composition of Data Marts. A Top-Down methodology is employed.

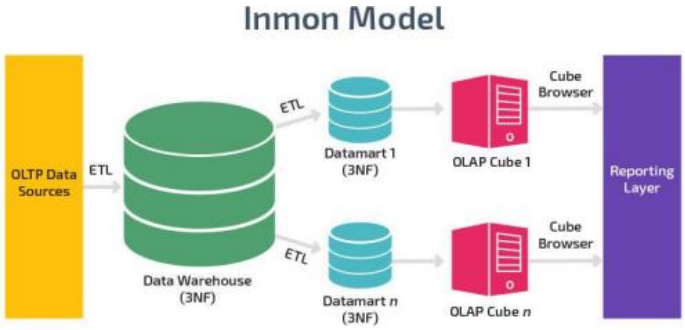


Figure 11: Inmon Model

- **The Kimball Model** - Dimensional Model employs a bottom-up methodology, wherein the Data Warehouse is constructed through the aggregation of multiple Data Marts, each pertaining to a distinct business domain.

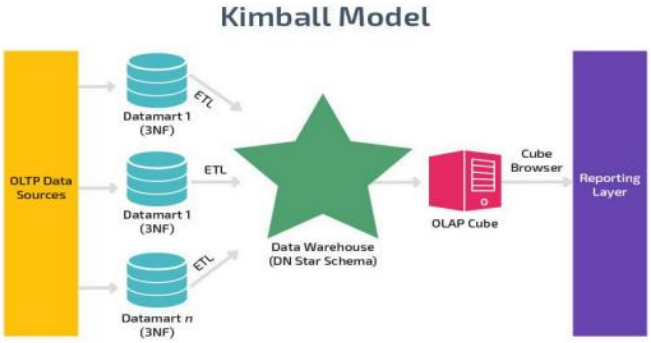


Figure 12: Kimball Model

The Inmon and Kimball methodologies have been shown to effectively implement Data Warehouses. Nonetheless, there exist optimal organisations when it has been employed a hybrid approach, integrating the Inmon model for the Data Warehouse while concurrently establishing business process-oriented Data Marts using the star schema for report production. It is inappropriate to generalise that one strategy is superior to the other; each possesses distinct advantages and limitations, and both are effective in varying contexts. The architect must choose a Data Warehouse strategy based on multiple variables; for any strategy to be effective, it must be thoroughly analysed, extensively discussed, and tailored to fulfil the organization's BI reporting requirements while aligning with the organization's culture (Bregata, 2019).

1.3.7 Cloud Data Warehouse Solutions

Data Warehouses have benefited from the advent of cloud technologies, since the availability of massive computational and storage resources has empowered their availability, scalability, and reliability, with a variety of cost-effective solutions. In contrast to on-premises designs that necessitate comprehensive infrastructure management, cloud platforms offer three service options based on the degree of flexibility and accountability (Borzi, 2023).

- **Infrastructure-as-a-Service (IaaS):** This cloud model offers access to remote, on-demand storage and computational resources, along with fundamental services like virtualization and networking tools. IaaS is considered the most cost-effective architecture, as clients are responsible for managing higher-level components such as the operating system, applications, and middleware (Red Hat, 2023).

- **Platform-as-a-Service (PaaS):** Building on the features of IaaS, PaaS provides a more advanced solution by including the operating system and additional software resources. This eliminates the need to set up and manage the entire infrastructure. Customers focus solely on developing, deploying, and maintaining their software applications (Google Cloud, 2023).
- **Software-as-a-Service (SaaS):** SaaS solutions grant on-demand access to fully managed applications. The service provider oversees the entire underlying infrastructure, enabling users to access software primarily through a web-based subscription model, powered by advanced technologies (Gartner, 2020).

A variety of cloud-based services for Data Warehousing and ETL (Extract, Transform, Load) processes are available, reflecting the three main patterns of cloud architecture. Prominent examples of ETL solutions include:

- Oracle Data Integrator
- Microsoft Azure Data Factory
- Google Cloud Dataflow
- Pentaho Data Integration
- Apache Spark
- Apache Kafka

The first three services are proprietary SaaS (Software as a Service) platforms that require subscriptions for use. In contrast, Pentaho Data Integration Community Edition, Apache Spark, and Apache Kafka are open-source tools. These open-source solutions demand supporting infrastructure in the form of PaaS (Platform as a Service) or IaaS (Infrastructure as a Service). Hybrid models, combining storage and processing capabilities, are frequently used to address both technological limitations and cost concerns associated with complex projects (Borzi, 2023).

2 DATA GOVERNANCE

Organizations are engaged in a fierce struggle for the best use of data in the quickly evolving world of today. The amount, pace, variety, variability, and validity of data collected, stored, and processed by businesses in electronic systems are all increasing quickly due to technology advancements. Among the contemporary data application domains are analytics, process mining, and Artificial Intelligence, which facilitate data-driven decision-making and process innovation for a competitive edge. To maximize the value of data utilization in an efficient, secure, and compliant manner, companies must implement Data Governance, which includes standards, policies, responsibilities, and relationships for managing data. New laws pertaining to data protection (such as the General Data Protection Regulation of the European Union) and secure and moral data processing (such as the European Union's Artificial Intelligence Act) further increase the pressure for compliance and conformity in organizations' management of their data assets (Karol Bliznak, 2024).

2.1 Definition

Organizations can manage data as a strategic asset by implementing Data Governance, which includes the procedures, standards, measurements, and regulations that guarantee the effective and efficient use of information. The swift growth of data volumes and the intricacy of data environments emphasize the importance of Data Governance even more. Every day, businesses from manufacturing and retail to healthcare and financial services produce and process enormous volumes of data. (Chukwurah, 2024)

The Data Management Association (DAMA) postulates Data Governance as *"the exercise of authority, control and shared decision-making (planning, monitoring and enforcement) over the management of data assets."* (Ladley, 2012) This definition places Data Governance as a planning/control overlay over data management. In the meanwhile, Data Governance is defined by the Data Governance Institute (DGI) as *"a system of decision rights and accountabilities for information-related processes, executed according to agreed-upon models which describe who can take what actions with what information, and when, under what circumstances, using what methods."* (Karol Bliznak, 2024).

When it comes to assisting enterprises with new data protection rules (e.g. the General Data Protection Regulation (GDPR) of the European Union) and secure and moral data processing, Data Governance is essential. To reduce risks, stay in compliance with regulations, and gain the trust of stakeholders—including partners, customers, and regulators—effective Data Governance is fundamental (Gade, 2024).

Data Governance is the framework that organizations use to ensure that their data is managed as a valuable asset. This includes developing rules and procedures to ensure data quality, privacy, security, and usability. Effective Data Governance integrates data-related processes with corporate objectives, thereby assisting in compliance, risk mitigation, and decision-making. This is especially important in today's data-driven environment when data accuracy and accessibility have a direct impact on an organization's competitiveness. Data Governance entails responsibility and oversight. It entails determining who owns the data, how it may be accessed, and how it is safeguarded during its lifetime. (Gade, 2024)

Governance rules are based on essential principles such as data integrity, stewardship, access control, and privacy, resulting in a foundation that supports both operational and strategic goals. With well-implemented governance, organizations can not only meet regulatory requirements but also maximize the value derived from their data by ensuring it is high-quality, relevant, and accessible to the right people at the right time.

In the age of AI and big data, advancements in scientific research on Data Governance are widely desired. Organizations may also benefit from a rigorous scientific study of Data Governance tactics and motivators, essential success criteria, and problems associated with actual Data Governance implementation (Karol Bliznak, 2024). The imperative for effective Data Governance is underscored by the inefficiencies observed in current data practices. Industry data suggest that:

- 40% of business professional time is spent fixing and validating data, before they can effectively put it to use.
- 59% of decision makers say it takes months or years to meet new complex requests to turn data into business intelligence.
- 80% of an analyst’s time is spent preparing data, instead of analysing data (Buglio, 2018).

2.2 Frameworks for effective Data Governance

A Data Governance framework is a structured methodology that organizations adopt to effectively manage and safeguard their data assets. At its foundation, a Data Governance framework encompasses key components designed to ensure proper data management throughout its lifecycle. These components include policies, processes, roles and responsibilities, and metrics.

Policies serve as the cornerstone of the framework, establishing the guidelines and standards for data management, access, and protection. They typically address critical areas such as data quality, privacy, and security, setting rules for data handling and ensuring compliance with regulatory and legal requirements.

Processes outline the specific procedures required to implement these policies effectively. These may include data classification, lineage tracking, and incident response protocols, which are essential for maintaining data integrity and security.

Another integral component is the definition of roles and responsibilities, ensuring that all aspects of data management are overseen by designated individuals or teams. Key roles often include the Chief Data Officer (CDO), data stewards, data custodians, and data owners, each with distinct responsibilities. The CDO usually leads the Data Governance initiative, setting its strategic direction and aligning it with organizational objectives. Data stewards are responsible for ensuring the quality and integrity of data within their domains, while data custodians manage the technical infrastructure that houses the data.

The Analytics Maturity Model provides a structured approach to evaluating an organization's data management capabilities across various dimensions, such as Data Governance, data quality, data integration, and data analytics. It offers a roadmap for organizations to improve their data maturity levels, enabling them to make more informed decisions, enhance operational efficiency, and drive business value. The model typically consists of several maturity levels, ranging from ad-hoc or reactive approaches to optimized and proactive data management practices (Kazlow, 2024), that reflect the degree of integration and use of data within the business ecosystem:

- **Descriptive Analytics:** Retrospective analysis that allows understanding of what happened in the past. This level represents the starting point for many companies, focusing on data collection and visualization.

- **Diagnostic Analytics:** This level goes further by determining the causes of past events through the identification of correlations and patterns in the data.
- **Predictive Analytics:** The use of statistical models and Machine Learning techniques to predict future outcomes based on historical trends. This level enhances the organization's ability to anticipate events and decisions.
- **Prescriptive Analytics:** Provides recommendations on optimal actions to take based on predictive insights. At this level, data becomes an active tool for guiding strategic choices.
- **Analytics Culture:** The most advanced level is characterized by the full integration of analytics into business processes, with strong leadership support and a widespread data-driven culture.

Implementing this model requires an accurate assessment of the organization's current maturity level and the development of a roadmap to achieve desired goals. Each step involves investments in infrastructure, training, and advanced tools.

2.3 KPIs

A primary KPI is the reduction of overhead achieved through automation and governance accelerators. However, quantifying the exact benefits of Data Governance remains challenging, making a qualitative assessment often more practical. The most effective way to gauge its impact is by tracking improvements in business initiatives supported by Data Governance, rather than relying on strict mathematical formulas.

Since many business initiatives are closely tied to Data Governance, one method to assess its significance is to consider whether these initiatives could proceed without it. If the absence of Data Governance would lead to the cancellation of key business projects, it highlights its critical role in enabling strategic objectives.

Metrics are used to measure the effectiveness of the Data Governance framework. These can include key performance indicators (KPIs) such as data accuracy rates, compliance rates, and incident response times. By monitoring these metrics, organizations can assess the performance of their Data Governance efforts and identify areas for improvement (Chukwurah, 2024). Here are some key categories and specific metrics to consider:

1. Data Quality Metrics: These metrics evaluate how trustworthy and usable the data is. Key indicators include:

- Accuracy: Percentage of data matching a trusted source.
- Completeness: Percentage of records with all required fields populated.
- Consistency: Rate of contradictions in data across different systems.
- Timeliness: Age of the data relative to its intended update frequency.
- Uniqueness: Percentage of duplicate records detected.
- Validity: Percentage of records that conform to defined formats (Atlan, 2024) (Secureframe, s.d.) (Chu, 2025).

2. Data Security Metrics: these metrics assess the effectiveness of security measures in place to protect data:

- Access Control Effectiveness: number of unauthorized access attempts detected.
- Data Breaches: Frequency and impact of security incidents involving data (Lin, 2024) (Bowman, 2025).

3. Data Usage Metrics: these metrics track how effectively data is being utilized across the organization:

- Adoption Rate: Percentage of business units actively using data assets.
- Data Access Frequency: Number of times data assets are accessed or queried by departments (Secureframe, s.d.) (EWSolution, 2024).

4. Compliance Metrics: these metrics ensure that the organization adheres to relevant regulations and standards:

- Regulatory Compliance Rate: Percentage of compliance with applicable laws and regulations.
- Audit Findings: Number and severity of issues identified during compliance audits (D.Foote, 2024) (Dresse, 2023).

5. Data Governance Training Metrics: these metrics evaluate the effectiveness of training programs related to Data Governance:

- Training Completion Rate: Percentage of employees who have completed Data Governance training.
- Knowledge Assessment Scores: Average scores from assessments conducted post-training (Atlan, 2024) (Chu, 2025).

6. Stewardship Activity Metrics: these metrics measure the effectiveness of data stewardship efforts:

- Stewardship Actions Taken: Number of meaningful actions performed by data stewards, such as validations or corrections over a specified period (EWSolution, 2024) (Secureframe, s.d.).

7. Overall Program Effectiveness: this includes broader metrics that assess the impact of the governance program on business outcomes:

- Business Value Realization: Measurement of how Data Governance initiatives contribute to business objectives, such as increased revenue or improved operational efficiency (Bowman, 2025) (D.Foote, 2024).

By implementing these metrics, organizations can gain valuable insights into their Data Governance frameworks, enabling them to identify areas for improvement, demonstrate value to stakeholders, and ensure compliance with regulations. Regular monitoring and assessment against these KPIs will help maintain a robust governance program aligned with organizational goals.

Category	KPI	Description
Data Quality Metrics	Accuracy	Percentage of data matching a trusted source.
	Completeness	Percentage of records with all required fields populated.
	Consistency	Rate of contradictions in data across different systems.
	Timeliness	Age of the data relative to its intended update frequency.
	Uniqueness	Percentage of duplicate records detected.
	Validity	Percentage of records that conform to defined formats.
Data Security Metrics	Access Control Effectiveness	Number of unauthorized access attempts detected.
	Data Breaches	Frequency and impact of security incidents involving data.
Data Usage Metrics	Adoption Rate	Percentage of business units actively using data assets.
	Data Access Frequency	Number of times data assets are accessed or queried by departments.
Compliance Metrics	Regulatory Compliance Rate	Percentage of compliance with applicable laws and regulations.
	Audit Findings	Number and severity of issues identified during compliance audits.

Data Governance Training Metrics	Training Completion Rate	Percentage of employees who have completed Data Governance training.
	Knowledge Assessment Scores	Average scores from assessments conducted post-training.
Stewardship Activity Metrics	Stewardship Actions Taken	Number of meaningful actions performed by data stewards, such as validations or corrections over a specified period.
Overall Program Effectiveness	Business Value Realization	Measurement of how Data Governance initiatives contribute to business objectives, such as increased revenue or improved operational efficiency.

Table 6: Metrics in Data Governance

2.4 Industry Standards and Models for Data Governance Framework

Industry standards and models offer essential direction for the development and sustenance of successful Data Governance frameworks. A prominent standard is the Data Management Body of Knowledge (DAMA-DMBOK), established by the Data Management Association (DAMA). DAMA-DMBOK provides an extensive overview of data management methodologies, encompassing Data Governance, and delineates optimal methods for treating data as a valuable asset. It offers a systematic framework for Data Governance, covering essential domains such as data architecture, quality management, and security.

Another significant framework is the Control Objectives for Information and Related Technologies (COBIT), created by ISACA. COBIT emphasises the alignment of IT objectives with business goals, offering a framework for the administration and management of enterprise IT. COBIT's principles and practices are particularly relevant to Data Governance, highlighting the significance of accountability, risk management, and performance evaluation. It offers a systematic framework for Data Governance that guarantees conformity with organisational objectives and regulatory standards.

The ISO/IEC 38500 standard, created by the International Organisation for Standardisation (ISO) and the International Electrotechnical Commission (IEC), establishes a framework for the governance of information technology within corporations. Although not solely dedicated to Data Governance, ISO/IEC 38500 provides principles and recommendations that are significantly pertinent to the field. It underscores the significance of responsibility, transparency, and compliance, offering a comprehensive framework for organisations to tailor to their requirements (Chukwurah, 2024).

2.4.1 Data Management Body of Knowledge (DAMA-DMBOK)

The DAMA-DMBOK, developed by the Data Management Association (DAMA), is a comprehensive framework that outlines best practices for data management, including Data Governance. It emphasizes the importance of treating data as an asset and provides a structured approach across ten core knowledge areas:

- **Data Governance**: Establishes policies, procedures, and standards for effective data management.
- **Data Architecture**: Focuses on designing and maintaining the data infrastructure.
- **Data Quality Management**: Ensures data accuracy, completeness, consistency, timeliness, validity, and uniqueness.
- **Data Security**: Protects data from unauthorized access and ensures compliance with regulations.
- **Data Integration and Interoperability**: Ensures seamless data exchange across systems.

DAMA-DMBOK serves as a valuable resource for organizations to create a robust Data Governance framework by providing guidelines that prevent data silos and facilitate integration across various teams and verticals (Atlan, 2024) (OptimizeMRO, 2024).

2.4.2 Control Objectives for Information and Related Technologies (COBIT)

COBIT, developed by ISACA, focuses on aligning IT goals with business objectives. It provides a framework for governance and management of enterprise IT, emphasizing accountability, risk management, and performance measurement. COBIT's principles are highly applicable to Data Governance, ensuring that data management practices align with organizational goals while mitigating risks associated with data handling. COBIT outlines several key components:

- **Governance Framework**: Establishes clear roles and responsibilities within the organization.

- **Performance Measurement**: Provides metrics to assess the effectiveness of governance initiatives.
- **Risk Management**: Identifies potential risks related to Data Governance and establishes controls to mitigate them.

By applying COBIT principles, organizations can enhance their Data Governance frameworks to ensure they meet both regulatory requirements and strategic business objectives (Mosley, 2008).

2.4.3 ISO/IEC 38500

The ISO/IEC 38500 standard offers a framework for the corporate governance of information technology. While it is not exclusively focused on Data Governance, it provides principles that are highly relevant to the field. The standard emphasizes accountability, transparency, compliance, and ethical behaviour in managing IT resources. Key principles include:

- **Responsibility**: Ensuring that decision-makers are accountable for their actions regarding data management.
- **Strategy Alignment**: Aligning IT governance with organizational strategies to ensure that data supports business goals.
- **Risk Optimization**: Managing risks associated with information technology in a manner that aligns with the organization's risk appetite.

Organizations can adapt these principles to enhance their Data Governance efforts by promoting accountability and transparency in their data management practices (Spiller, 2021).

Incorporating these industry standards into a Data Governance framework allows organizations to establish best practices that enhance data quality, security, and compliance. By leveraging the structured approaches offered by DAMA-DMBOK, COBIT, and ISO/IEC 38500, organizations can create a cohesive strategy for managing data effectively as a valuable organizational asset. This not only facilitates better decision-making but also supports long-term business success in an increasingly data-driven environment.

2.5 Cloud Data Governance

As enterprises increasingly migrate to cloud environments, effective Data Governance has emerged as a crucial element for ensuring data protection, compliance, and efficient management. Cloud Data Governance encompasses the creation and implementation of policies, processes, and standards that regulate the responsible management of data assets across diverse cloud platforms. This requires continuous investment in governance structures that can adapt to the changing legal and regulatory environments related to cross-border data flows and jurisdictional complications. By emphasising a systematic methodology for Data Governance in the cloud, organisations may enhance the security of their data assets while promoting innovation and growth. The transition to cloud infrastructures, frequently defined by multi-cloud or hybrid models, has substantial advantages, including scalability and cost efficiency (Gade, 2024).

This flexibility increases the risks associated with data breaches and compliance infractions. Organisations must establish a complete governance framework that handles essential problems such as data ownership, security, compliance, and operational transparency. In contrast with traditional governance approaches that emphasised the regulation of on-premises systems, cloud governance necessitates flexible techniques to proficiently manage the complexities of a distributed data environment. It is feasible to examine the core challenges and prospects presented by cloud Data Governance, while also offering insights into best practices that companies may adopt to safeguard their data and enhance their operational resilience. This intricate yet essential process will assist organisations in manoeuvring a more data-centric landscape, enabling them to fully leverage their data while ensuring compliance, security, and future readiness.

2.5.1 Challenges

Traditional governance approaches frequently fail when applied to cloud environments because of the dynamic nature of cloud resources and the complexity of multi-cloud and hybrid configurations. Let us look at the most major obstacles that organizations encounter while managing data in cloud settings, including security concerns, regulatory demands, ownership ambiguities, and integration issues:

- **Data Visibility and Control**: the cloud distributes data across multiple platforms and regions, obscuring visibility and control. Organizations must

implement tools and procedures that enable them to monitor and manage data from numerous places while adhering to governance rules. Organizations frequently lack complete visibility into where their data lives, which poses a unique problem for cloud compliance. This makes it difficult to ensure compliance with data residency regulations or provide an audit trail when authorities want information. Cloud providers do provide compliance support, but it is the client's obligation to configure and maintain compliance settings. For enterprises, this entails a continuous commitment to monitoring and responding to regulatory changes, which can be both time-consuming and complex, particularly in multi-cloud environments where each provider interprets compliance differently (Gade, 2024) (Ratman, 2024) (Islam, 2024).

- **Balancing Security and Accessibility**: governance systems for cloud environments must achieve a balance between security and accessibility. Security measures, such as encryption and multi-factor authentication, are necessary but should not obstruct authorized access. Governance policies must be adaptable enough to accommodate authorized users while keeping threats at bay.

  Data security is one of the most critical issues in cloud computing. Organizations face a variety of dangers when data is stored offsite and accessed via the internet, including data breaches and unauthorized third-party access. While cloud providers frequently provide comprehensive security frameworks, enterprises are still responsible for securing sensitive data and controlling risks specific to their settings.

  Cloud data is especially vulnerable to attack due to the interconnected structure of cloud services, which can disclose security flaws. When apps, networks, and databases are integrated, a breach in one area can swiftly spread to others. In this setting, organizations must apply a layered approach to security by encrypting data both at rest and in transit, implementing strong identity and access management (IAM), and continuously monitoring for threats. However, the diverse security tools and services offered by various cloud providers can create a fragmented approach, making it challenging to establish cohesive security policies (Gade, 2024) (Ratman, 2024) (Ofori, 2024).

- **Compliance**: Compliance is another challenge, particularly as regulatory frameworks become stricter and more diverse across industries and geographies, necessitating ongoing adaptation (Gade, 2024). Controlling data that may be spread across numerous jurisdictions, each with its own set of privacy laws and regulatory norms, is a significant difficulty in cloud Data Governance. This dispersion of data assets causes regulatory issues, particularly in highly regulated areas like finance and healthcare. Compliance with such as the General Data Protection Regulation (GDPR) or the California Consumer Privacy Act (CCPA) necessitates a thorough understanding of data storage locations, usage patterns, and access rights. Noncompliance has significant effects, including costly penalties, reputational harm, and potential legal ramifications, thus regulatory adherence is a critical component of any cloud Data Governance strategy (Gade, 2024).
- **Vendor Management**: many cloud services rely on third-party vendors, which requires you to trust them with sensitive data. Governance frameworks must contain policies for assessing vendor adherence to data management requirements. This involves performing frequent audits and ensuring that providers meet their contractual duties for data security and privacy (Boyd M Knosp, 2022) (Ofori, 2024).
- **Scalability of Governance Policies**: cloud usage frequently increases rapidly, necessitating governance mechanisms that can expand proportionately. To deal with rising data volumes and complexity, flexible frameworks must be developed and automated for tasks such as data classification and access management (Koppichetti, 2023).
- **Complexity in Data Ownership and Responsibility**: in traditional on-premises environments, data ownership is straightforward. However, in cloud environments, data ownership and responsibility can become ambiguous, especially as cloud providers take a “shared responsibility” approach. While the cloud provider is typically responsible for securing the infrastructure, the company using the cloud is responsible for managing its data and controlling access to it. This shared responsibility model requires organizations to carefully

navigate their roles and establish clear governance policies to avoid misunderstandings.

Cloud providers offer tools and configurations to secure data, but it is up to organizations to implement and monitor them. Inadequate understanding of these responsibilities can lead to vulnerabilities, as businesses may unknowingly neglect areas they assume are covered by the provider. To ensure clarity, organizations must develop governance frameworks that clearly define roles, responsibilities, and accountability measures in their agreements with cloud providers (Gade, 2024).

- **Data Access Management in Multi-Tenant Environments**: multi-tenancy, in which numerous clients share the same cloud infrastructure, is a critical component of cloud computing that provides significant cost savings. However, it presents distinct data access issues. In a multi-tenant environment, each company must keep its data segregated and inaccessible to other tenants while yet ensuring secure access for its employees and third-party partners. Cloud providers often provide multi-layered access controls, but it is up to each business to customize these restrictions based on its governance requirements. In highly regulated industries such as finance and healthcare, where data access is strictly controlled, managing access in multi-tenant setups becomes even more difficult. Granular access control policies and constant monitoring of access records are critical to preventing unwanted access, but these techniques can become laborious. Access controls may become inappropriate over time, necessitating a regular assessment to ensure they meet both the needs of the company and regulatory obligations (Boyd M Knosp, 2022) (Ratman, 2024).
- **Data Integration and Consistency Issues**: data may be stored in different formats & databases, often scattered across various regions and services. This lack of uniformity can lead to data silos, making it difficult to achieve a sole source of truth. For instance, an organization may have customer data stored in one cloud platform and transactional data in another. Ensuring that these datasets are synchronized & consistent requires sophisticated integration and data replication strategies, often involving tools for Data Lakes, Lakehouses, or data mesh architectures (Onyinye Obioha Val, 2024).

2.5.2 Opportunities

The adoption of cloud technology offers notable advantages, including increased scalability, flexibility, cost efficiency, and accessibility (Ocean, s.d.). Cloud solutions introduce advanced tools and automation capabilities that elevate governance standards, enabling scalable governance initiatives through automated data cataloguing, real-time monitoring, and AI-driven compliance checks. Unlike traditional on-premises models, cloud-native solutions enhance data lineage, quality management, and access control. By leveraging these cloud-specific governance tools and processes, companies can achieve improved visibility and control over their data, aligning data management policies with business objectives while mitigating risks (Gade, 2024):

- **Advanced Tools for Automation and Monitoring**: Cloud platforms' scalability allows enterprises to adopt governance solutions on a much greater scale than before, leveraging advanced technologies such as Artificial Intelligence (AI) and Machine Learning (ML) to automate policy enforcement and data classification (Gade, 2024). This automation also helps to eliminate human mistake, which remains a major risk element in data management (Ladley, 2012). The cloud automates governance processes, making it easier to implement policies uniformly. Tools for monitoring data activity, detecting abnormalities, and enforcing access regulations are readily available, enabling for more efficient governance processes which grow with corporate expansion (Ticong, 2024).

  Machine Learning may also improve data quality and integrity by identifying abnormalities or patterns that may suggest data corruption or inaccuracies. Cloud environments offer innovative tools for automation, Machine Learning, and Artificial Intelligence, which can be used to improve Data Governance projects. Automation also decreases the manual workload for Data Governance teams, from automatically identifying sensitive material to enforcing compliance laws based on predefined rules. With these features, cloud-based governance allows for proactive governance actions, such as automatically detecting or resolving possible concerns without the need for manual intervention. This proactive approach can reduce risks before they become concerns, giving enterprises more control and security over their data ecosystems (Ticong, 2024) (Cavanell, 2023) (Maffeo, 2023).

- **Enhanced Collaboration & Data Sharing**: Data Governance in the cloud also facilitates increased collaboration, allowing teams from different countries and time zones to securely and efficiently access and interact with shared data across departments and locations, enabling data exchange while still retaining governance restrictions (Gade, 2024). This can foster creativity while also allowing cross-functional teams to securely access and use shared data assets (Paypro Global, 2024).
- **Cost Savings**: governance policies can assist firms in implementing cost-effective storage strategies, such as archiving inactive data or using tiered storage systems (Petricca, 2024). This enables businesses to manage budgets while keeping access to critical data. Cost efficiency is a central appeal of cloud technology, and Data Governance is no exception. Traditional on-premises governance solutions sometimes incur substantial expenses due to hardware, maintenance, and storage requirements. Cloud-based governance, on the other hand, often functions on a pay-as-you-go basis, allowing organizations to pay for only the resources they require at any given time (Gade, 2024). Cloud-based Data Governance improves resource efficiency by allowing enterprises to benefit from pooled services, decreased downtime, and quicker updates (Ladley, 2012).
- **Improved Data Analytics**: cloud Data Governance frameworks that emphasize quality and accessibility help enhance data analytics capabilities (Karol Bliznak, 2024). Organizations that have a clear structure for data quality, integrity, and access can obtain greater insights from their data, allowing for more effective decision making (Osservatorio Cloud Transformation, 2024).
- **Enhanced scalability and flexibility**: one of the most significant benefits of cloud-based Data Governance is scalability. The cloud alleviates much of this load by enabling firms to swiftly and easily adapt governance frameworks in response to data expansion, business changes, or regulatory needs (Bhuvana Jayabalan, 2024). This scalability is especially useful in businesses where data demands fluctuate or peak periodically, such as retail during the holidays or finance during tax filing season. Cloud-based governance frameworks adapt to

these developments, allowing firms to manage growing workloads while maintaining day-to-day operations (Bhuvana Jayabalan, 2024) (Maffeo, 2023).

- **Improving Data Visibility and Transparency**: Data visibility and transparency improve when cloud platforms offer centralized data catalogues and governance dashboards that support a uniform view of data assets (Gade, 2024). Organizations can use these features to follow data across departments and systems, providing insights into who is utilizing what data and how it is being used. This transparency allows firms to better manage data access and decreases the risk of data silos, which occur when essential information is separated within specific teams or systems, making it difficult to administer. Improved data visibility fosters confidence both within the organization and among external stakeholders. When there is a clear, accessible record of where data lives, how it's managed, and who has access, everyone from data users to compliance officers can build confidence in the organization's data management practices (Ladley, 2012). This granular approach strikes a balance between data accessibility and security, allowing teams to make data-driven decisions while maintaining data integrity and compliance (Mazzi, 2021) (Perel, 2023).

2.5.3 Why Data Governance is Crucial in the Cloud?

As businesses transition to the cloud, they face additional issues in ensuring data is maintained in accordance with the same governance rules as on-premises data. Here are a few reasons why Data Governance is especially important in cloud environments:

- **Compliance with Regulations**: many industries are subject to severe regulations governing data privacy and security, such as GDPR in Europe and HIPAA in the healthcare sector. Moving data to the cloud frequently includes data crossing geographical and legal boundaries, which complicates compliance. Strong governance standards help to ensure that data handling in the cloud meets regulatory requirements, lowering the chance of costly infractions (Mathew, 2024).

- **Data Security & Privacy**: cloud systems have numerous advantages, but they also expose data to new forms of security threats. Data stored on the cloud is accessible from anywhere, which raises the danger of illegal access or data breaches. Proper Data Governance ensures that security mechanisms are in place

to protect sensitive information, both at rest and in transit. Governance also specifies privacy rules to secure customer and company data while promoting trust and compliance (Evren Eryurek, 2019).

- **Quality & Integrity of Data**: when data is distributed across different cloud services and locations, ensuring consistency and accuracy becomes difficult. Cloud environments can raise the likelihood of data duplication, inconsistency, and corruption. Governance frameworks include systems for standardizing data formats, creating data quality checks, and establishing procedures for periodical audits, guaranteeing that data is correct and useable (Lashch, 2023).
- **Data Accessibility & Usability**: while the cloud makes data available globally, consumers may find it difficult to identify, access, or analyse the data they require without defined governance regulations. Governance procedures aid by standardizing data organization and access restrictions, as well as speeding data retrieval workflows. When employees have regular access to quality data, they can make better educated decisions, which boosts productivity (Evren Eryurek, 2019).
- **Cost Management**: storing and managing data in the cloud can become costly if not governed effectively. Unnecessary data duplication, prolonged data storage, and inefficient access practices can drive up expenses. With effective governance, organizations can track data usage, archive outdated data, and enforce cost-efficient storage practices, making cloud data management both practical and affordable (Vincent, 2024).

2.5.4 Best Practices for Implementing Data Governance in the Cloud

Cloud adoption has changed the data landscape for enterprises. While the cloud provides flexibility and scalability, it necessitates a rigorous approach to Data Governance. Navigating Data Governance in the cloud presents new difficulties and opportunities for enterprises that manage sensitive or regulated data. Companies can maintain control, ensure compliance, and maximize the value of their cloud-based data assets by applying best practices.

- **Defining Clear Data Ownership and Responsibility**: one of the most fundamental aspects of Data Governance in the cloud is to establish clear ownership and accountability for data assets. Without specific roles, data can

easily fall through the cracks, leading to compliance risks, inefficiencies, or even data breaches. Data in the cloud is often spread across various regions, teams, and even third-party platforms. To avoid ambiguity, every data set should have a designated owner who is responsible for its quality, security, and compliance.

Data ownership clarifies who has the authority to make decisions about access levels, data lifecycle, and retention policies. In addition, establishing ownership in practice define data ownership by linking data sets with roles rather than individuals. For example, a "Data Steward" role can be assigned to each department, responsible for ensuring the integrity and security of departmental data (Jolas, 2023).

- **Establishing strong Data Governance policies**

Establishing strong Data Governance policies is crucial for maintaining regulatory compliance and meeting organizational needs. These policies must be tailored to various regulatory frameworks, such as GDPR, HIPAA, and other sector-specific rules, ensuring the integrity and efficiency of data management operations. Organizations must develop rules that specifically handle data access, sharing, and retention, making them accessible and understandable to all stakeholders.

A strong control system is necessary for successful enforcement of compliance policies, ensuring governance is not only theoretical but also actively practiced throughout the business. Access control systems protect sensitive data by limiting access to authorized personnel, while data encryption ensures data integrity. Regular audits are essential for monitoring compliance, finding weaknesses, and ensuring adherence to set standards. Automation improves compliance monitoring by increasing efficiency, consistency, and dependability in governance procedures, making it easier to track access logs, identify unwanted access attempts, and manage data retention dates.

Cloud providers offer solutions designed to improve data security, compliance, and governance, saving time and resources while strengthening the Data Governance system. Identity and Access Management (IAM) services allow enterprises to define and regulate access to resources, while role-based access control (RBAC) restricts access to sensitive data to personnel with the

necessary authorization. Modern cloud platforms provide extensive encryption at multiple levels, safeguarding data from illegal access and ensuring adherence to regulatory requirements.

Data masking techniques provide further security by obscuring critical information in non-production situations, ensuring a suitable level of security for various types of information. Real-time monitoring and alert systems supplied by cloud platforms, such as Amazon CloudTrail, Google Cloud's Operations Suite, and Azure Monitor, enable enterprises to monitor access logs, discover abnormalities, and receive alerts for suspicious activity. This comprehensive strategy ensures data protection while remaining compliant and operationally efficient in dynamic cloud environments (Vincent, 2024).

- **Creating a data-centric culture**

  Across an organization is essential for the success of Data Governance, as it extends beyond being a purely technical endeavour and requires a significant cultural shift. Employees must understand the value of Data Governance and actively participate in maintaining it. This begins with investing in data literacy initiatives that educate employees on data privacy laws, security protocols, and their roles in protecting and managing data effectively. Regular training sessions, accessible resources, and clear communication of Data Governance policies are vital to fostering this understanding.

  To ensure that governance is not viewed as an additional or separate task, it should be seamlessly integrated into everyday workflows. For example, embedding data quality checks within operational processes and incorporating compliance reminders into software tools can make Data Governance a natural part of daily work. Encouraging accountability across all levels of the organization is equally important. Employees should feel empowered to flag data quality issues, report unusual access, and propose improvements, ensuring that Data Governance becomes a shared responsibility and an integral aspect of the organizational culture.

Data Governance in the cloud is both a requirement and a strategic advantage, but it presents new issues and opportunities. As more businesses migrate to cloud environments, the complexity of managing data across distant systems, various geographies, and even hybrid

infrastructures become clearer. However, when organizations handle Data Governance in the cloud strategically and with an eye toward agility, the benefits far outweigh the obstacles. Maintaining uniform security and privacy requirements is one of the most challenging aspects of cloud-based Data Governance.

2.6 Data-Driven Organizations

The concept of a data-driven organization is increasingly vital in today's business landscape, where data serves as a key asset for strategic decision-making and operational efficiency. A data-driven organization is one that prioritizes data in its decision-making processes. This approach involves using data analytics to inform business strategies, optimize operations, and enhance customer experiences. The transition to a data-driven model necessitates a fundamental shift in how organizations perceive and utilize data.

- Anticipating Market Trends: Data-driven organizations leverage predictive analytics to forecast market trends and consumer behaviour. By analysing historical data, businesses can make informed predictions about future demands, enabling them to adjust their strategies proactively.
- Optimizing Operational Efficiency: Organizations that adopt a data-driven approach can streamline their operations by relying on objective metrics rather than intuition or past experiences. This leads to improved resource allocation and increased productivity.
- Personalizing Offerings: Through the intelligent use of customer data, companies can tailor their products and services to meet specific customer needs. This personalization enhances customer satisfaction and loyalty, driving sales and revenue growth.
- Promoting Innovation: Integrating data into product development processes encourages innovation. By analysing user feedback and market trends, organizations can develop new products that better meet customer expectations.

Creating a robust data culture is essential for the successful implementation of a data-driven strategy. A data culture refers to an organizational mindset where decisions are primarily based on data rather than intuition or anecdotal evidence. It is critical for the successful execution of a data-driven strategy because it promotes data-driven decision-making over intuition or

anecdotal evidence. Leadership commitment (Siatec, 2024), ongoing training and education for employees at all levels (MJV, 2023), encouraging experimentation and learning from errors (Siatec, 2024), and data accessibility are all critical components of cultivating a data culture (Talarico, 2024). Top management should actively support the transition to a data-driven culture, addressing resistance from middle management and employees. Continuous education instils confidence in using analytical tools and evaluating data. An effective data culture encourages employees to try new ideas based on data analysis, creating an atmosphere in which innovative solutions might develop. Ensuring data accessibility through user-friendly dashboards and reporting tools is critical for developing a data-driven culture. Adopting a data-driven approach can be challenging for organizations due to cultural resistance (Girardo, 2024), data quality issues (Regesta Lab, 2024), and system integration (Digital4, 2021). Traditional decision-making processes often face resistance, while high-quality, accurate, and reliable data is crucial for effective decision-making. Poor quality data can lead to misguided decisions and erode trust in analytics. A cohesive strategy for organizing and accessing valuable data is essential for effective decision-making in these organizations.

Businesses should take the following important measures to successfully become a data-driven organisation:

- Establish Clear Goals: Define what being data-driven means for the organization and set measurable objectives that align with overall business goals (Siatec, 2024).
- Develop Governance Frameworks: Put in place strong governance frameworks that define the roles, duties, and procedures for efficiently managing and using data.
- Invest in Technology: Use advanced analytics technologies, such as Machine Learning platforms and business intelligence software, to gain a better understanding of the company's activities (Gandini, 2024).
- Monitor Progress: Regularly evaluate the efficacy of the organization's data projects by using performance metrics and feedback loops to verify that improvements are ongoing (Regesta Lab, 2024).

- Promote a Culture of Continuous Learning: Support continuing training and development opportunities relating to data literacy throughout the organisation (Girardo, 2024).

In conclusion, becoming a truly data-driven organization requires more than just implementing new technologies; it necessitates a cultural transformation that permeates all levels of the business. By prioritizing data in decision-making processes, fostering a supportive environment for experimentation, and ensuring high-quality governance practices are in place, organizations can unlock the full potential of their data assets.

2.7 Market Landscape: Gartner Magic Quadrant

The Gartner Magic Quadrant is a pivotal analytical framework for assessing Data Governance solution providers, segmenting them into Leaders, Challengers, Visionaries, and Niche Players based on their execution capabilities and strategic vision.

- Leaders demonstrate superior scalability and integration, delivering comprehensive, AI-driven platforms that enhance governance, compliance, and data quality. Informatica, Collibra, and Talend are market dominators, leveraging cutting-edge technologies to optimize data management.
- Challengers, including IBM and SAP, exhibit robust execution but possess a relatively constrained vision, focusing on enterprise-scale deployments and multi-cloud ecosystems to ensure operational resilience.
- Visionaries, such as Ataccama and Alation, are at the forefront of innovation, integrating AI and automation to advance Data Governance paradigms, though their market penetration remains emergent.
- Niche Players, exemplified by Egnyte and OvalEdge, provide industry-specific, cost-effective solutions that prioritize functionality, compliance, and affordability within specialized sectors.

This comprehensive analytical approach empowers organizations to strategically select vendors by aligning technological innovation, scalability, and financial feasibility with their long-term Data Governance objectives.

Figure 13: Gartner Magic Quadrant

3 COMPANY DATA INTEGRATION FRAMEWORK

Mediamente Consulting created a Data Integration framework for ETL and Data Warehousing, allowing for a standardised and readily maintainable data flow using metadata tables. The firm structure consists of three layers, including an additional management layer. The Staging Area stores raw data from sources and feeds it into subsequent ETL operations. Relational Data Storage is a centralised store for reconciled and normalised data. The Dimensional Data Storage layer integrates data from several sources and prepares it for loading into the Data Warehouse or other dimensional analysis storage systems. The management layer holds metadata and process information (Chiarello, 2020).

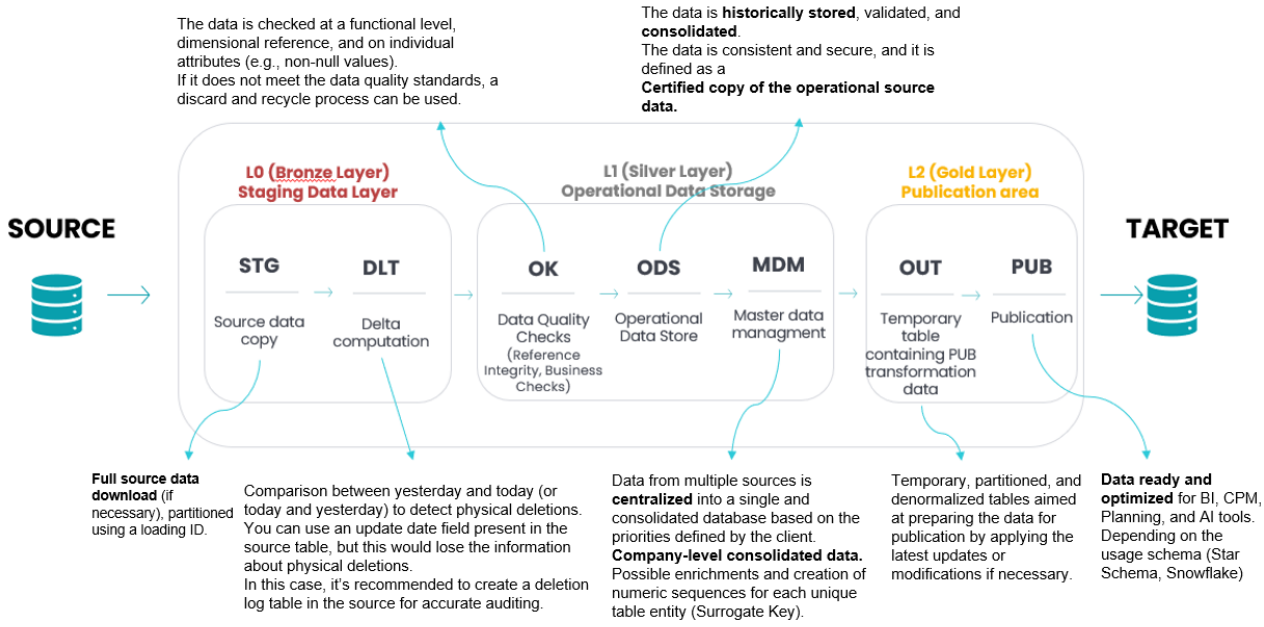


Figure 14: The company framework

3.1 Metadata Layer

Metadata, a term that means "data about data," are information that describe the structure, content, and meaning of the data stored in a DWH. They serve as a dictionary or instruction manual, providing context and understanding to users. Having well-defined metadata is essential for several reasons:

- **Usability**: They allow users to understand the data and find what they need for analysis. Imagining a DWH with cryptic table and column names; metadata help decipher their meaning.
- **Data quality**: Metadata can help identify and resolve data quality issues such as inconsistencies or missing values.
- **Maintainability**: When changes are made to the structure of the DWH or the definitions of the data, clear metadata facilitate system updates and ensure ongoing accuracy.
- **Governance**: Metadata help document data ownership, access rights, and usage policies, promoting Data Governance within the organization (Eliasy, 2024).

Metadata is data that describes other data and plays a significant role in data warehousing. It indicates the sources, value, use, and functions of the data stored in the Data Warehouse and describes how the data is altered and transformed as it passes through distinct levels of the architecture. Metadata aims to uniquely and accurately characterize information regarding the status of a process, preventing the initiation of multiple instances, ensuring processes are launched at appropriate times, indicating successful or erroneous completion, and specifying the time frame during which the process extracts data (Bregata, 2019). Metadata tables are connected to the Data Warehouse and are used extensively for data ingestion and analysis.

There are two categories of metadata:

- internal metadata: relevant for administrators, detailing sources, transformations, feeding policies, logical and physical schemas, constraints, and user profiles.
- external metadata: pertinent to users, encompassing definitions, quality, units of measure, and significant aggregations.

Metadata is housed in a specific container accessible to all other components within the architecture. Depending on the scope of the metadata, it is possible to distinguish two further categories:

- Global metadata: contains metadata pertinent to all levels and processes, facilitating synchronization across various phases at a common temporal or detail level.

- Process metadata is categorised according to the feeding system and the specific process in which it is engaged.

The company framework includes two metadata tables: FLOW_MANAGER and TABLE_MANAGER, each with additional technical fields such as JOBID, INS_TIME, and UPD_TIME.

- The **FLOW_MANAGER table** stores information about the history and unfolding of ETL iterations, detailing the current state and outcome of each ETL layer for every ingestion flow. Its purpose is to synchronize different ETL stages and processes and monitor pipeline advancement.

  Each record is uniquely identified by the iteration JOBID, an IDENTITY, a GROUP, and its LEVEL. The IDENTITY defines and aggregates flows with a common working area, while the GRP_NAME group assembles tables into functional entities. The LEVEL is a numerical representation of the ETL stage performed. Records gather information about the current STATUS of the process with a numeric value in the domain defined in the table below:

Value	Status
0	The job is complete without errors
1	The job is currently running
2	Job data is ready to be loaded in the next stage
-3	There are errors; the job is aborted

Table 7: numeric values of STATUS field

In this example, the LEVEL and the GROUP_NAME are incorporated in the same field:

IDENTITY	GRP_NAME	JOBID	STATUS	START_DATE	END_DATE
PROGETTO_TIROCINI_CARLO	L0_ANAGRAFICHE	20241118152300	0	2024-11-18 15:22:59.340	2024-11-18 15:23:01.250
PROGETTO_TIROCINI_CARLO	L0_ANAGRAFICHE	20241118153417	0	2024-11-18 15:34:17.120	2024-11-18 15:34:18.997
PROGETTO_TIROCINI_CARLO	L0_ANAGRAFICHE	20241118154732	0	2024-11-18 15:47:32.450	2024-11-18 15:47:40.697
PROGETTO_TIROCINI_CARLO	L0_ANAGRAFICHE	20241118154751	0	2024-11-18 15:47:50.507	2024-11-18 15:47:52.823
PROGETTO_TIROCINI_CARLO	L0_ANAGRAFICHE	20241118155112	0	2024-11-18 15:51:11.700	2024-11-18 15:51:14.750

Figure 15: FLOW_MANAGER table example

- The **TABLE_MANAGER table** tracks information about the iteration flow in relation to a table. Its primary key is characterized by four fields: IDENTITY, GRP_NAME, TABLE_NAME, and JOBID. These fields provide insights on ETL processes in relation to each table. NUM_ROWS is a counter of the number of rows produced as output from the current stage, which may indicate ingestion or integration errors or failure of data quality checks. INS_TIME stores the timestamp at which the table population has been completed.

IDENTITY	GRP_NAME	TABLE_NAME	JOBID	NUM_ROWS	INS_TIME
PROGETTO_TIROC	L0_ANAGRAFICHE	TEST_CARLO	20241120123820	11	2024-11-20 12:38:20.237
PROGETTO_TIROC	L0_ANAGRAFICHE	TEST_CARLO	20241120124513	10	2024-11-20 12:45:13.430
PROGETTO_TIROC	L0_ANAGRAFICHE	TEST_CARLO	20241120124631	11	2024-11-20 12:46:31.497
PROGETTO_TIROC	L0_ANAGRAFICHE	TEST_CARLO	20241120145334	11	2024-11-20 14:53:34.283
PROGETTO_TIROC	L0_ANAGRAFICHE	TEST_CARLO	20241120145705	8	2024-11-20 14:57:04.890

Figure 16: TABLE_MANAGER example

Metadata tables can be subject to business rules and supplementary checks, playing a fundamental role during the deployment and maintenance phases of a project. In specific major projects, further metadata tables can be implemented in the model to furnish a more extensive view on data sources, staging tables, Data Warehouse schemas, and execution flow. For instance, data types can be widely described in documentation tables, and an error management table can provide a collection of errors obtained during data quality checks. The error management table can be queried to retrieve a subset of errors deemed critical by the customer from a business perspective and then processed.

3.2 BRONZE LAYER-Staging Area

The L0 level in Data Warehouse Management (DWH) denotes the preliminary stage of data ingestion, during which information is extracted from source systems, such as relational DBMSs and customer files. Data ingestion is performed in batches on a periodic schedule for each table or document within a job identified by a JOBID which describes an ETL iteration instance with a timestamp in the format 'YYYYMMDDHHmmss'. Four types of tables are implemented in the ingestion stage: staging tables, delta tables, delta_his tables and error tables.

- **staging tables (STG)**: these tables contain data loaded from the source tables. Each loading is identified by a JOBID column to partition by time the rows.

- **delta tables (DLT)**: they store updates from sources and provide new records through three mechanisms:
    - FULL imports a copy of the entire table or file for each iteration. A FULL load is performed when an Initial Load is required for tackling new sources or populating a Data Warehouse for the first time.
    - CDC (Change Data Capture) the tables are subjected to a change data capture mechanism that automatically intercepts the delta of the data for each table compared to the previous extraction, and the replication process to L0 obtains the data already to be loaded. Recommended for replicating very large tables so as not to have to download tons of unnecessary data.
    - MINUS: To better explain the MINUS operations on STG tables it is possible to exploit the example below. Starting from the original STG table and partitioning by JOBID, we obtain two tables which refers to the yesterday snapshot 'SAR_STG_CARLO_IERI' and 'SAR_STG_CARLO_IERI1'. The third table 'SAR_STG_CARLO_OGGI' contains the current rows. From two MINUS operations we obtain 'IERI_MENO_OGGI' which contains the deleted rows and 'OGGI_MENO_IERI' which contains new rows.

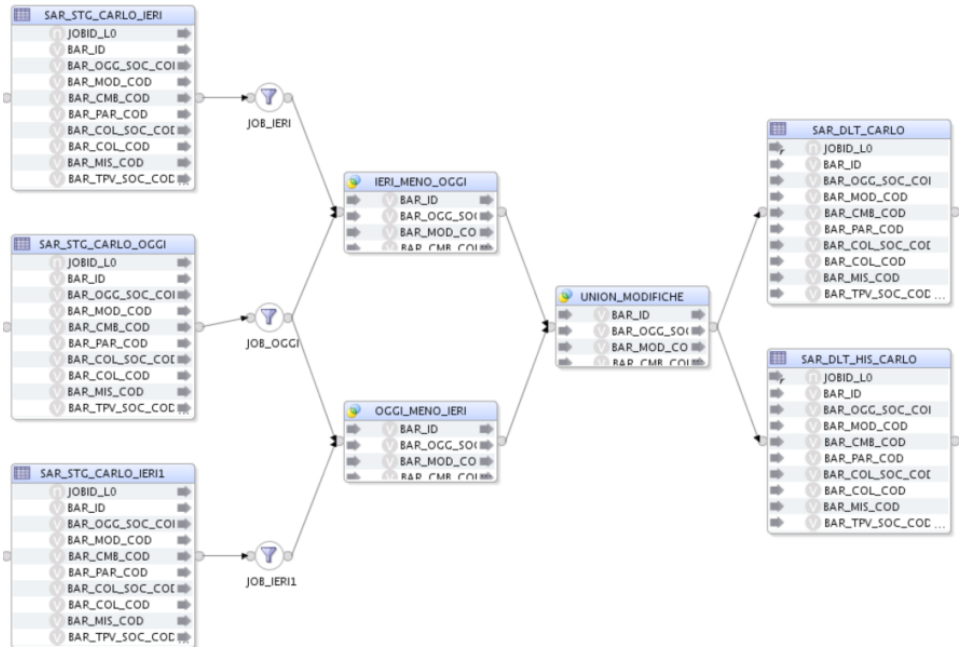


Figure 17:An example of a MINUS operation performed in an Oracle Data Integrator environment

The union then gathers the results and redirects the data flow to the corresponding DLT table called ‘UNION_MODIFICHE’. The output DLT table ‘SAR_DLT_CARLO’ schema includes an additional field ‘FLAG_NEG’ set to ‘1’ to perform a logical deleting (initialized originally to ‘0’ for validated records). DLT tables are partitioned by JOBID with a retention policy determined by the amount of ingested data.

- **DLT_HIS tables**: To address possible data losses or partial loading during the ETL process due to errors, a historicized version of the DLT table is required. A straightforward approach is to keep the history directly on the DLT tables and partition by JOBID. Another solution is to create DLT shadow tables marked by the prefix "DLT_HIS" which are not subjected to retention policies. In the example above, ‘SAR_DLT_HIS_CARLO’.

3.3 SILVER LAYER-Relational Data Store

The L1 section of the ETL process is crucial for ensuring data integrity, correctness, and dependability. It involves complex and computationally expensive procedures and transformations, such as data quality checks and business rules. The source tables at this stage are the DLT tables that insert the cleaned data into:

- **OK table:** they are used to evaluate data freshness, accuracy, completeness, consistency, and trustworthiness in a specific context, empowering business decisions and Data Governance programs.
- **ERR tables**

  They are included in the L0 layer, but their discussion and relevance are examined within the L1 data quality step. They store records that do not meet the data quality rules and are discarded temporarily for the next iteration. Error management is treated within ERR tables by extending the original schema with a description of the error and defining a retention period after which the ERR table records are permanently discarded.

- **ODS tables**

  Implement the conventional L1 layer, providing a centralized and normalized data model that encompasses all aspects of the company requirements. L1 tables aim to provide a consolidated and historicized version of the relational source data ready for integration. Since Relational Data Store needs to be historicized, ODS tables are populated with MERGE statements in an Incremental Update. In order to allow the records update, a primary key constraint must be defined on the physical key of the table. Moreover, tables are not subject to compression since frequent updates are performed, especially when many iterations are carried out in a single load process.

- **Master Data Management (MDM) tables**

  They fulfil two critical roles in the ETL process: integration from various sources and data enrichment. Data integration in the current stage consists of gathering matching data by joining two or more ODSs, accumulating attributes from many sources in a central table. MDM tables with descriptive attributes (dimension tables) serve the other purpose of assigning a surrogate key (SK), a unique identifier assigned to each record of a table with the aim of superseding the natural key. The most prominent gain from the use of a surrogate key is the increase in performance, as querying tables through surrogate keys is faster than compound or more descriptive keys, especially in expensive join operations. SK are defined with compact data types, such as integers, with a growing pattern (Greco, 2018).

3.4 GOLD LAYER-Publication Area

The Publication Area level serves as the central repository for business data storage and analysis, supporting strategic decision-making and Business Intelligence (BI) operations. Data is extracted from the silver layer, which collects and transforms operational data, and is then consolidated into publication tables. These tables retain historical data and are optimized for reporting and multidimensional analysis. In traditional relational databases, silver layer tables are generated through a series of JOIN operations between tables, using primary keys directly instead of surrogate keys, ensuring stronger referential integrity. In modern data warehouses,

the gold layer also supports Data Marts and Star Schema models to simplify data access. The design of the gold layer is continuously adapted based on customer needs and visualization tools used, ensuring efficiency in data processing and representation for business insights.

- OUT tables: prepare records for the publishing phase by collecting the required information for the multidimensional schema and transforming the records to fit the Data Warehouse model. They acquire the surrogate keys, preparing records for the visualization layer, especially for snowflake schemas. By merging various sources into one, multiple representations of the same value can be met, and the general pattern is choosing a common representation to prevent misalignment errors.

3.4.1 Data Visualization

Through data visualization, information derived from large-scale data structures becomes more accessible and easier to interpret. By leveraging graphical representations, relationships and patterns within the data can be identified, making complex insights comprehensible immediately. The visualization process ultimately transforms raw data into meaningful representations using shapes, colours, and interactive elements, ensuring that users can quickly and effectively understand trends and make informed decisions.

The tools employed are designed for business professionals who will use these dashboards to guide strategic decision-making. Business Intelligence (BI) frequently relies on data visualization, as it needs to be accessible even to those who are not experts in data analysis. The use of colours, proportional scaling, and graphical elements plays a crucial role in conveying whether key metrics are progressing positively or require attention. Every data-driven project generates valuable insights, but these insights hold significance only when they can be effectively understood and evaluated. Charts, tables, lists, and summary statistics are fundamental tools that enable analysts to process information and extract actionable knowledge.

However, the goal of data visualization is not merely to display data but to provide business users with a practical decision-support tool. Several key principles underpin effective data visualization. First, it must be user-oriented, ensuring that the language and presentation style are intuitive and cater to the knowledge level of the intended audience. Second, it should highlight key performance indicators (KPIs) by focusing on essential metrics rather than overwhelming users with excessive figures. Additionally, visualizations should reflect organizational structure, assigning responsibility for each KPI to a specific individual or

department for accountability. Another important aspect is the ability to compare data effectively, as raw numbers alone can be misleading without historical context or benchmark comparisons.

Furthermore, a well-designed visualization tool should offer a logical and comprehensible layout that enhances usability for different types of users. What works best for an analyst may not be ideal for an executive who requires a more streamlined summary. Lastly, data integrity and certification are crucial. Users must trust that the visualized data is accurate, consistent, and up to date; otherwise, they will be reluctant to rely on it for decision-making.

Various tools are available for data visualization, each offering unique functionalities and suited to different business needs. The choice of a tool may follow either a targeted or an open-ended approach. In a targeted approach, a company specifies a preferred tool based on its existing infrastructure and requirements, making it essential for the implementation team to understand that tool's strengths to maximize its effectiveness. Conversely, an open-ended approach allows the consulting firm or BI specialists to select the tool they deem most suitable for the given analytical context. In this case, evaluating the available options in the market and selecting the best tool for the specific data and business needs becomes a crucial step in ensuring successful implementation.

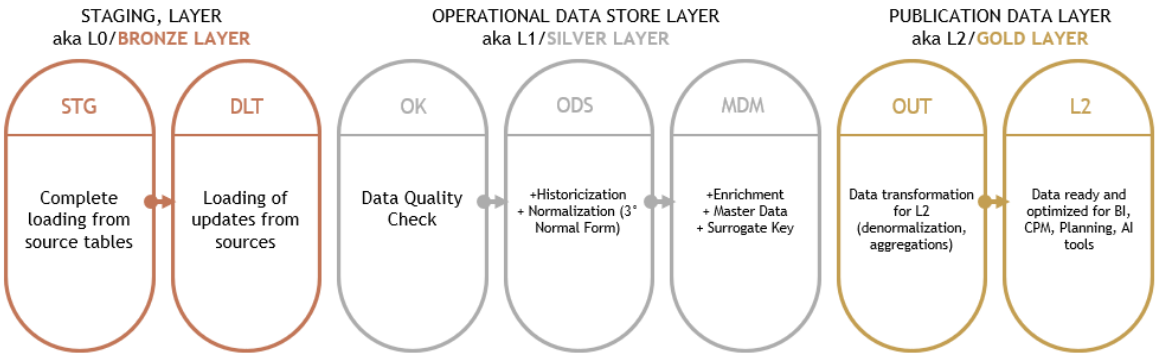


Figure 18: Company Framework

4 IBM KNOWLEDGE CATALOG

4.1 IBM

International Business Machines Corporation (IBM), also known as "Big Blue," is one of the world's major technological companies. IBM was founded on June 16, 1911, in Armonk, New York, and specialises in hardware, software, and information technology services. Its most significant innovations are the first electromechanical calculator, the first PC with the MS-DOS operating system, and the Watson Artificial Intelligence system. Today, IBM is active in a variety of sectors, including cloud computing, Artificial Intelligence, and quantum computing, and is well-known for its ongoing research and development initiatives (Wikipedia, 2025) (Pereira, 2024) (Cultur-e, 2018).

4.2 IBM Cloud Pak for Data

IBM Cloud Pak for Data is a complete, cloud-native platform for optimising data management and implementing Artificial Intelligence (AI) in enterprises. The architecture of IBM Cloud Pak for Data is anchored around three primary pillars: Collect, Organize, and Analyse. Each stage represents a fundamental phase in the data lifecycle, ensuring that raw, disparate datasets are transformed into actionable intelligence. Designed with a modular structure, IBM Cloud Pak for Data combines cutting-edge tools and services, each serving distinct yet complementary roles within a cohesive framework.

The Cloud Pak offer a smooth user experience through Watson API which guarantees a fluent use of IBM Knowledge Catalog, Watson Studio and Data Refinery. They can operate interchangeably within the platform thanks to the same user interface that makes difficult to distinguish what services and tools the user is using due to the consistent experience offered by Watson API.

Figure 19: IBM Cloud Pak Framework

Collectively, they constitute a scalable ecosystem designed to ensure that its constituent tools interact seamlessly, enabling an efficient and scalable data workflow leveraging on these key features:

- Unified Architecture: IBM Cloud Pak for Data unifies diverse data management technologies and tools, including as Watson AI, data virtualisation, and Data Governance, to create an information architecture that adapts to changing business needs (Lorusso, 2021) (IBM, s.d.).
- Flexibility and Scalability: The platform may be built on-premises, in the public, or private cloud, and it adapts smoothly to business requirements. It leverages Red Hat OpenShift to manage containers and provides a consistent user interface to facilitate data access (Vierrath, s.d.) (IBM, s.d.).

  Red Hat OpenShift is a Kubernetes-based container orchestration platform used for IBM Cloud Pak for Data, ensuring applications operate reliably across hybrid and multi-cloud infrastructures. (IBM acquired Red Hat in 2019, which develops OpenShift, an enterprise platform built on Kubernetes. Kubernetes, created by Google and now managed by the Cloud Native Computing Foundation (CNCF), is the container orchestration engine on which OpenShift adds advanced management and security features. OpenShift simplifies Kubernetes for enterprises by providing automation, integrated CI/CD, and enterprise support). It offers scalability, unified operations, automation, flexibility, and seamless integration with CI/CD pipelines. OpenShift dynamically allocates computing resources to meet workload demands,

providing optimal application performance. It also automates tasks like scaling, updates, and load balancing, reducing manual overhead.

- Data lifecycle automation: IBM Cloud Pak for Data automates data discovery and organisation, allowing for more dependable decision-making. This strategy shortens the time required to prepare data for analysis (Seidor, s.d.) (IBM, s.d.).
- AI and analytics support: The platform enables the simple integration of Machine Learning models and AI applications, transforming AI into an essential component of daily company operations. Watson's sophisticated features enable significant insights from collected data (Lorusso, 2021) (Vierrath, s.d.).

4.3 Best Practices for Data Governance program

To ensure data readiness and reliability, governance architecture is structured upon three principal domains:

1. Understand the data:
    - Data Profiling consists in analysing data with statistics to identify issues and ensure alignment with business initiatives. For doing this can be productive a partnership with the IT team.
    - Data Lineage tracks the origins and transformations of data to understand who handles it and where the data comes from. Can be a business lineage or a technical lineage.
    - Data Catalog: Create a comprehensive catalog for easy data discovery, allowing end users to find out data quickly, saving time. Data Governance needs to facilitate intuitive access to data. A catalog is composed by the following glossary:
        1. Business terms: define the entities in the domain. It is possible to import a file of business terms from external sources.
        2. Classifications: they are structured labels used to categorize data based on sensitivity, compliance, business domain, or data type. They help identify and tag PII (Personally Identifiable Information), financial data,

healthcare records, and confidential information. Classifications can be applied manually or automatically using Machine Learning.

3. Categories: allow users to easily affect the security around their governance artifacts. Categories must be effectively used for search quicker assets around the organization. A category is like a folder or directory to organize the governance artifacts and the users that can manage or use these artifacts. Categories can be organized in a hierarchical structure, including subcategories. The more the hierarchies, the more specific will be the control over the different artifacts. In the Access Control tab it is possible to control which users can view or manage categories and their governance artifacts by adding users as collaborators to categories.
4. Data classes: describe the format of the data.

2. Protecting the data:
    - Data Security means to establish policies and rules. Policies state who have access to a data. In the data security step, data owner and data steward participate. Policies can also refer to how a data is represented and written (for example a phone number which has a specific format).
    - Data Compliance ensure adherence to internal and external regulations.
    - Data Lifecycle answers to the following questions: “How long I need to keep that data? Which mechanism I must use to store that data?”. It is possible to store as much data as needed, but it is better to not overspend that capability. In summary, it defines retention policies and storage mechanisms.
3. Curate the data:
    - Data Quality Management: it ensures the right data is used for each business initiative.
    - Data Integration: collecting data from various sources must fit together coherently through processes and tools.
    - Master Data Management (MDM): It is like data integration, but it takes on special responsibilities for certain entities like customer, suppliers, products. It

ensures that the same data can be accessed by different entities. Master Data Management is essential when data is across various sources, and without an easy integration a good catalog is useless. “The best MDM you could do, it’s the MDM you don’t have to do, because it was already done”.

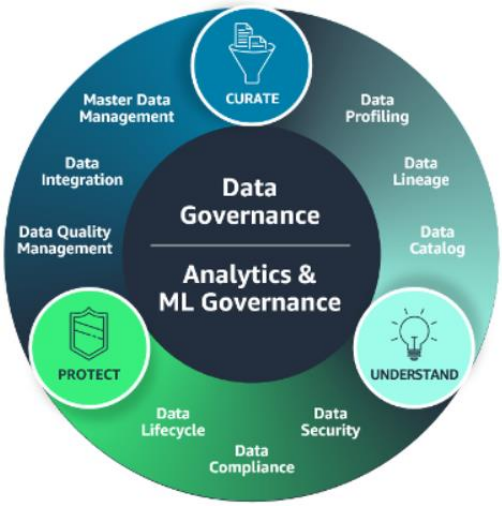


Figure 20: Data Governance Model

4.4 IBM Knowledge Catalog – Data Integration and Cataloguing

Data Integration and Cataloguing is performed by IBM Knowledge Catalog: raw data is ingested into IBM Knowledge Catalog that underpins the organization phase.

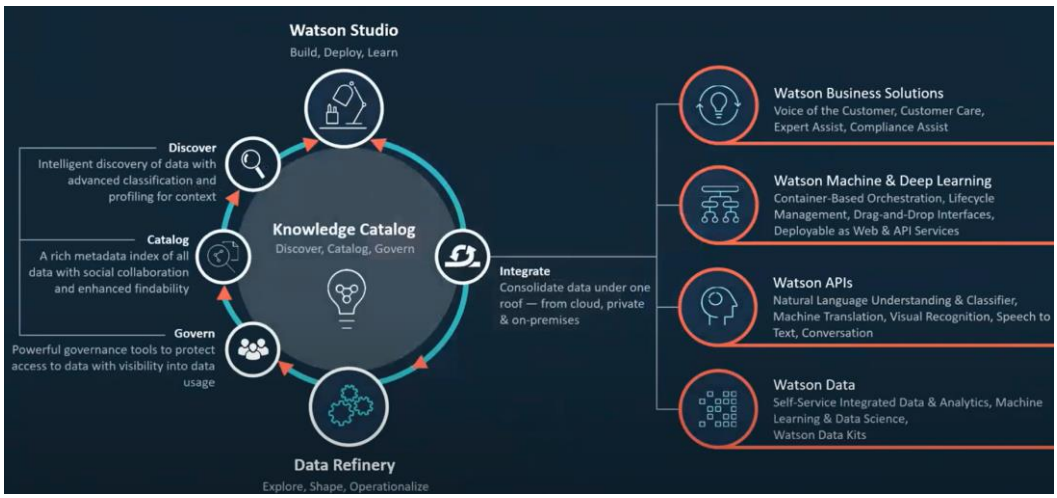


Figure 21: IBM integration features

IBM Knowledge Catalog (IKC) is a comprehensive Data Governance and cataloguing solution that enables organizations to effectively manage, curate, and protect their data assets. It is incorporated within IBM Cloud Pak for Data and functions as a central repository for data

discovery, quality management, and adherence to regulatory standards such as GDPR and CCPA. IKC offers intelligent, self-service functionalities for data and model discovery, fostering collaboration and expediting data-driven activities. It enhances the administration and organisation of data and knowledge assets, rendering information more accessible and usable for Artificial Intelligence and data analysis endeavours. IKC is offered as software as a service (SaaS) delivering a comprehensive enterprise metadata repository enhanced with active metadata (IBM, 2020) (Grossi, 2020). Its functionality encompasses:

- Metadata Enrichment: Automated tagging and documentation to create detailed metadata repositories. This ensures that datasets are searchable and contextualized for diverse user needs.
- Governance Mechanisms: IBM Knowledge Catalog is designed to help organizations comply with regulations such as GDPR and CCPA through several mechanisms:
    - Dynamic Data Masking: Sensitive information is automatically masked in real-time when accessed by unauthorized users. This feature is vital for maintaining privacy while allowing legitimate users access to necessary information.
    - Audit Trails: Detailed logs track all user interactions within the platform, enabling organizations to demonstrate compliance during audits. These audit trails provide transparency and accountability in how data is accessed and used.
    - Policy Enforcement: Governance policies can be embedded directly into the catalogue environment using business terms, classifications, and rules that align with regulatory requirements. By embedding these policies into daily operations, organizations can ensure ongoing compliance without additional overhead. Policies are textual information that describe the various kind of policies itself. They can have a hierarchical structure. Can be created in a text file and then imported into the project. Policies are enforced by the rules. The rules are composed by a criterion plus an action.

- Data Lineage Tracking: Comprehensive visualization of data flows, detailing origins, transformations, and destinations. This transparency supports auditability and trust.
- Access Control Protocols: Configurable role-based permissions that safeguard sensitive data while granting appropriate access to authorized stakeholders.

Figure 22: IKC functionalities

Through these capabilities, IBM Knowledge Catalog establishes a robust framework for managing enterprise data while seamlessly integrating with analytical tools. Furthermore, IBM Knowledge Catalog fosters team collaboration by providing a centralized platform where team members can access, share, and curate data assets through:

- Shared Workspaces: Teams can create catalogues tailored to specific projects or departments, enabling members to collaborate on datasets in real-time. For example, marketing teams can curate governed datasets for analytics projects while ensuring compliance with governance policies. This capability is critical for fostering a collaborative environment where insights can be shared and leveraged across different functional areas.
- Annotations and Feedback: Users can annotate datasets with comments or notes, facilitating discussions about data quality or usability. This feature ensures that insights are shared across teams and that all stakeholders have a clear understanding of the data. Such collaborative features enhance the overall quality of data analysis by incorporating diverse perspectives.
- Role-Based Collaboration: By assigning roles such as Viewer or Editor within a catalogue, team members can collaborate securely while maintaining

appropriate access controls. This security framework allows organizations to manage who can view or edit sensitive data effectively.

These capabilities promote transparency and efficiency in data-driven projects, making IBM Knowledge Catalog an invaluable tool for cross-functional teams.

4.4.1 Advantages of IKC

IBM Knowledge Catalog offers several advantages compared to traditional data management systems:

- Advanced Metadata Management**:** The platform uses AI to enrich metadata automatically by adding context, labels, or descriptions to thousands of assets simultaneously. This reduces manual effort and enhances the discoverability of data. As noted by IBM, this automated enrichment process significantly improves the user experience by allowing faster access to relevant datasets.
- Integration with AI/ML Tools: IKC is designed to support AI and Machine Learning workflows by providing high-quality, curated datasets that are ready for analysis. The ability to integrate seamlessly with Machine Learning frameworks positions IKC as a vital component in organizations' analytics strategies.

  IBM Cloud Pak offers several tools to facilitate the integration of unstructured data: Watson business solutions can recognize voice of customer, focusing on customer care solutions and compliance assistant. In addition to Watson business solutions, the integration step can involve also Watson machine and deep learning, a very useful tool with drag and drop interfaces, lifecycle management and container-based-orchestration. In order to manage with natural language understanding and classifier, machine translation, visual recognition and speech to text services, Watson provides a Watson APIs that can perform these activities. Furthermore, Watson data comes in for data analytics and Machine Learning and data science. So, the integration part around a Data Governance program can be performed exploiting several tools in the cloud Pak platform of IBM. Remembering that the data integration step objective is to bring metadata in the platform, not data itself.

- Integration with other Business Applications: It interfaces with ETL tools, automating the ingestion of structured and unstructured data from databases, data lakes, and cloud storage platforms. It additionally supports Business Intelligence Platforms such as Cognos Analytics, augmenting analytical capabilities throughout the organisation. IKC supports more than 50 connectors for third-party applications such as Salesforce, Microsoft SQL Server, Dropbox, Cloudera, and Looker, and may be tailored with custom JDBC drivers for more integrations. This comprehensive integration enables organisations to utilise their current technology infrastructure while gaining advantages from IKC's sophisticated cataloguing features (SmallNet Consulting, 2025).
- Comprehensive Governance Framework: Unlike many competitors, IKC integrates governance artifacts directly into the cataloguing process, ensuring compliance with regulations like GDPR and CCPA without additional tools. This built-in governance simplifies compliance management for organizations.
- Data Virtualization: IKC eliminates the need for extensive ETL processes by enabling users to access data where it resides through virtualization tools. This capability not only saves time but also reduces the complexity associated with traditional data integration methods. Furthermore, through data virtualization users can access data from a unique interface, improving the operational efficiency and the facilitating the analysis of data.

4.4.2 Challenges of IKC

While IBM Knowledge Catalog offers significant benefits, its implementation may present some challenges:

- Complex Permissions Management: Configuring granular role-based access controls (RBAC) requires careful planning to ensure that sensitive data is protected while maintaining usability for authorized users. Organizations must invest time in defining roles clearly to avoid potential security risks.
- Data Migration: Migrating legacy datasets into IKC may require extensive preparation, including metadata enrichment and quality assessments. This process can be resource-intensive but is crucial for ensuring that all data is accurately represented in the new system.

- User Training: Ensuring that all team members understand how to use IKC effectively may necessitate comprehensive training programs tailored to different roles (e.g., Data Stewards vs. Analysts).

Organizations should address these issues proactively by developing clear implementation plans.

4.4.3 Economic impact of IKC

IBM Knowledge Catalog offers advanced capabilities that enable enterprises to manage, govern, and optimize their data assets. The Forrester study, commissioned by IBM, evaluates the total economic impact (TEI) of IBM Cloud Pak for Data. Analysing four companies, the report projects a return on investment (ROI) between 274% and 459% over three years, thanks to improvements in infrastructure efficiency, Data Governance, and data science/AI capabilities. The benefits and costs are detailed, considering different impact scenarios (low, medium, high). The TEI method by Forrester provides a framework for evaluating the investment, considering risks and flexibilities. IBM Knowledge Catalog economic advantages span several dimensions:

1. Infrastructure and Management Cost Optimization

- Resource Efficiency Through Automation: Leveraging containerized architectures and Kubernetes, IBM Knowledge Catalog enhances hardware utilization and minimizes operational overhead. According to empirical evidence from Cloud Pak for Data users, automation has facilitated infrastructure management cost reductions between 65% and 85%.
- Unified Licensing: By consolidating disparate solutions into a single platform, organizations achieve significant reductions in licensing and maintenance costs. Forrester's projections indicate potential infrastructure and licensing savings of up to $7.2 million over three years.

2. Enhanced Productivity of Data Professionals

- Streamlined Data Accessibility: Advanced data cataloguing and virtualisation capabilities facilitate universal access to enterprise data, reducing dependency on manual extract-transform-load (ETL) processes. Forrester's findings highlight a reduction in ETL-related tasks by 25% to 65%, enabling data professionals to focus on high-value analytical activities.

- Accelerated Model Deployment: Simplified workflows for Machine Learning and AI model development significantly enhance time-to-market, translating into increased revenue generation through faster delivery of impactful analytics solutions.

3. Risk Mitigation and Regulatory Compliance

- Governance and Security: IBM Knowledge Catalog ensures robust adherence to regulatory standards such as GDPR and CCPA, facilitating effective policy enforcement and comprehensive data lineage tracking. This reduces exposure to regulatory penalties.
- Minimized Data Breach Risks: Strengthened access controls and auditing mechanisms mitigate vulnerabilities, safeguarding enterprise reputation and reducing the financial implications of security incidents.

The implementation of IBM Knowledge Catalog delivers measurable economic benefits that extend beyond operational efficiencies:

- Net Present Value (NPV): Forrester estimates a three-year NPV of $9.7 million, with a return on investment (ROI) as high as 459% (Forrester applied a risk-adjusted approach with a discount rate of 10%).
- Data-Driven Decision-Making: A unified source of truth enhances the reliability of strategic insights, enabling organizations to make informed, data-backed decisions.
- Avoidance of Opportunity Costs: Improved access to high-quality data empowers enterprises to capitalize on emerging market opportunities and adapt to dynamic business environments.

4.5 Data Refinery – Data Preparation

Data Preparation is performed by a Data Refinery flow: datasets are cleaned and transformed using the Data Refinery tool of IBM Watsonx, focused on data preparation, cleansing, and transformation. Data Refinery empowers users to transform raw datasets into analysis-ready formats, accelerating the process of insight generation. IBM Watsonx is an AI platform comprising Watsonx.ai, Watsonx.data, and Watsonx.governance, which includes a development

environment for AI models, a data management platform for AI, and a tool for governing AI models.

Data refinery is not DataStage, but it is an easy-to-use tool which enable users to connect assets of different types, calculations without coding. Data Refinery access and explore data that resides across a myriad of endpoints: on-premises, or on-cloud. Connectors to IBM, non-IBM and third-party data sources.

In Data Refinery users can visualize the data thanks to interactive explorations and data visualizations to gain deeper insights from data using, also statistical techniques are built into that tool for automatically detect and label data types, anomalies and sensitive fields. Data refinery helps users to explore and prepare data as a preliminary step before the data cataloguing.

In a Data Refinery stage is possible to uncover patterns and refine data through shaping and cleansing techniques. One of the key features of data refinery is the possibility to operationalize for repeatability, scalability and accountability. A very powerful way to improve the repeatability is to create scheduled data flows for repeatable outcomes. The operationalization can also monitor and inspect the results of a data refinery stage.

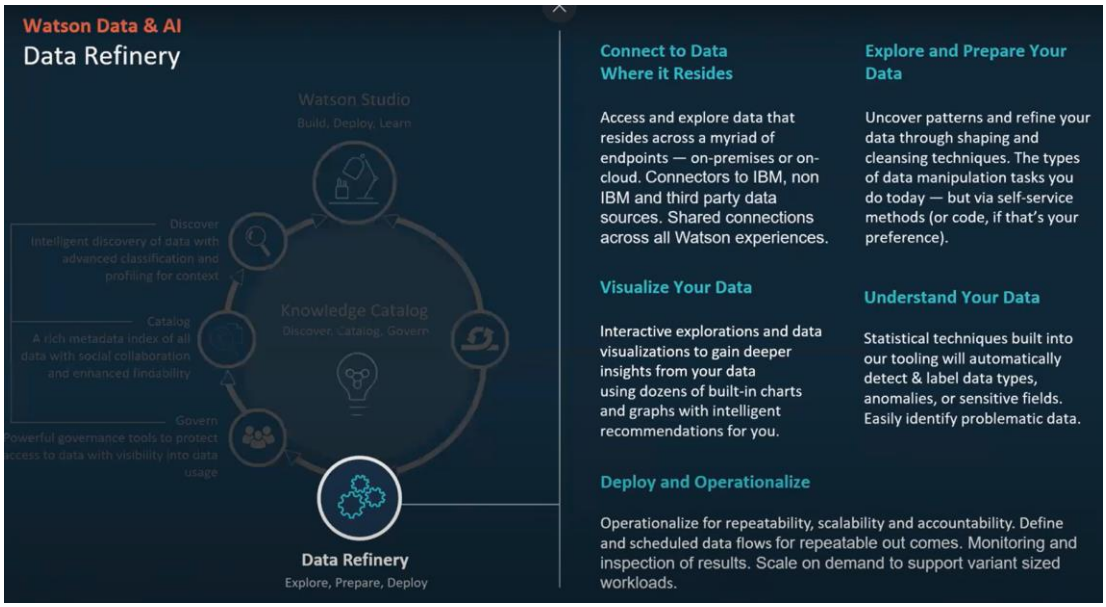


Figure 23: Data Refinery features

After data is analysed and explored in data refinery, data is subject to the operations of govern, catalog and discover:

- Govern: powerful governance tools to protect access to data with visibility into data usage. This phase embraces the setup of the policy management. Through a policy management is possible to deny or allow access to data consumers or apply a data masking on sensible data. It is possible to define a hierarchical data access to mitigate unauthorized exposure. Policy manager acts thanks to data classification through tags.
- Catalog: search, explore, consume data asset. When an asset is published, it is available to consumers. When a data consumer access the catalog, he can search through tags among the assets of a catalog. Based on the role and permissions he will be able to see or access to certain assets, implementing structured ownership designations to regulate data stewardship. An asset can be a dataset, a notebook or a metadata.
- Discovery: analyse, identify, classify assets. Data discovery is made for scaling governance with MI, AI and automation, to understand which data is available and where it resides. After having found all data, utilizing Machine Learning algorithms to automate classification of data by type.

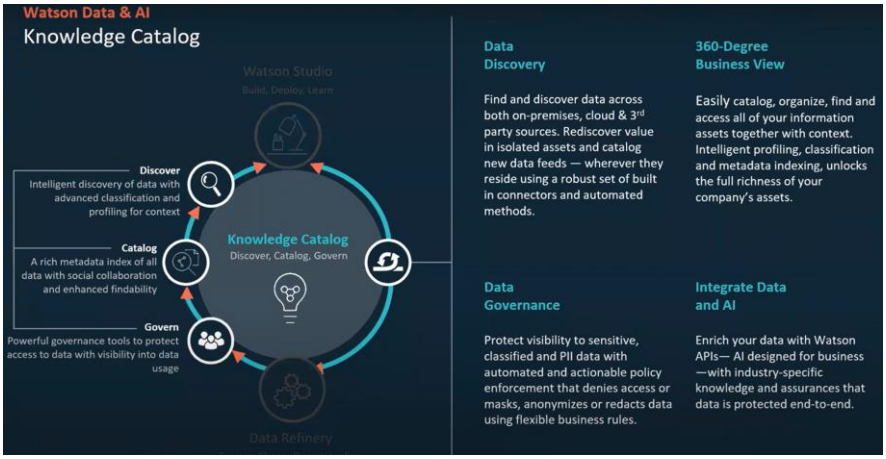


Figure 24: Govern, Catalog and Discover steps

4.5.1 Roles

Among the users throughout the platform, we must meticulously define governance framework that requires precise delineation of roles:

- Data consumer: Analysts and business intelligence specialists leveraging data insights for market expansion and strategic planning. without being burdened by data management responsibilities. In according to the role, some access can be denied, or some data can be masked. Furthermore, the loading of assets in specific projects can depend on the role and permissions of the users.

  The main idea is to adopt a decentralized organization (modern data community) because the aim of a good governance is to avoid burdening data consumers of responsibilities for data management, empowering them to leverage on data they need, without being burdened of management responsibilities.

- Data steward: Custodians of data integrity, ensuring adherence to governance protocols. Assign individuals responsible for understanding and managing data relevant to specific business activities. A data steward is a person who knows which data is appropriated for a certain business activity and to support it (it is a day-by-day involvement).

- Data Owner: Senior executives overseeing policy formation and regulatory enforcement. Define individuals responsible for setting data policies and access controls. They decide who has access to that data.

- Data Producers: Cross-functional units responsible for maintaining transactional and operational datasets. Empower teams to develop and manage their own data products while adhering to governance policies.

- Centralized Governance Oversight Committee: A supervisory entity tasked with ensuring enterprise-wide adherence to governance standards. Facilitate collaboration and ensure semantic consistency across data products developed by different teams. Its role is crucial especially when the main object is not to centralize the analyst activities, but to focus on data products. When data is integrated various channels and sources managed by data producers (responsible for their area of business) data must be joined together (“are the join going to

work?") and must have a semantic sense, this ensures a cohesive and fluent customer experience. The semantic sense is often missed in a centralized organization.

Users	Description
Manager	Oversees catalogues, governance rules, and access policies.
Reporting Administrator	Manages report generation and metadata visualization.
Data Steward	Manages data, improves quality, and publishes assets in catalogues.
Data Engineer	Creates data connections, prepares, and publishes data for analysis.
Data Scientist	Accesses catalogues to retrieve datasets for AI model training.
Governance Artifacts Administrator	Administers governance artifacts and manages permissions.
Data Quality Analyst	Analyses data quality and identifies reliability issues.

Table 8: Main users in IKC

Role	Permissions	Description
Administrator	Access data protection rules, manage governance rules, manage governance workflows, find a resource using global search and tagging, view IAM access details, administer governance rules, execute data quality details, manage data quality assets, add catalog assets to data history	Assign this role to individuals who configure and administer IBM Knowledge Catalog or watsonx.ai Studio and perform the following tasks: - watsonx.ai Studio users with this role can participate in any project as administrators and view all active projects in the account. IBM Knowledge Catalog users with this role must make decisions about organizations, workflow, and importing user governance resources, as well as deciding which users can perform specific activities and which catalogues to create. The Administrator role includes all permissions granted in other roles, except for the following permission: Manage reporting.
Reporting Administrator	Manage reports	Assign this role to individuals who need to create reports on assets in catalogues. Note: Users with this role can send all metadata from a project, catalog, or category to an external database, regardless of membership or access permissions in existing projects, catalogues, and categories. Assign this privileged role with caution.
Data Steward	Access catalogues, access governance resources, manage	Assign this role to individuals who need to perform the following tasks: Implement the governance framework

	data protection rules, add catalog assets to projects	by creating governance resources. Curate data by importing metadata, enriching metadata, performing data quality analysis, and publishing data assets in catalogues.
Data Engineer	Access governance resources, manage data protection rules	Assign this role to individuals who create connections and then prepare and publish data assets in catalogues.
Data Scientist	Access catalogues, access governance resources, add catalog assets to projects	Assign this role to individuals who need to perform the following tasks: Find data assets in catalogues and use the data to train models in projects. Document and manage models in the catalog.
Governance User Resource Administrator	Administer governance resources	Assign this role to individuals who need to perform the following tasks: View and modify all governance user resources across all categories, modify categories (including changing contributors and permissions), execute all API calls for governance user resources, and set rules and rule conventions.
Data Quality Analyst	Drill down into problem details, execute data quality rules, manage data quality assets	Assign this role to individuals who need to configure and execute data quality analysis and evaluate analysis results.
Data Source Administrator	Create data source definitions and view a list of all account connections	Assign this role to individuals who need to create data source definitions.

Data Source Creator	Create data source definitions	Assign this role to individuals who need to create data source definitions.
Policy Decision Operator	Evaluate policy decisions	Assign this role to individuals who evaluate data access requests on behalf of other users.
Lineage Administrator	Access data lineage, create data source definitions	Assign this role to individuals who need to import lineage metadata and manage imported lineage data.

Table 9:Predefined Roles for IBM Knowledge Catalog

4.6 IBM Watson Studio – Advanced Analytics

Advanced Analytics is performed by IBM Watson Studio which is an Artificial Intelligence (AI) platform included in Watsonx that allows enterprises to incorporate sophisticated cognitive functionalities into their operational procedures. Curated datasets are accessed in IBM Watson Studio for Machine Learning model development and exploratory analysis. Once the data is catalogued, we can use Watson studio capabilities to do auto classification, profiling and statistics. In the Watson studio environment, there is a strong use of machine and deep learning to craft models and compare results. A direct integration with Watson Machine Learning can allow users to exploit the latest deep learning techniques.

Learning Deployment is performed by IBM Watson Machine Learning enabling real-time decision-making and integration into enterprise systems. WML is integrated within Watson Studio and constitutes a component of the Watsonx ecosystem, emphasising the development, deployment, and governance of AI models in corporate settings.

Watson Studio enables collaborative work among data scientists and supports diverse coding languages such as Python, R, and Scala. IBM Watson provides a diverse array of AI-driven services that empower organizations to harness the full potential of their data. On the other hand, one of the key factors that drive to success a Watson tool is the absence of coding that enable users to design and train Machine Learning models without having knowledges on coding. In the Watson studio tool, there is the possibility to have access to multiple way to

visualize data, but moreover it fosters collaboration across teams by bringing together data and talent under one proof.

One of its standout features is Natural Language Processing (NLP), which allows Watson to understand and interpret human language. This capability enables it to process extensive amounts of unstructured text, making it particularly valuable for sophisticated applications such as chatbots, virtual assistants, and automated customer care systems (Kalibbala, 2023), as well as conducting sentiment analysis and emotion monitoring for clients (New Gen Apps, 2018).

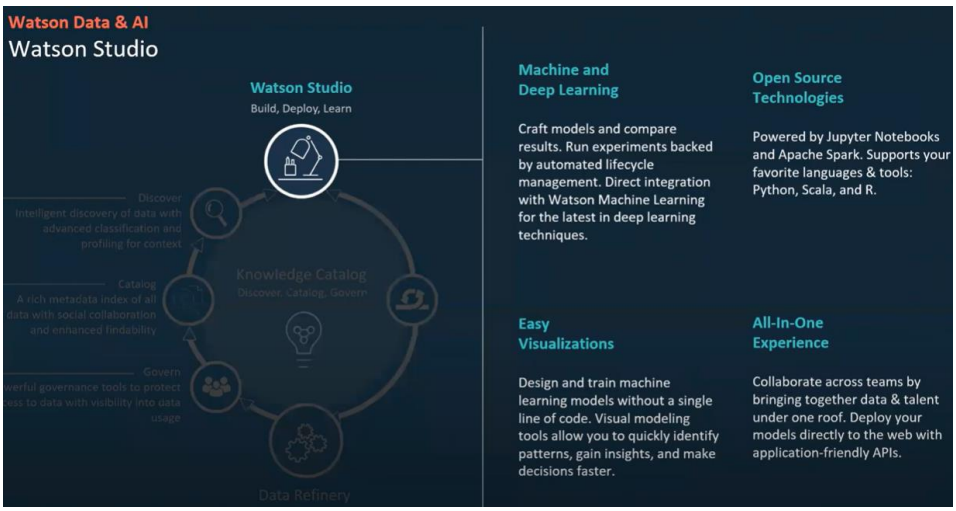


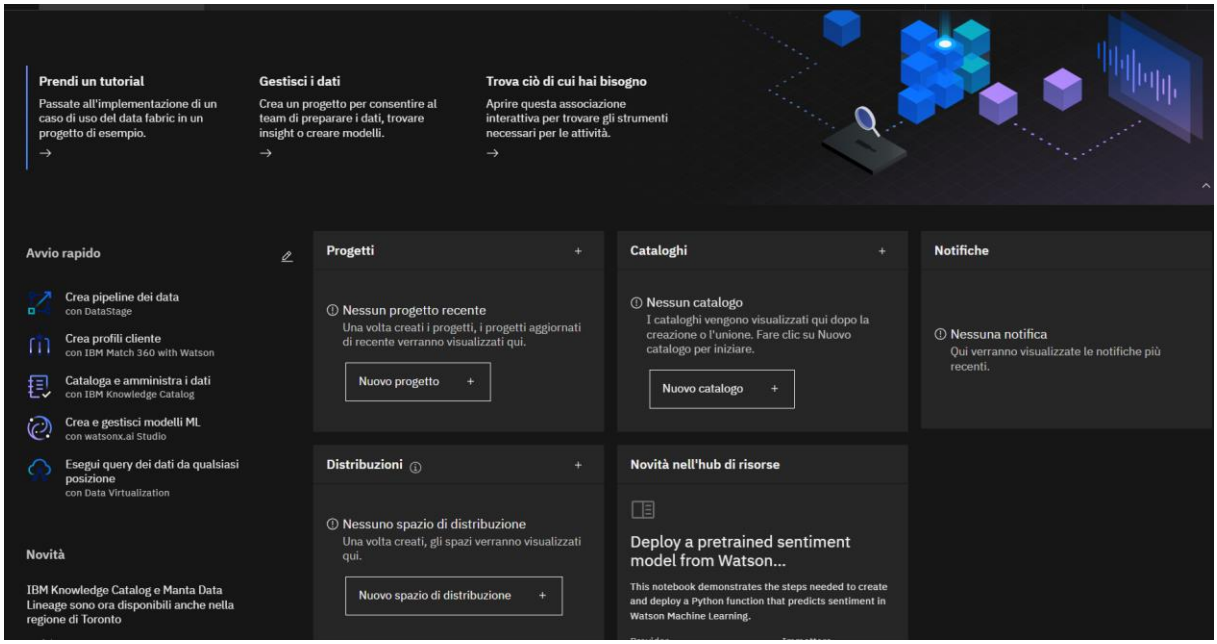
Figure 25: IBM Watson features

Furthermore, Watson employs Machine Learning techniques and predictive analytics to identify patterns in data, offering valuable recommendations that organizations can use to enhance model performance continually (Kalibbala, 2023; Tot, 2024). Watson's integration with IBM Cloud Pak for Data enables seamless deployment of AI models in both cloud and on-premises environments. This collaboration enhances data accessibility and organization while facilitating advanced analytics (Kalibbala, 2023; Carter, 2024). The platform's customization and scalability are also noteworthy. Organizations can tailor Watson's functionalities to meet specific requirements through its adaptable APIs. Support from Red Hat OpenShift ensures that Watson can scale effectively, making it suitable for multi-cloud and hybrid environments (IBM, n.d.; Tot, 2024). Collectively, these capabilities enable organizations to derive significant value from their data, improve business decisions, and automate complex processes.

5 PROJECT WORK

In contemporary data-driven business landscapes, effective Data Governance is paramount, particularly within data-intensive sectors such as the fashion industry. The proliferation of data necessitates robust governance frameworks to ensure data quality, compliance, and strategic alignment with business objectives. In this chapter, I will present a case study detailing my collaboration with a fashion company to implement a robust Data Governance framework leveraging IBM Knowledge Catalog.

The objective is to outline the procedural steps and key considerations for applying Data Governance principles in a real-world scenario, with the anticipation of integrating actual data and incorporating feedback from company personnel. Below, is presented an exhaustive analysis of each procedural step, integrating theoretical frameworks and industry best practices.



This is IBM Cloud Pak platform interface from which we can create and manage projects, catalogues and resources. The first step of the project is to create a project in IBM Cloud Pak for Data, adding a name and a description of the project. The project name is “PROJECT WORK TESI CRESCENZI”. It is possible also to add tags useful to identify projects and accelerate research. After having created the project instance we can move towards the next steps.

5.1 Understand the Data

Implementing a robust Data Governance framework necessitates a methodologically rigorous approach that aligns the Data Governance initiatives with corporate objectives, ensuring regulatory compliance, information security, and operational accessibility.

Starting from individuating the project to perform, it is necessary to find out which data is needed and who will manage that data. We undertake a critical examination of the motivations underpinning Data Governance adoption, encompassing regulatory adherence, operational refinement, and advanced analytics for retail intelligence. For example, if the company wants to begin an optimization pricing model, then it would start to think which data is necessary to perform this initiative (data price history, locations, products and so on).

A well-structured Data Governance initiative requires an initial assessment to define:

- Key Business Priorities: Identifying how governance enhances data-driven decision-making, regulatory compliance, and operational efficiency.
- Baseline Data Maturity Analysis: Conducting a diagnostic evaluation of existing data handling practices, their deficiencies, and potential areas for optimization.
- Performance Metrics and Key Performance Indicators (KPIs): Establishing quantifiable benchmarks to evaluate governance success over time and demonstrate its value to stakeholders and defining quantifiable metrics to assess governance effectiveness, including compliance rates, data accuracy improvements, and policy adherence levels. These indicators provide quantifiable insights into business performance, enabling decision-makers to optimize inventory management, forecast sales trends, and ensure financial efficiency.

  The KPIs utilized in this study are as follows:

  - Net Sales = Σ (Total Sales – Returns): This metric captures the total revenue generated after deducting product returns, ensuring an accurate representation of actual sales performance.

- Discount Rate = Σ (Discount Applied) / Σ (Total Sales) * 100: This metric assesses the percentage of total revenue lost due to promotional discounts, helping the pricing team optimize future campaigns.
- Standard Production Cost = Σ (Unit Cost * Quantity Sold): Represents the total cost incurred in producing the goods sold within a specific time frame.
- Gross Margin = (Net Sales - Standard Cost) / Net Sales * 100: This percentage-based KPI is crucial for evaluating profitability across different product lines.
- Budgeted Sales = Forecasted Sales Volume * Expected Price: Used for strategic financial planning, this metric estimate projected revenue based on market expectations.
- Forecasted Sales = Machine Learning-based prediction using historical data: This KPI leverages AI-driven predictive analytics to estimate future sales volumes based on past trends and market conditions.

The systematic tracking of these KPIs ensures that the company maintains alignment with its business objectives, facilitating data-driven decision-making. The update frequency of each KPI is tailored to operational needs, with some indicators, such as Net Sales and Discount Rate, requiring daily monitoring, while others, such as Budgeted Sales and Forecasted Sales, being assessed over longer periods.

KPI Name	Formula	Update Frequency	Data Source	Users Involved
Net Sales	Σ (Total Sales – Returns)	Daily/ Weekly	ERP Database	Sales Managers, Finance Team
Discount Rate	Σ (Discount Applied)/ Σ (Total Sales) * 100	Daily	ERP Database	Business Intelligence Analysts, Pricing Team
Standard Production Cost	Σ (Unit Cost * Quantity Sold)	Monthly	ERP Database	Finance, Cost Control Team
Gross Margin	(Net Sales - Standard Cost) / Net Sales * 100	Weekly	ERP Database	Executive Board, Finance Department
Budgeted Sales	Forecasted Sales Volume * Expected Price	Annually/Semi-Annually	Planning Tools	Executive Board, Sales Strategy Team
Forecasted Sales	Machine Learning-based prediction using historical data	Quarterly	Forecasting Tool	Sales Planning Team

Table 10: KPIs

- Ongoing Skill Development and Knowledge Transfer: Institutionalizing governance knowledge through specialized training programs, workshops, and certification initiatives to build internal expertise.

The preliminary analysis of the sales data from the Data Mart revealed crucial challenges, as highlighted by the IT team's responses. Specifically, the presence of errors in the data and incomplete information, stemming from different loading streams for time zones (USA, EMEA, Asia), requires immediate attention. Although integration between systems has been improved with the creation of a new DWH, the imperfect synchronization of data from various sources remains an obstacle.

To address data quality issues, it will be necessary to implement stricter quality controls on incoming data, with automated validations and alerts in case of anomalies. Additionally, a data normalization process and direct communication with customers to resolve inconsistencies will be essential to ensure the accuracy and reliability of sales data. Advanced data profiling techniques help detect outliers and patterns that influence sales trends.

5.1.1 Data Profiling

A fundamental step in implementing a robust Data Governance framework is the initial assessment of the available data, ensuring its quality, structure, and alignment with business objectives. This phase establishes the foundation for subsequent governance processes, facilitating efficient data integration and utilization.

The dataset provided by the company consists of 34 Excel files containing descriptions of the tables that compose the sales Data Mart of the company. This dataset is structured according to a Snowflake Schema. Below is represented the snowflake schema, drafted on excalidraw.com.

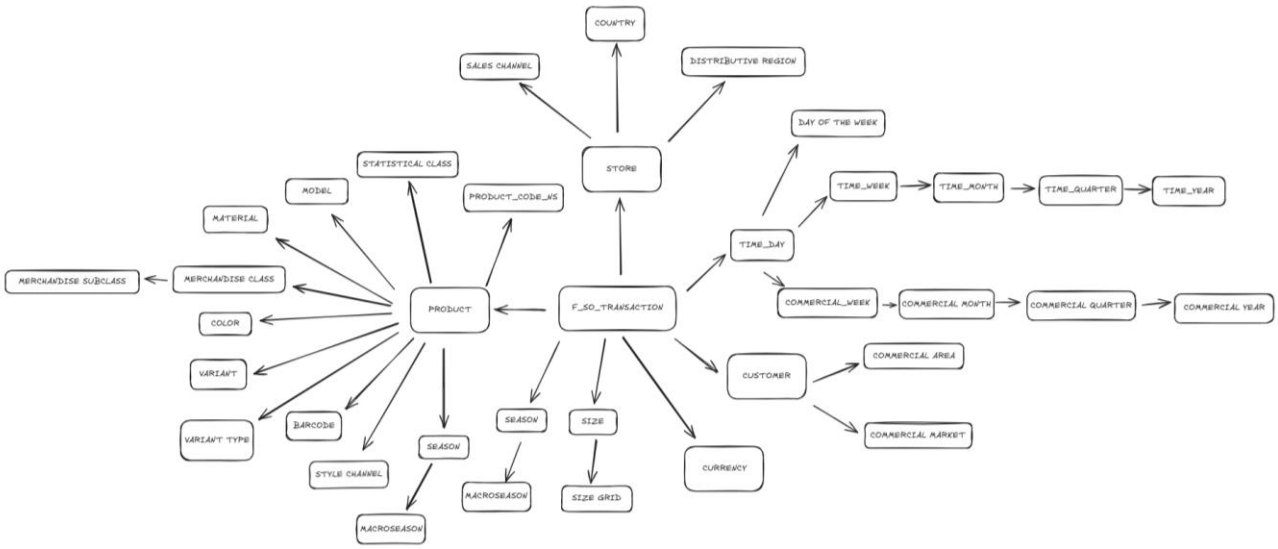


Figure 26: Snowflake Schema

The dataset is composed of two primary categories of tables:

- Dimensional Tables: These contain descriptive attributes that provide context to transactional data, enhancing analytical capabilities. Notable dimension tables include:
    - Product: Contains information on sold products.
    - Model, Material, Colour, Variant: Provide further product-level granularity.
    - Store, Sales Channel, Country, Customer: Define store locations, sales channels, and customer demographics.

- Time: A temporal dimension with detailed time attributes spanning days, weeks, months, quarters, and years.
- Additional supporting tables such as Currency, Statistical Class, Style Channel, and Barcode.

- Fact Table:
  - F_SO_TRANSACTION: Serves as the central table in the schema, capturing sales transactions. It includes foreign keys referencing dimensional tables and stores measurable business metrics such as sales volume, revenue, and applied discounts.

COLUMN_NAME	DATA_TYPE	NULLABLE	DATA_DEFAULT
JOBID_L2_INS	NUMBER	Yes	
JOBID_L2_UPD	NUMBER	Yes	
COMPANY_TRANSACTION_COD	VARCHAR2(400 CHAR)	No	
HEADER_ID	VARCHAR2(400 CHAR)	No	
STORE_SK	NUMBER	Yes	
STORE_REGISTER_SK	NUMBER	Yes	

Figure 27: Example of the structure of the excel tables

Understanding this schema is essential for defining strategic objectives and ensuring that the data is properly structured for governance and analytical workflows. The subsequent steps will focus on integrating this dataset into a comprehensive governance system, leveraging IBM Knowledge Catalog for metadata management and compliance.

The interview with the IT team confirmed that the Data Mart is fed by several sources, including POS, e-commerce, ERP, and APEX. The data is updated three times a day, divided by geographical areas (Americas, EMEA, Asia). However, errors such as missing data and inconsistencies between systems frequently occur. To improve data quality, the company has implemented reference checks, alerts on missing sales, standardization of processes, and a direct communication system with customers. It will be important to integrate these existing controls with IBM Knowledge Catalog for more centralized and automated data quality management.

5.1.2 Data Lineage

Data Lineage tracks the origins and transformations of data to understand who handles it and where the data comes from. Can be a business lineage or a technical lineage. Being conscious about where data come from, can help to understand and identify the source of wrong data. By eliminating them, data can become more reliable. Data Lineage is necessary to perform these operations in few minutes, especially relying on IBM tools such as Knowledge Catalog.

Is important to set up the architecture mechanism necessary to support this business initiative. In this analysis is crucial to individuate data storage which are shared among different business units in the company to avoid the work repeating which is typical without a good planning. The aim is to create a shared project to serve multiple initiatives within the company and in this way more BU can serve themselves from the same architecture and sources without replicating twice or more the same data. This occur a balance between centralization and decentralization of analyst activities to ensure also a relationship among different analyst teams which are spread all around the company in different departments. Decentralization is a right solution but shouldn't occur in an excessive splintering.

For balancing the centralization with the decentralization, it is necessary to understand the role of analytics, who works and for what purpose, avoiding working on the same job without coordination between different analyst teams and finally find the analytics project that wants to coordinate to build a community (shared reporting and analytics capability helpful to avoid recreating it twice).

Mapping data lineage and flow across operational units ensures a coherent, scalable governance infrastructure tailored to company's multi-channel ecosystem, identifying opportunities for shared data assets across different business units to avoid redundancy and promote collaboration across teams.

IBM Knowledge Catalog integrates data from disparate data sources (third-party solutions, cloud platforms, on-premises systems) and harmonizes disparate data repositories to uphold governance uniformity. A key challenge that we can encounter in this phase is represented by Data Silos and fragmentation: the lack of interoperability between various business units results in disparate data repositories, leading to inconsistencies in data access, redundancy, and difficulties in achieving a unified data view. Bridging these silos is imperative to establish seamless Data Governance and one way to overcome that challenge is the implementation of a Data Integration Mechanisms. Implementing advanced data aggregation and harmonization

techniques to consolidate disparate data sources, ensuring that data silos are effectively eliminated, and interoperability is achieved.

The IT team emphasized the importance of tracking data lineage to understand the transformations from raw data to final reporting. Currently, they use audit systems such as "last modified" and "modified by," as well as control logs on uploads and accesses.

The implementation of IBM Knowledge Catalog will allow for the centralization and automation of this process, tracking not only the transformations but also the origin of the data, the transformation rules, and the update frequency. Furthermore, it will be essential to define the ownership of each table and attribute for more effective governance.

5.1.3 Data Catalog

After the data integration step, an admin can decide to perform immediately an analysis on data without proceeding to catalogue it, but to ensure data homogeneity across company's functional divisions is contingent upon a well-integrated metadata framework:

- Systematic Metadata Cataloguing: constructing an enterprise-wide metadata repository enhances data traceability and utilization. Employ IBM's data integration solutions to ingest metadata into the platform. This approach allows for efficient cataloguing and governance without immediately requiring large-scale data migration. This allows for analysis and governance before full data consolidation.
- Automated Data Classification: Deploying AI-driven classifiers such as Watson APIs facilitates real-time metadata tagging and semantic structuring. Leverage Watson APIs for natural language understanding, machine translation, visual recognition, and speech-to-text services to enhance metadata enrichment and classification.
- Defining Enterprise Data Taxonomy: Establishing standardized business definitions curtails interpretative discrepancies and fortifies governance across retail, inventory, and customer management domains.

By structuring a Data Catalog, organizations ensure that data assets are well-documented, standardized, and readily available for analysis while maintaining compliance with governance policies. The creation of a structured Data Catalog involves the definition of metadata attributes for each table, including source descriptions, update frequency, and data classification.

We proceed to define the glossary of the catalog:

- Categories: allow users to easily affect the security around their governance artifacts. Categories are used to quickly search assets around the organization. So, when we create categories, we must focus on the point of view of the final users that are going to look for assets. We have created a hierarchical structure of the categories for the project. The lite plan of IBM Knowledge Catalog allows the creation of up to 10 categories and thus considering that the lite plan is shared within the company among other users, we must adapt and create only 6 categories, where the main one is “PROJECT WORK TESI CRESCENZI” and other 5 subcategories. We can control the user access for each category, knowing that if a user has access to the primary category, he will have access to the subcategories too.

Category Name	**Description**
CUSTOMER & MARKET SEGMENTATION	Category related to customer data and market segmentation for targeted analysis.
FINANCIAL DATA	Category that includes revenue, profit margins, and cost analysis.
PRODUCT INFORMATION	Category containing product-related metadata, including identification, attributes, and classification.
SALES & TRANSACTIONS	Category containing sales-related metadata, including channels, financial data, and statistical classification.
SUPPLY CHAIN & LOGISTICS	Category related to product distribution and logistics metadata.

Table 11: Subcategories

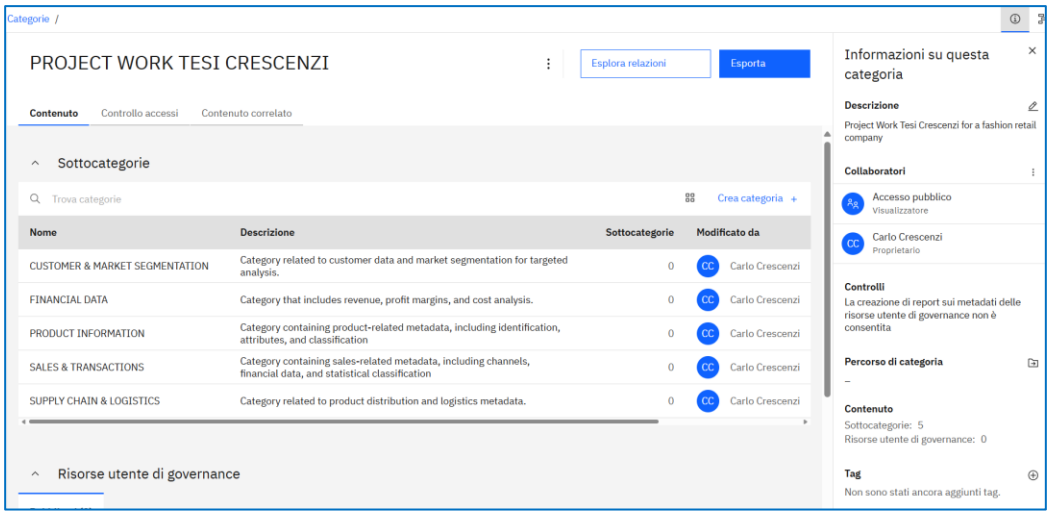


Figure 28: Subcategories

- Business terms: There are two ways to add business terms in the platform. The first one consists in uploading an external csv file, while the second one is performed by using the GUI of the platform specifying name description and the related category. As we mentioned before, business terms define the entities in the domain. In the figure below there are some of the business terms that we created in the platform, each of them is associated to a primary category. In each row is also specified the last user who has modified it, and the modification date.

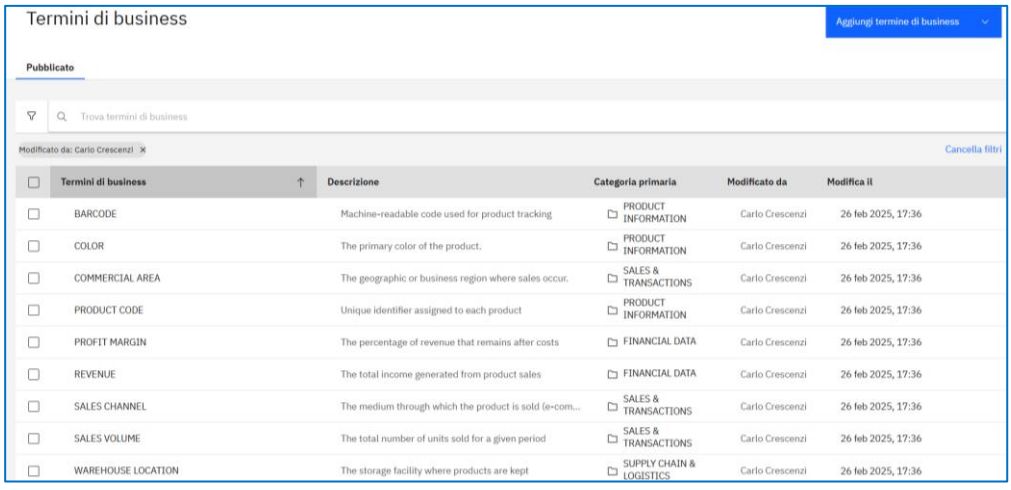


Figure 29: Business terms

- Classifications: they are structured labels used to categorize data based on sensitivity, compliance, business domain, or data type. Classifications can be applied manually or automatically using Machine Learning. We created the classification that will be linked to the data of the table. As for the business

terms, each classification specifies the associated category, the date and the creator.

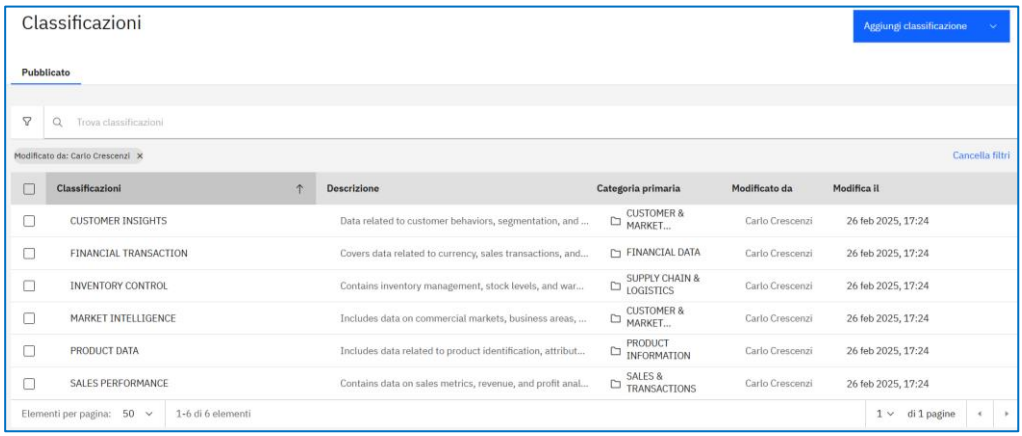


Figure 30: Classifications

- Data classes: describe the format of the data. We have created only these two additional data classes because in the platform where already exist a sufficient number and typologies of data classes exhaustive for defining also our data classes of the project.

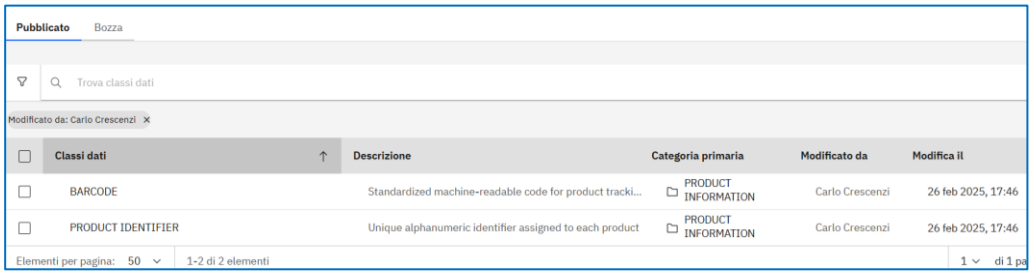


Figure 31: Data Classes

In the Cataloguing phase is crucial also establishing primary keys and relationships between tables to ensure seamless data integration. Key insights from the dataset analysis:

- Primary keys and relationships among tables
    - Each dimensional table contains a unique primary key (e.g., PRODUCT_PK for Product, CUSTOMER_PK for Customer).
    - The F_SO_TRANSACTION fact table contains foreign keys linking to dimension tables, establishing relationships that support analytical reporting.
    - The Time dimension tables (Time Day, Time Month, Time Year) provide a hierarchical structure for temporal analysis.

- Necessity for automated metadata management
    - To improve usability, each dataset should be catalogued with descriptive labels and categorized based on business domains.
    - IBM Knowledge Catalog can streamline this process by automating metadata ingestion and enabling advanced searchability of the data assets.

5.2 Protecting the Data

The key challenge is balancing accessibility with compliance and security because fashion companies must balance the need for widespread data accessibility with stringent regulatory frameworks such as GDPR, CCPA, and other privacy laws that dictate how data should be handled. Achieving this equilibrium necessitates utilizing automated policy enforcement mechanisms, data masking, and access controls to mitigate security risks while ensuring compliance with industry-specific regulations and a regularly revisiting governance policies to align with new regulations, emerging technologies, and evolving business priorities. Sales data contains sensitive information, such as customer data and confidential pricing. It is crucial to ensure compliance with GDPR, CCPA, and tax regulations. Data access is restricted based on the user's role and Business Unit. Strengthening data security and privacy requires data anonymization or masking mechanisms at the visualization level.

5.2.1 Data Compliance

Sales data contains sensitive information, including customer details, confidential pricing structures, and fiscal data, all of which must comply with GDPR, CCPA, and tax regulations. Ensuring compliance requires a structured governance framework that regulates access control, data masking, auditing mechanisms, and classification policies.

To enforce these regulations, the organization implements role-based access control, which limits data visibility based on user roles and business units. There are two levels of access classification: a region/global distinction and restrictions based on the user's role and department. Access requests are formally managed through an IT ticketing system or email approval workflow, ensuring that only authorized personnel can retrieve specific datasets.

Sensitive data, such as customer records and pricing information, can be masked dynamically depending on the level of analysis required and the user's role. IBM Knowledge Catalog supports this approach by implementing data masking rules at the data visualization level. For example, personally identifiable information such as customer names, addresses, and emails are masked for non-finance users, while pricing data and discount structures are obfuscated for unauthorized personnel. Additionally, the organization has established audit mechanisms that track modifications to sales data using logs of last modifications and user activity monitoring.

The compliance strategy also involves data classification and lineage tracking. By cataloguing and classifying datasets in IBM Knowledge Catalog, the organization ensures that all transformations, data sources, and processing steps are transparent and auditable. This guarantees that data used for decision-making aligns with internal governance policies and external regulatory requirements.

5.2.2 Data Security

A structured governance framework ensures role clarity and accountability. The IT team confirmed that sensitive data can be masked or not, depending on the degree of analysis and the user's role, and that audit systems are in place to monitor data access and modifications. Access to sales data is subject to restrictions based on the user's role and the Business Unit they belong to. There are two access classifications: at the region/global level and based on the role and BU. Data access requests are managed through IT tickets or email.

Key actions include:

- Designation of Core Governance Roles: Appointing data stewards, governance officers, and executive sponsors to oversee governance execution and enforcement and to foster accountability, regulatory coherence and responsibilities for data management within the project team. Collaborating with cross-functional teams, including IT, compliance, marketing, and supply chain departments, to continuously optimize governance strategies and adapt to emerging challenges.
- Cross-Departmental Integration: Developing governance policies that incorporate inputs from business, IT, and compliance teams, ensuring a holistic and balanced approach.
- Governance Charter Development: Defining governance principles, accountability matrices, decision-making hierarchies, and escalation procedures.

Ensuring secure access to sales data while maintaining usability for authorized users is a crucial aspect of Data Governance. To this end, a structured access request process has been analysed to regulate data visibility and prevent unauthorized modifications.

Process for Requesting Access to Sales Data:

1. Request Submission
    - Employees requiring access to sales data must submit a formal request via the IT ticketing system or by email to the IT governance team.
    - The request must specify the required level of access (read-only or full access) and justify the business need.
2. Approval Workflow
    - The request is reviewed by the Regional IT Manager, who verifies compliance with company policies.
    - If the request pertains to sensitive data, additional approval is required from the Data Governance Officer.
    - The approval process typically takes 24 to 48 hours.
3. Access Levels and Role-Based Permissions
    - Read-Only Access: Granted to Business Intelligence Analysts and Planning Specialists for analytical purposes.
    - Limited Modification Access: Sales Managers and Finance Team members receive partial modification permissions.
    - Full Access: Restricted to executive-level personnel and IT administrators managing the system.

The access request and approval workflow described above was not implemented in IBM Knowledge Catalog due to technical constraints and platform limitations.

IBM Knowledge Catalog does not provide a manual approval system for data access requests. Unlike traditional IT governance models where access is granted through ticketing systems or direct approvals, the platform operates through predefined role-based policies.

The Lite Plan does not include advanced approval workflows. While higher-tier IBM Cloud plans may offer integrations with IT ticketing and approval systems, our implementation was constrained by the features available in the Lite version. Governance and access control are fully automated. Instead of requiring manual approval from an IT Governance Officer, data

access is controlled through predefined policies and governance rules, ensuring compliance with security regulations while reducing administrative overhead.

To enforce secure access to sales data, IBM Knowledge Catalog has implemented an automated role-based access model, eliminating the need for manual approval workflows. Rather than relying on individual requests, data visibility and modification permissions are automatically assigned based on governance policies and user roles:

- Full Access: Granted to executive-level personnel and IT administrators managing the system. They have complete access to all financial, operational, and customer data.
- Limited Modification Access: Assigned to Sales Managers and Finance Team members, who can view and edit sales data but cannot access all financial details.
- Read-Only Access: Business Intelligence Analysts and Planning Specialists can view aggregated sales trends and customer behaviour but cannot modify data.
- Restricted Access: Interns and external consultants can only access high-level business trends and aggregated KPIs, with no visibility into individual transactions.

By leveraging automated access control, IBM Knowledge Catalog ensures that data security and regulatory compliance are maintained while allowing authorized stakeholders to work with relevant datasets. Unlike traditional access request processes that require IT governance intervention, access permissions are automatically enforced based on predefined governance rules, eliminating the need for manual approval workflows.

To enhance data privacy and ensure compliance with regulatory frameworks such as GDPR and CCPA, the company has implemented robust data masking policies. These policies restrict the visibility of sensitive information based on user roles, preventing unauthorized access to critical data fields. Implementation of Data Masking:

- Customer Data: Personally Identifiable Information (PII) such as customer names, addresses, and emails are masked for non-finance users. Example:
    - Original Data: "John Doe, 123 Main Street, johndoe@email.com"
    - Masked Data: "J*** D**, *** Main St, j********@****.com"

- Pricing Information: Confidential pricing and discount structures are obfuscated for unauthorized users.
    - Original Data: "Product A – $199.99, Discount: 15%"
    - Masked Data: "Product A – $XXX.XX, Discount: XX%"
- Order History: Specific transactional details are redacted for non-authorized personnel.

Governance policies define high-level security frameworks that regulate data access, protection, and compliance. These policies provide the foundation for enforcing security measures and ensuring that data handling aligns with regulatory requirements.

- Role-Based Data Access & Permissions Policy: This policy ensures that access to sales and financial data is strictly role-based, preventing unauthorized modifications while maintaining visibility for authorized users. It eliminates the need for manual access approvals by automating permissions through predefined governance rules. The policy helps in implementing structured access levels, ensuring that only users with the appropriate role can access or modify data.
- Data Masking & Privacy Compliance Policy: This policy enforces data masking techniques to protect personally identifiable information (PII) and financial data, ensuring compliance with privacy laws such as GDPR and CCPA. Sensitive customer details, such as names, addresses, emails, pricing, and discount information, are dynamically masked for unauthorized users. By ensuring that sensitive data is only visible to those who need it, this policy mitigates risks of data breaches and unauthorized exposure.
- Data Access Request & Approval Workflow Policy: Initially designed to support a manual access request system, this policy was adapted to IBM Knowledge Catalog's automated framework. Instead of requiring IT approval workflows, access control is now automatically enforced based on predefined user roles and governance rules. This approach eliminates administrative overhead while maintaining strict access control mechanisms.
- Data Classification & Protection Policy: This policy defines classification levels for company data, ensuring that sensitive information is properly categorized and protected. Each data asset is assigned a classification, such as Sales &

Transactions, Customer & Market Segmentation, or Financial Data, which determines how it is secured and accessed. This policy ensures that security controls, such as data masking or access restrictions, are applied automatically based on the asset's classification.

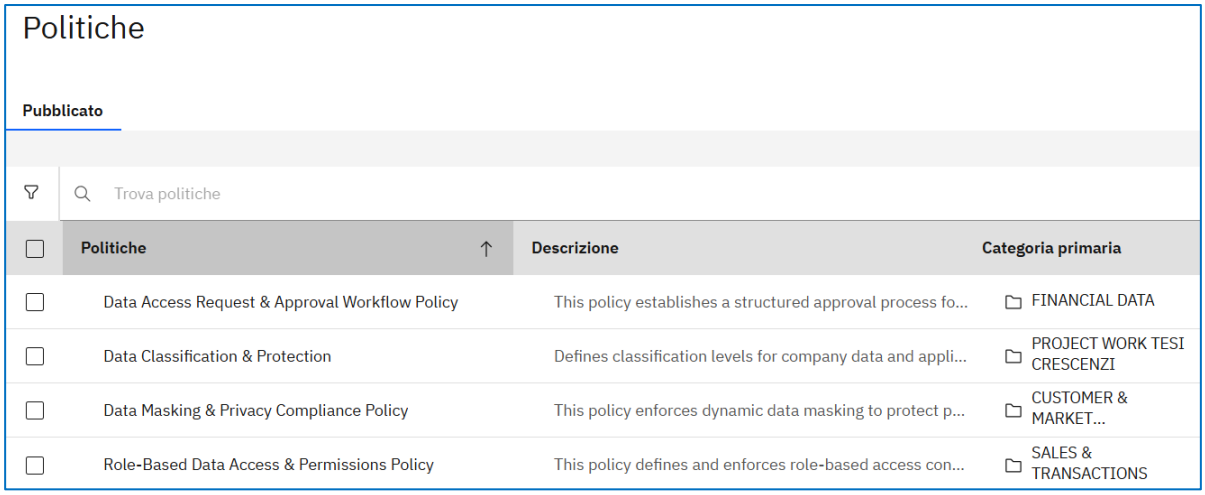


Figure 32: Policies

Governance rules enforce the governance policies by specifying how data access, security, and compliance should be managed. These rules act as the operational backbone of IBM Knowledge Catalog's security framework.

- Enforce Role-Based Access Control (RBAC): This rule defines and enforces structured access permissions based on user roles. Instead of requiring manual access approval workflows, users are automatically assigned predefined access levels, such as Full Access, Limited Modification, Read-Only Access, or Restricted Access. This approach ensures that only authorized personnel can access and modify sensitive financial and sales data, reducing security risks while maintaining usability.
- Enforce Data Masking for Privacy Compliance: To protect personal and financial data, this rule applies dynamic data masking for unauthorized users. Depending on their role, employees either see masked data or complete details. This rule is crucial for ensuring GDPR and CCPA compliance, particularly in industries handling customer-sensitive information.

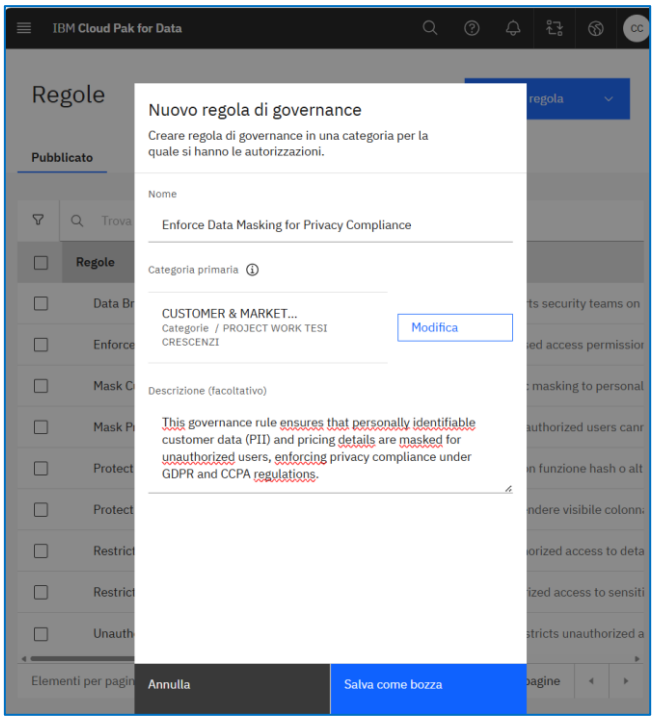


Figure 33: Enforce Data Masking rule

- Enforce Data Access Approval Process: While IBM Knowledge Catalog does not support manual approval workflows, this rule automates access permissions based on predefined policies. Instead of requiring users to submit access requests, permissions are granted or denied automatically, based on their assigned role and access level. This rule eliminates the need for manual intervention while ensuring that access is tightly regulated and compliant with governance policies.

Data protection rules enforce security controls such as restricting access, masking data, and monitoring unauthorized attempts. These rules are designed to proactively prevent security threats and ensure that only authorized personnel can access critical business information.

- Restrict Access to Sales Transactions: This rule prevents unauthorized users from accessing detailed sales transaction data. It ensures that only Finance Team, Sales Managers, and Executives can access individual transaction records, while all other users see only aggregated sales trends. This restriction helps prevent unauthorized data modifications and ensures that transactional integrity is maintained.

- Mask Customer Data: This rule ensures that personally identifiable information (PII) such as customer names, emails, and addresses is masked for unauthorized

users. Only employees with financial clearance (e.g., Finance Team and Executives) can access unmasked customer data. This rule significantly enhances data privacy and reduces exposure risks in case of a security breach.

- Mask Pricing and Discount Data: Confidential pricing and discount structures are protected by this rule, ensuring that unauthorized users cannot see sensitive financial details. Only the Finance Team and Executives have access to unmasked pricing and discount information, while all other users see obfuscated values. This rule ensures that pricing confidentiality is maintained while allowing general visibility for analytical purposes.
- Data Breach Alert: This rule enhances security monitoring by detecting unauthorized access attempts and generating alerts for review. By tracking suspicious login activities, IBM Knowledge Catalog administrators can proactively identify potential data breaches and take preventive action before sensitive data is compromised.
- Unauthorized Access Detection: This rule monitors and logs all unauthorized access attempts to protected data. By tracking access logs and failed attempts, the system can flag potential security threats and ensure that unauthorized personnel are unable to extract sensitive information.

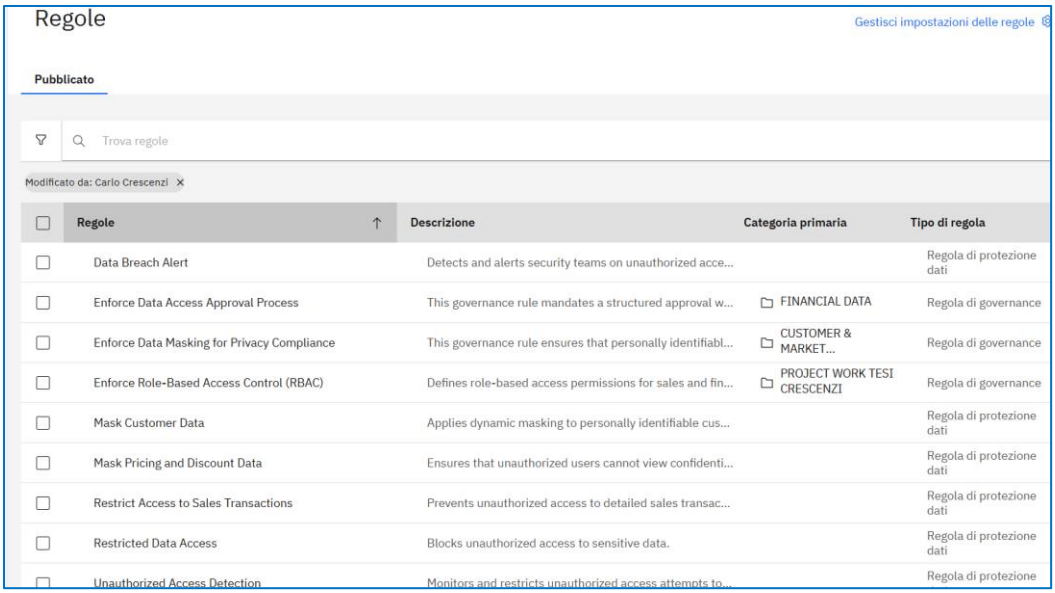


Figure 34: Rules

This governance framework ensures that IBM Knowledge Catalog is fully aligned with data security regulations, providing a structured, efficient, and scalable solution for managing sales and financial data access.

In our IBM Knowledge Catalog implementation, we did not create custom roles due to lack of administrative permissions. The IBM Knowledge Catalog Lite plan does not provide the necessary privileges to define or manage user roles at the system level. Role creation typically requires administrative rights that are only available in higher-tier IBM Cloud plans or for account administrators.

However, instead of defining formal roles, we structured access control using user groups. These groups function similarly to roles, allowing us to implement Role-Based Access Control (RBAC) effectively within the catalog. By associating user groups with predefined governance rules and policies, we ensured that data access restrictions were properly enforced.

- Executive Board: This group consists of senior executives and key decision-makers who require full access to sales and financial data. They can view and modify all financial, operational, and customer data without restrictions. Their permissions include modifying sales reports, approving pricing strategies, and overseeing governance policies to ensure compliance with business objectives. Since they handle critical business decisions, their access level is the highest in the organization.
- Finance Team: The Finance Team requires access to all sales data, including cost and margin details, to perform financial analysis. They have limited modification rights, allowing them to create financial projections, assess discount effectiveness, and analyse revenue trends. While they can view detailed financial records, they do not have full modification privileges to prevent accidental data alterations.
- Sales Managers: Sales Managers need limited modification access to sales data to monitor regional performance and adjust pricing strategies. They can view sales data segmented by region, evaluate promotional success, and track sales performance. Their role involves strategic execution rather than financial analysis, so they do not have visibility into full revenue or customer transaction details.

- Business Intelligence (BI) Analysts: BI Analysts work primarily with aggregated sales trends and customer behaviour data to perform forecasting and generate business insights. They have read-only access to relevant datasets but cannot modify or access transactional details. Their access ensures they can provide data-driven insights without compromising the integrity of the raw data.
- Marketing Team: The Marketing Team uses customer segmentation data to design targeted marketing campaigns. Since they do not require access to sensitive financial or transactional data, their visibility is limited to anonymized customer insights. They can analyse market trends, customer behaviour, and overall sales performance, but personally identifiable information (PII) remains masked to comply with data privacy regulations.
- Interns / External Consultants: This group has the lowest level of access in the system. Interns and external consultants can only view high-level business trends and aggregated key performance indicators (KPIs) without seeing individual transactions, customer data, or detailed sales records. Their access is strictly restricted to ensure that confidential company information remains protected.

Although custom roles could not be created, the access groups defined in IBM Knowledge Catalog function as role equivalents and provide structured Role-Based Access Control (RBAC):

- Instead of manually approving access requests, permissions are assigned at the group level.
- Data masking, transaction access restrictions, and security alerts are automatically applied based on the user's group.
- Governance rules enforce strict data access control, ensuring that each group only accesses relevant datasets without compromising security.

This structured access model ensures that even without formal role creation, data governance remains secure and well-organized in IBM Knowledge Catalog.

User Role	Access Level	Data Visibility	Example of Permitted Actions
Executive Board	Full Access	All financial, operational, and customer data	Modify sales reports, approve pricing strategies
Finance Team	Limited Modification	Full sales data (with cost and margin details)	Create financial projections, assess discount effectiveness
Sales Managers	Limited Modification	Sales data with regional segmentation	Adjust pricing models, evaluate promotional success
BI Analysts	Read-Only Access	Aggregated sales trends, customer behaviour	Perform trend analysis, generate insights
Marketing Team	Read-Only Access	Anonymized customer segmentation data	Plan targeted marketing campaigns
Interns / External Consultants	Restricted Access	Aggregated KPIs without individual transaction details	Analyse high-level business trends

Table 12: Group-Based Access Control (GBAC) Model

5.2.3 Data Lifecycle

Answers the following questions: “How long do I need to keep that data? Which mechanism must I use to store that data?”. It is possible to store as much data as needed, but it is better not to overspend that capability. In summary, it defines retention policies and storage mechanisms. The data lifecycle defines how long sales data should be stored, archived, or deleted in alignment with business, compliance, and operational needs. While the system allows data to be retained indefinitely, an effective strategy must balance storage optimization, cost efficiency, and regulatory compliance.

The organization follows a structured data retention policy, ensuring that transactional sales data is kept for at least five years to meet tax requirements before being archived or deleted. This policy prevents unnecessary data accumulation while maintaining historical records for compliance and analysis. Data updates occur three times a day, synchronized across different time zones (Americas, EMEA, and Asia), ensuring that the most recent transactions are continuously integrated into the system.

To enhance storage efficiency, the company differentiates between frequently accessed data and historical archives. Frequently used datasets remain in primary storage, while older data is moved to cold storage or external archives. This approach prevents excessive costs and improves system performance. Additionally, automated validation checks are implemented to maintain data integrity, ensuring that missing or inconsistent records are flagged and corrected in real-time.

IBM Knowledge Catalog plays a crucial role in data lifecycle governance, enabling organizations to define expiration rules that automatically archive or delete data based on predefined conditions. This ensures that outdated data is not only removed from active storage but is also properly documented and tracked before deletion, maintaining transparency and compliance with regulatory frameworks.

By implementing a robust data lifecycle strategy, the company ensures that only relevant and necessary data is retained, optimizing resources while maintaining compliance with legal and business requirements.

5.2.4 Data Integration and Cataloguing

A critical component of the Data Governance initiative is the ingestion and cataloguing of data into IBM Knowledge Catalog, ensuring a structured and compliant data management process.

The IBM Cloud Pak platform allows the integration of data into IBM Knowledge Catalog through two ingestion methods. The first one is a Database Connection: Directly linking IBM Knowledge Catalog to a structured relational database to enable real-time synchronization and metadata extraction. This method allows for automated updates thus reducing the manual intervention. Having a database connection means also that we could use structured query executions for enhanced data discoverability. In this project we don't use this option, which requires a connection to the original database.

The second method involves a Local Excel Upload: Ingesting data from Excel files, ensuring accessibility for non-technical users while enabling metadata enrichment. This way provides flexibility in data ingestion for non-relational sources and facilitates metadata annotation. In this project, we perform only this option to ingest data into the platform, due to legal enforcement that deny the possibility to connect to the original data source of the company. The tables that we are going to use contains the original data, but they are replicated from the sources.

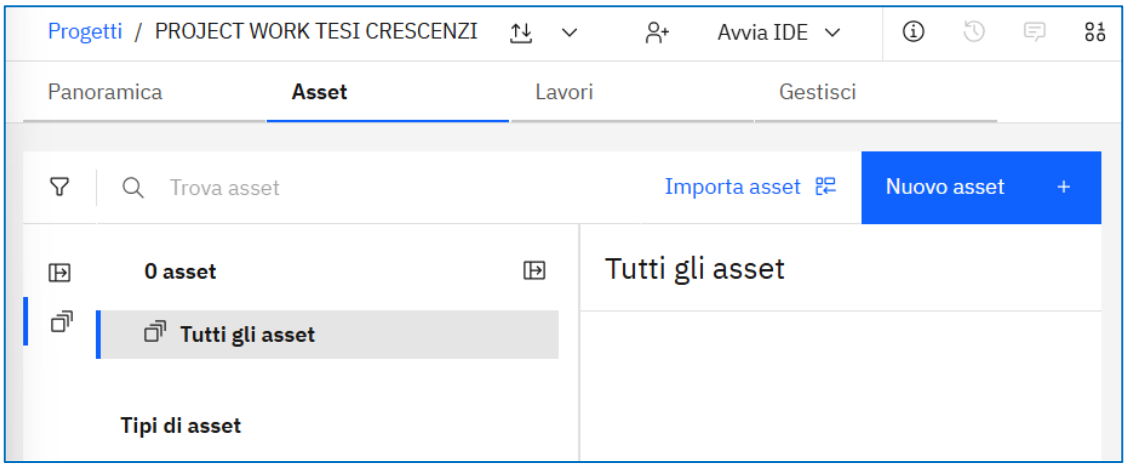


Figure 35: Asset panel

"Import Asset" option uploads asset that already exist, while "New Asset" creates an asset. In our case we already have assets to import, so we select the first option.

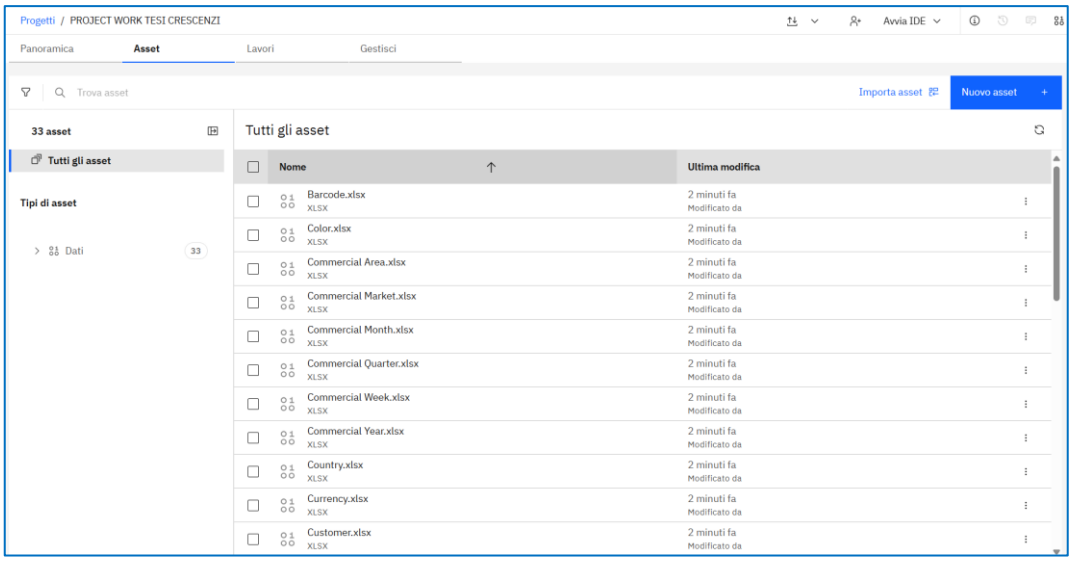


Figure 36: Assets imported

These are the assets that we have imported in the platform. All of them are excel tables. The platform offers the possibility to change the name of the assets and moreover to publish them in a catalog or to go to the data refinery step.

Below we show the preview of the data. From the panels on the top, we can check the profile of data that creates statistics on columns. Knowledge Catalog can recognize the data type and creates the statistics more appropriable for a certain data type.

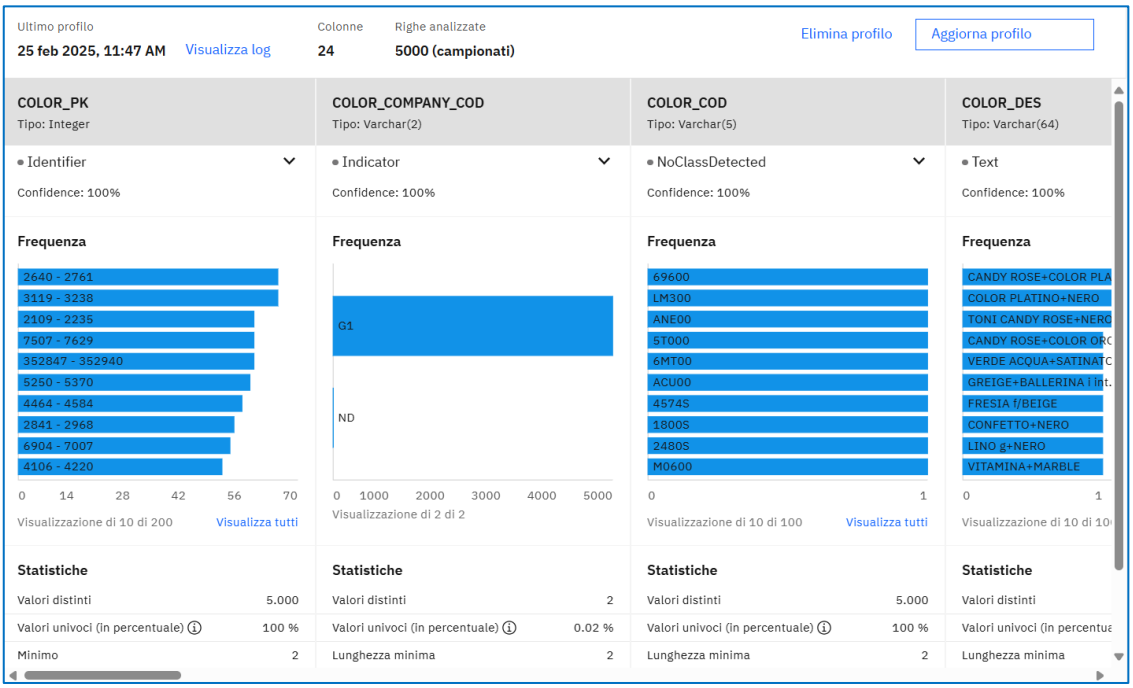


Figure 37: Profile panel

In the Visualization panel, the user can choose among different models to visualize the distribution of data in a column. In this example we selected the COLOR_COD column. The tool suggests the most appropriable graphs to visualize the column selected. If we need, we can save and download the graphs adding a title and a description. A common operation that we can always do in every situation is the possibility to detect and describe an operation, like in this case. This emphasizes the objective of IBM Knowledge Catalog, which aim to categorize and label all data and track operation to gain affordability to data and operations.

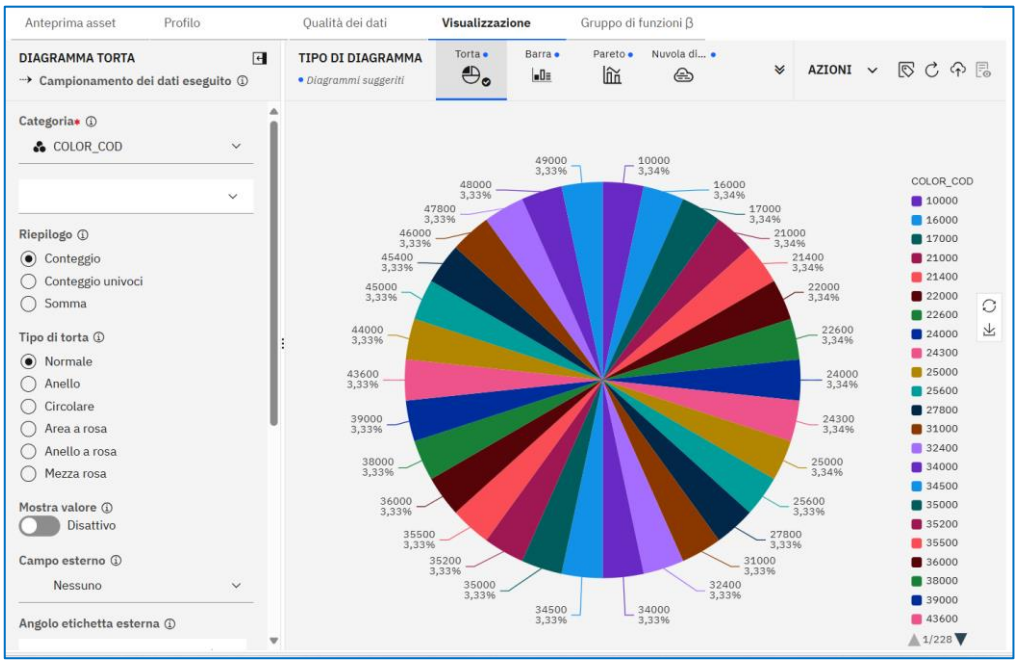


Figure 38: Visualization panel

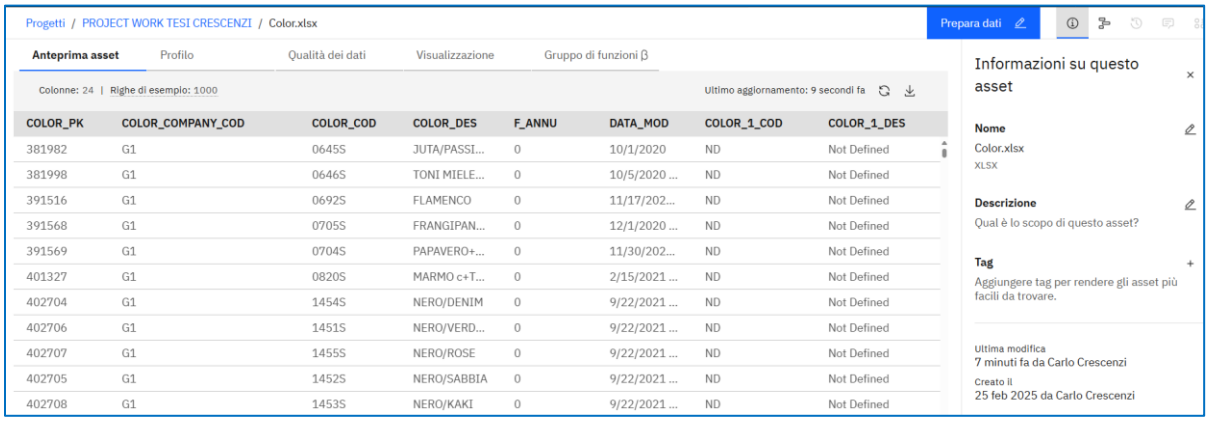


Figure 39: Asset preview panel

For example, if we would like to prepare data for the COLOR table, this is the scenario that is presented. This is the scenario for metadata enrichment:

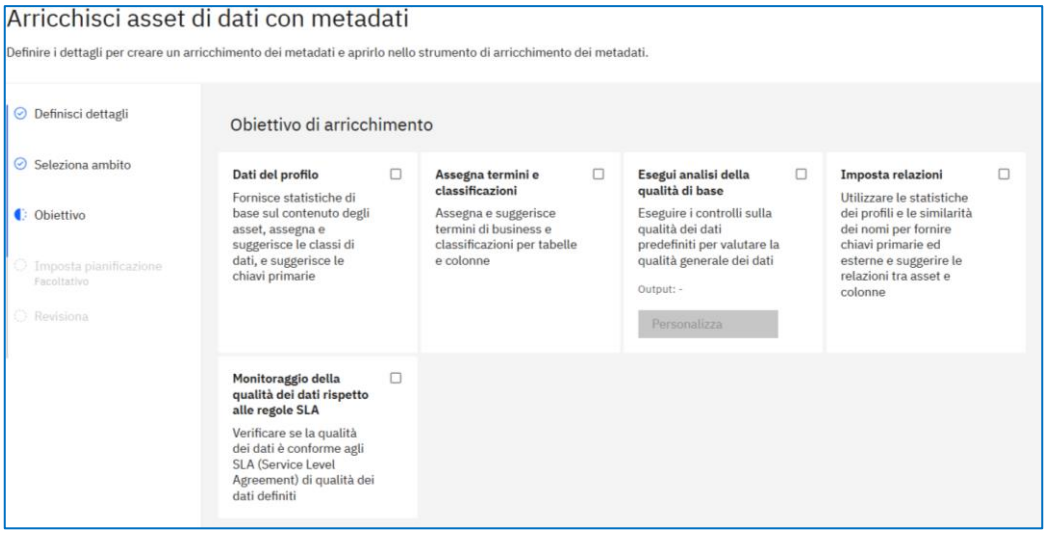


Figure 40: Metadata enrichment panel

Upon ingestion, IBM Knowledge Catalog will automatically extract and classify metadata, applying business terms and data classes, enhance searchability through intelligent data discovery mechanisms and ensure security compliance, restricting access to sensitive attributes based on governance policies.

The combination of database ingestion and local file uploads ensures a hybrid approach to Data Governance, allowing seamless integration while maintaining flexibility in data onboarding processes.

In order to further refine sales analytics and improve strategic decision-making, the company has planned an expansion of its Sales Data Mart by incorporating additional data sources. This initiative aims to enhance forecasting accuracy, streamline order planning, and optimize inventory distribution across different regions. The newly integrated sources will include:

- Merchandise Planning Data: This dataset will provide insights into the allocation and movement of products within retail stores, enabling better demand forecasting and reducing stockouts.
- Advanced Forecasting Applications: Predictive modelling tools leveraging historical sales data and machine learning algorithms will be integrated to improve demand prediction and ensure a proactive supply chain strategy.
- Regional Sales Trends Reports: Detailed reports will be generated at a regional level, highlighting the top-performing product categories and market demand fluctuations.

- Customer Purchase Behaviour Insights: Data sourced from loyalty programs and digital interactions will be incorporated to enhance personalization strategies and targeted marketing.

By incorporating these new data sources, the company will significantly enhance its analytical capabilities, ensuring that sales forecasts and stock planning are more precise and aligned with actual demand. Additionally, integrating predictive analytics into the Sales Data Mart will provide a proactive approach to business strategy, reducing reliance on reactive decision-making.

After having integrated the assets in the platform, we proceed to load them in the main catalog where we can assign classifications and business terms. IBM Knowledge Catalog can make autoclassification of the assets, but due to the low number of assets that we must import, we compute a manual classification and assigning of the business terms for each asset.

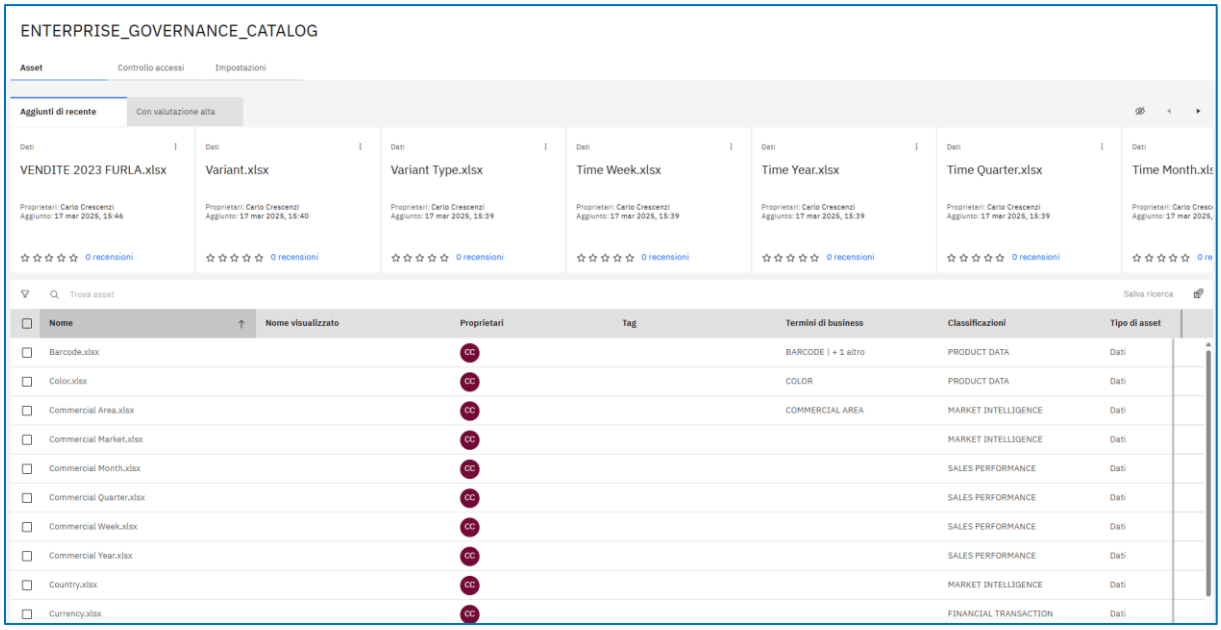


Figure 41: Catalog assets

An IBM Knowledge Catalog acts as a centralized repository for managing, organizing, and governing enterprise data assets, ensuring accessibility, compliance, and security while enabling efficient data discovery. By integrating metadata enrichment, automated classification, and business term assignments, the catalog provides a structured approach to data standardization and governance, reducing inconsistencies across departments. One of its most critical functionalities is data lineage tracking, which allows organizations to trace the origin, transformations, and flow of data across systems, ensuring auditability and regulatory compliance. Additionally, the catalog supports data quality management by identifying errors,

inconsistencies, and missing values, ensuring that business intelligence and analytics are based on reliable and validated datasets.

A fundamental distinction in IBM Knowledge Catalog is the difference between a catalog and a project. A catalog is designed for governance and enterprise-wide data management, providing a structured framework where data assets are curated, classified, and secured according to governance policies. It is intended for long-term data organization and controlled access, ensuring that different business units can discover and utilize standardized datasets while maintaining compliance.

In contrast, a project serves as a collaborative workspace where users can actively work with data, develop models, and perform analyses. Projects are temporary and task-focused, allowing teams to ingest, clean, and manipulate datasets without altering the core governance structure maintained in the catalog. While the catalog enforces data security, lineage, and standardization, a project provides flexibility for innovation and development, with users having more freedom to modify data and apply transformations. The synergy between catalogues and projects ensures that organizations can maintain a balance between structured governance and agile data exploration, enabling teams to efficiently leverage trusted data assets for analytics and decision-making.

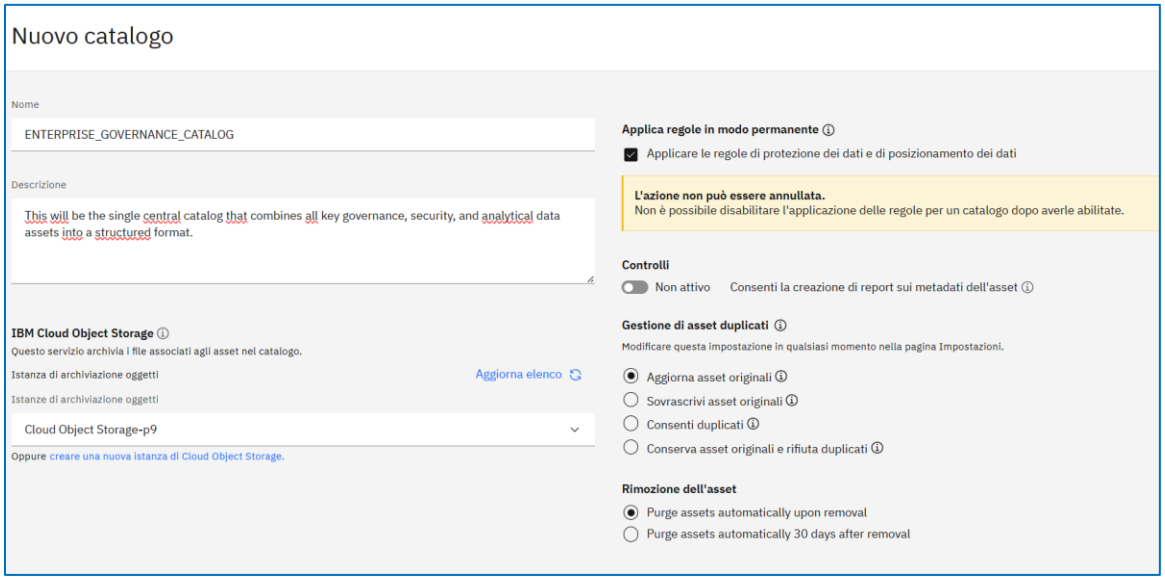


Figure 42: Catalog configuration

5.2.5 Data Refinery

The next step, after having performed the data integration one, is Data Refinery. In other cases, analyst can decide to move directly to data deployment with Watson Studio, but in most cases Data Refinery can be a necessary step to explore, shape and operationalize the data. Refining and structuring organizational data are essential for compliance and analytical robustness:

- Advanced Data Profiling Techniques: Connect Data Refinery to various data sources (on-premises, on-cloud, IBM, non-IBM) to facilitate data visualization and statistical analysis for conducting heuristic and statistical data analyses to identify systemic inconsistencies and anomalies.
- Precision-Driven Data Cleansing: Utilize built-in statistical techniques to detect data types, anomalies, and sensitive fields.
- Automated Data Processing Pipelines: Establish scheduled data flows to ensure repeatability, scalability, and accountability in data refinement processes. Deploying profiling tools to assess and rectify inconsistencies, ensuring data integrity before further processing and analysis. As companies expand their operations and integrate emerging technologies, governance frameworks must be scalable and adaptable to changing business landscapes without disrupting existing operations. Establishing automated workflow mechanisms to track policy adherence and continuously monitor Data Governance processes, ensure a sustained compliance and efficiency.
- Anomaly Detection and Outlier Analysis: Leveraging predictive analytics to detect deviations, facilitating proactive governance adjustments.

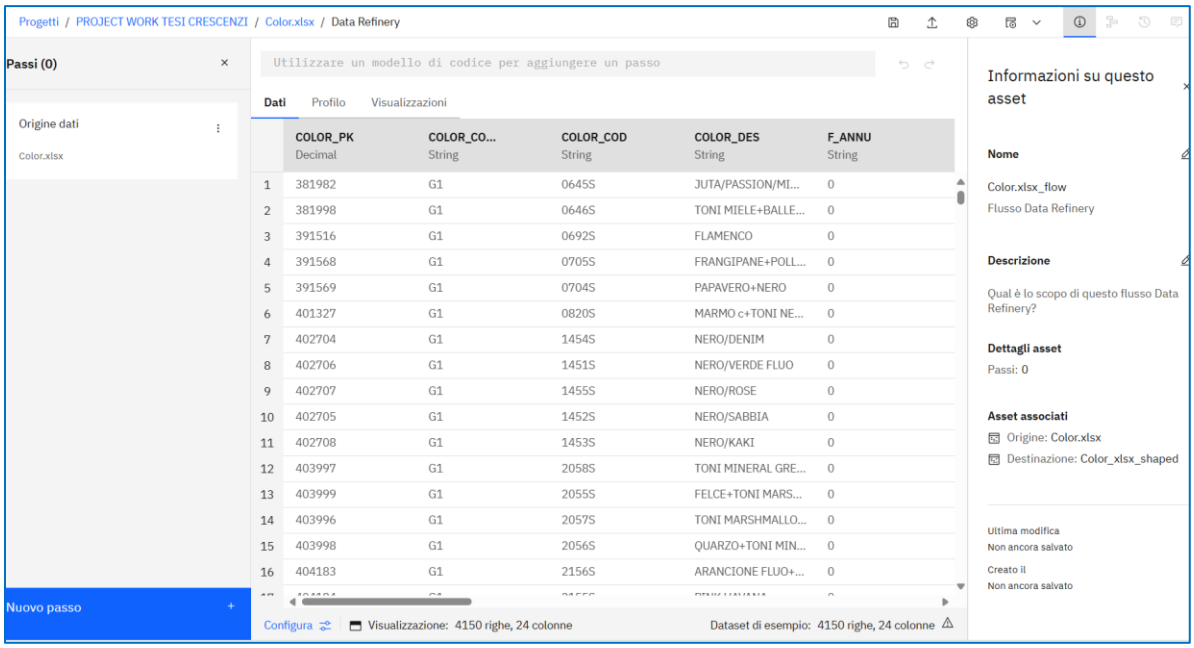


Figure 43: Data Refinery Scenario

This is the Data Refinery scenario. On the left there is the section from which it is possible to add Data Refinery steps. Once the steps are all defined, they can be saved as "work" and can be applied on the same asset with a certain time frequency. For example, in the COLOR table the last column doesn't provide any additional information to the table because all the rows contain the same value as we can check from the visualization or profile panel. So, we would choose to delete the last column every time that the asset is imported in the project. Selecting "add step", then we choose the "delete column" option. As we previously said, a Data Refinery flow execution can be scheduled and applied with a certain frequency. The Data Refinery flow is identified by a name and a description. In the saving phase there is a resume of the assets involved and the scheduling information. Once the flow is created, in the "asset" panel there will be the input asset, the Data Refinery flow and the output asset. All the workflows are listed in the "work" section of the project.

Figure 44: Input asset, workflow and output asset

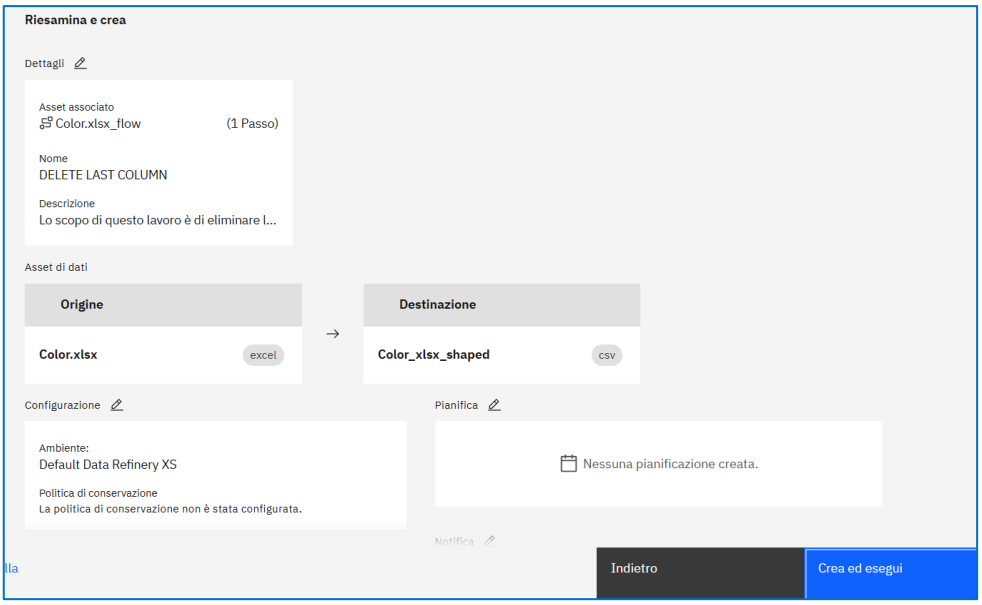


Figure 45: Flow saving panel

Once the flow is performed, we can view the log to check whether something went wrong. Furthermore, if necessary, we can modify the flow adding or deleting steps. The Data Refinery flow is listed in the asset panel. The Data Refinery steps are particularly useful when the assets present anomalies. Through this process is possible to define clear rules that allow users to operate with affordable data.

5.2.6 Integration of Machine Learning for Continuous Improvement

Utilizing Machine Learning algorithms to automate anomaly detection, trend analysis, and predictive governance modelling, thereby reducing manual intervention and enhancing governance efficiency. The growing reliance on AI-driven insights within the fashion industry requires the development of governance models that ensure ethical AI usage, bias mitigation, and transparency in automated decision-making processes. In order to be compliant to these standards is necessary to leverage on AI-powered tools to enrich metadata, improve data lineage tracking, and enhance searchability, ultimately fostering a more efficient and organized Data Governance system.

The company leverages Machine Learning for:

- Anomaly Detection: Identifying irregularities in sales trends and forecasting models.
- Automated Data Validation: Streamlining quality assurance processes and reducing manual intervention.

- Predictive Analytics:
    - Demand Forecasting: Anticipating future stock requirements based on historical sales.
    - Promotion Optimization: Identifying the best timing and pricing strategies for promotions.
    - Customer Retention Analysis: Evaluating factors influencing repeat purchases and customer loyalty.
- Historical sales data and special edition product reports are utilized to enhance data refinement.
- Automated error alerts notify data stewards of missing transactions, allowing immediate corrective actions.
- Advanced data profiling techniques help detect outliers and patterns that influence sales trends.

5.3 Curate the Data

5.3.1 Data Quality Management

A significant challenge in the industry is the absence of standardized data collection, storage, and reporting practices. There is a pressing need to establish rigorous protocols that ensure data accuracy, completeness, and consistency across heterogeneous data sources while aligning with industry-wide standards. A possible solution can be represented by Data Quality Management that consists in employing systematic profiling, validation, and cleansing methodologies to enhance data reliability and integrity, facilitating more accurate and consistent decision-making. In order to sustain that Quality Data Management in the long run is necessary to ensure that governance frameworks are synchronized with enterprise growth trajectories, market dynamics, and technological advancements. It ensures the right data is used for each business initiative.

The sales data errors that occur most frequently are missing data and inconsistency between systems. The company systems that feed the Sales Data Mart include POS, e-commerce, ERP, and APEX. The data in the Data Mart is updated 3 times a day, dividing the flow into zones: Americas, EMEA, Asia. The quality controls implemented on sales data include reference checks, alerts for missing sales, process standardization, control over individual values, and direct customer communication management. Data management processes that could be further automated include:

- Automated validation includes reference checks and direct customer communication in case of data anomalies.
- Sales transaction validation is standardized across regions.
- Data cleansing processes use ML algorithms to identify inconsistencies and improve overall data accuracy.

To maintain data consistency and accuracy, the Sales Data Mart is updated three times per day to align with regional business operations across different time zones (Americas, EMEA, Asia). Each ingestion process follows a structured pipeline that ensures real-time availability of the most recent sales data:

- Point of Sale (POS) Systems: Captures in-store transactions in real-time.

- E-commerce Platforms: Records online sales and customer interactions.
- Enterprise Resource Planning (ERP) System: Manages financial transactions, invoices, and cost calculations.
- APEX System: Aggregates sales forecasts and inventory planning data.

5.4 CONCLUSIONS

Implementing a Data Governance framework with IBM Knowledge Catalog requires a well-calibrated approach that balances centralization and decentralization, as demonstrated by the case study analysis. It is essential to find the right balance between centralization and decentralization, while promoting effective communication between the various teams. Data Governance must, therefore, speed up analysis, preventing teams from duplicating the same efforts.

The optimal Data Governance model depends on factors such as company size, industry regulations, and operational complexity. A hybrid model, where governance policies are centralized but data ownership is decentralized, is often the most effective.

- Centralized Data Governance: Best suited for companies with strict regulatory requirements or those handling sensitive data (e.g., financial institutions, healthcare). This model ensures uniform security policies, compliance, and data consistency but can slow down decision-making.
- Decentralized Data Governance: Works well for large enterprises with multiple business units, allowing teams to manage their own data assets while adhering to global governance standards. This enables faster decision-making and flexibility but requires strong oversight mechanisms to avoid data silos and inconsistencies.
- Hybrid Approach: A governance structure where policies, security, and compliance measures are set at the corporate level, while data stewardship and operational management are decentralized. This allows business units to customize data processes based on their specific needs, while still ensuring alignment with enterprise-wide governance policies.

Several companies have successfully implemented hybrid Data Governance models that balance centralization and decentralization.

- Master Data Management (MDM) Best Practices
    - Companies like Unilever and Nestlé have centralized their data policies and security measures but decentralized data ownership to regional

teams. This ensures data consistency across different markets while allowing local teams to operate autonomously.

    - IBM and Microsoft use AI-powered governance tools to ensure that data remains compliant with global policies while allowing for flexibility in local decision-making.

- Self-Service Analytics with Centralized Oversight
    - Many organizations, including Google and Amazon, allow teams to access and analyse data independently while enforcing strict access control and security policies at a corporate level. This approach enables data-driven decision-making without compromising security or compliance.
- Decentralized Data Stewardship for Faster Insights
    - Some companies assign data stewards to individual business units while maintaining a central governance team that ensures compliance and quality control. This approach has been effective in banks and retail chains, where different branches operate independently but must adhere to corporate data policies.

For the company analysed in this case study, the best approach is a hybrid governance model, where governance policies are defined centrally, but business units manage their own data within predefined security and compliance frameworks. This structure allows for flexibility while ensuring data integrity and regulatory compliance.

The importance of Data Governance is a cornerstone; tools like IBM Knowledge Catalog increase its effectiveness in data management. A sophisticated Data Governance framework is essential to maintaining a competitive advantage in the market, ensuring regulatory compliance, data security, and operational intelligence. Even in situations where initial data resources are lacking, it is possible to implement an effective Data Governance framework by focusing on creating key processes, precisely defining roles and responsibilities, and establishing priorities for business initiatives. Adopting a Data Governance framework leads to improvements in data quality, simplifies the decision-making process, and ensures greater compliance with regulations.

5.5 Comparisons with Market Leaders

IBM Knowledge Catalog (IKC) is an advanced solution for Data Governance and data cataloguing, offering seamless integration with IBM Cloud Pak for Data and AI-powered functionalities. However, a comparison with leading competitors such as Informatica and Ataccama highlights both strengths and limitations.

One of the main advantages of IKC is its deep integration with the cloud and the IBM Watson ecosystem, enabling automation of data cataloguing and classification processes through AI and machine learning. Additionally, IKC provides a scalable platform ensuring efficient management of multi-cloud and on-premises environments. Another key strength is its data virtualization capability, which allows users to access data directly without physically moving it, reducing costs and improving performance.

However, compared to Informatica, IBM Knowledge Catalog may be less mature in Data Integration functionalities. Informatica is a well-established leader in data integration, offering advanced ETL and ELT tools, along with extensive support for data lineage and data quality management. If a company requires a comprehensive solution for data integration and seamless data movement across environments, Informatica might be a more suitable choice.

On the other hand, Ataccama is known for its all-in-one approach to Data Governance and Data Quality, with a strong focus on automation and anomaly detection in data. Its platform provides robust Master Data Management (MDM) and Data Quality capabilities, often with a more intuitive user experience compared to IKC. If an organization prioritizes data quality and advanced data profiling, Ataccama could be a more effective solution.

Among the main disadvantages of IKC compared to its competitors, a steeper learning curve and more complex permission configurations stand out, which may require greater initial effort for access management and security setup. Additionally, while IKC excels in integration with IBM tools, it may be less flexible for organizations using mixed technological ecosystems.

In summary, the choice between IBM Knowledge Catalog, Informatica, and Ataccama largely depends on business priorities. IKC is an excellent option for those seeking a cloud-native AI-integrated solution, whereas Informatica excels in advanced integration capabilities and Ataccama specializes in data quality and management. The adoption of IKC should be assessed considering the required level of automation, integration needs with other systems, and internal expertise for platform management.

5.6 Results

One of the critical aspects of evaluating the efficiency of a Data Governance project is the difficulty of defining and measuring KPIs in the short term. Unlike other IT or business management projects, the benefits derived from adopting a Data Governance framework do not immediately manifest in quantitative metrics. Data quality, regulatory compliance, and information availability are factors that evolve over time, requiring a sufficiently long time horizon for proper measurement.

Although this thesis outlines several KPIs useful for assessing the effectiveness of a Data Governance model, these KPIs cannot be applied to the project under examination, as their measurement requires historical data and continuous long-term monitoring. However, a qualitative metric that can provide an initial indication of the value of Data Governance is the reduction of wasted time spent by analysts in cleaning the data necessary for their analyses. A well-implemented Data Governance system significantly reduces the time required for data cleansing activities, improving operational efficiency and allowing data analysts to focus on data interpretation rather than preparation.

In summary, assessing the effectiveness of Data Governance cannot rely solely on traditional short-term KPIs but must include qualitative metrics and long-term analysis to fully understand its impact on data management and utilization within the organization.

The integration of a robust Data Governance framework is no longer a mere option for organizations managing complex data ecosystems but a necessity for ensuring data quality, compliance, and strategic alignment with business objectives. Throughout this thesis, the implementation of IBM Knowledge Catalog within an analytical workflow has been explored, highlighting both the benefits and challenges associated with its adoption.

While the introduction of governance mechanisms brings clear long-term advantages, such as improved data reliability and operational efficiency, their impact is not immediately quantifiable through traditional key performance indicators (KPIs). Instead, qualitative benefits, such as enhanced collaboration, reduced time spent on data preparation, and increased trust in data assets, serve as early indicators of success. Over time, these improvements translate into measurable business value, reinforcing the strategic importance of Data Governance.

5.7 Future Improvements

Looking ahead, the evolution of data-driven decision-making will further emphasize the need for adaptable and scalable governance solutions. Emerging technologies, including artificial intelligence and machine learning, will play a crucial role in automating governance processes and enhancing data stewardship. However, as governance frameworks continue to mature, organizations must balance automation with human oversight to maintain data integrity and ethical compliance.

In conclusion, the successful implementation of Data Governance requires a combination of technological innovation, organizational commitment, and continuous refinement of governance strategies. This study contributes to the ongoing discourse on enterprise data management, underscoring the necessity of integrating governance practices into analytical workflows to drive sustainable, data-driven growth.

Table Name	Primary Key	Description	N. of Columns
Product	PRODUCT_PK	Contains a list of all products sold by the company, including details such as product code, description, and merchandise categories.	237
Product_Code_NS	PRODUCT_CODE_NS_PK	Manages unique product identification codes, linking them to models, materials, and colours for detailed catalog management.	15
Model	MODEL_PK	Describes the different product models available, associating each model with company codes and descriptions.	21
Material	MATERIAL_PK	List of materials used in product manufacturing, including identification codes and detailed descriptions.	29
Color	COLOR_PK	Contains the list of colours available for products, with reference information for catalog management.	25
Variant	VARIANT_PK	Defines different product variants, distinguishing possible configurations of materials, colours, and	25

		other characteristics.	
Variant Type	VARIANT_TYPE_PK	Classifies different types of product variants, such as sizes, finishes, or customizations.	9
Statistical Class	STATISTICAL_CLASS_PK	Organizes products into statistical classes for market analysis and reporting.	11
Merchandise class	MERCHANDISE_CLASS_PK	Divides products into broad merchandise categories to facilitate catalog management.	13
Merchandise Subclass	MERCHANDISE_SUBCLASS_PK	Further details merchandise classes into more specific subclasses for better catalog organization.	16
Style Channel	STYLE_CHANNEL_PK	Identifies different style channels products belong to, such as design lines or collections.	8
Barcode	BARCODE_PK	Associates a unique barcode with each product for traceability and logistical management.	78
Size	SIZE_PK	Manages information on product sizes, essential for apparel and accessories.	15
Size Grid	SIZE_GRID_PK	Defines the size grids used for products, specifying available	12

		measurements in each category.	
Season	SEASON_PK	Indicates the reference season for each product, useful for seasonal collections and marketing strategies.	31
Macroseason	MACROSEASON_PK	Groups seasons into broader macro-categories, such as summer and winter, for a high-level collection view.	24
Store	STORE_PK	List of physical and online stores where products are sold, including identifying codes and locations.	237
Sales Channel	SALES_CHANNEL_PK	Describes the company's sales channels, such as e-commerce, physical retail, or wholesale distribution.	16
Country	COUNTRY_PK	Lists the countries where the company operates, useful for market analysis and expansion strategies.	36
Distributive Region	DISTRIBUTIVE_REGION_PK	Defines geographic regions for product distribution, optimizing logistics.	8
Currency	CURRENCY_PK	Manages currencies used for commercial transactions in	17

		different countries.	
Customer	CUSTOMER_PK	Collects information on registered customers, including purchase preferences and personal data.	168
Commercial Area	COMMERCIAL_AREA_PK	Segments the market into commercial areas for analysis and sales strategies.	11
Commercial Market	COMMERCIAL_MARKET_PK	Defines different markets in which the company operates, allowing targeted positioning strategies.	11
TIME_Day	DAY_PK	Contains detailed information on days, used for reporting and sales analysis.	16
Day Of the Week	DAY_WEEK_PK	Indicates the day of the week associated with each date in the database.	3
TIME_Week	WEEK_PK	Organizes the calendar into weeks for periodic sales analysis.	14
TIME_MONTH	MONTH_PK	Defines calendar months for financial reports and trend analysis.	14
Time Quarter	QUARTER_PK	Divides the year into quarters for business analysis and periodic balances.	10

Time Year	YEAR_PK	List of fiscal years recorded in the database for historical reference.	6
Commercial Week	COMMERCIAL_WEEK_SK	Commercial weeks defined for sales analysis and business strategies.	12
Commercial Month	COMMERCIAL_MONTH_SK	Commercial months used for strategic sales planning.	15
Commercial Quarter	COMMERCIAL_QUARTER_SK	Commercial quarters for performance comparisons and corporate reporting.	10
Commercial Year	COMMERCIAL_YEAR_SK	Commercial years for evaluating annual sales performance.	5
F_SO_TRANSACTION	JOBID_L2_INS	Central table of sales transactions, recording all operations carried out in various points of sale and distribution channels.	187

Table 13: Tables of Data Marts

6 BIBLIOGRAPHY

Adamson, C., 2010. *Star Schema: The Complete Reference*. s.l.:McGraw-Hill.

Anwar, M., 2024. *Il futuro dell'intelligenza artificiale nel data warehousing: tendenze e previsioni*. [Online] Available at: https://www.astera.com/it/type/blog/ai-in-data-warehousing/

Atlan, 2024. *Atlan*. [Online] Available at: https://atlan.com/know/data-governance/performance-metrics/

Atlan, 2024. *DAMA-DMBOK Framework: What It Is and How To Adopt It?*. [Online] Available at: https://atlan.com/dama-dmbok-framework/

AWS, n.d. *Qual è la differenza tra ETL ed ELT?*. [Online] Available at: https://aws.amazon.com/it/compare/the-difference-between-etl-and-elt/

Bhuvana Jayabalan, V. S. P. S. A. M. S. K., 2024. *Enhancing Cloud Security: Artificial Intelligence-based Data Classification Model for Cloud Computing*. [Online] Available at: https://ijisae.org/index.php/IJISAE/article/view/5611

Borzi, M., 2023. *Automated Data Integration and*. s.l.:s.n.

Bowman, K., 2025. *Pathlock*. [Online] Available at: https://pathlock.com/learn/data-governance-metrics/

Boyd M Knosp, C. K. C. D. A. D. E. V. B. T. R. C., 2022. *Understanding enterprise data warehouses to support clinical and translational research: enterprise information technology relationships, data governance, workforce, and cloud computing*. [Online] Available at: https://pubmed.ncbi.nlm.nih.gov/35289370/

Bregata, L., 2019. *Traditional ETL Best Practice & Business Analytics*. s.l.:s.n.

Buglio, R., 2018. *An overview of IBM's Watson Knowledge Catalog*. [Online] Available at: https://www.youtube.com/watch?v=6gwT-CCMEYo

BuzzyBrains, 2024. *What is Microsoft SSIS: Key Features, Benefits and More*. [Online] Available at: https://www.buzzybrains.com/blog/what-is-microsoft-ssis-key-features-benefits/

Carter, R., 2024. *What is IBM WatsonX? The Evolution of IBM Watson*. [Online] Available at: https://aitoday.com/technology/what-is-ibm-watsonx-the-evolution-of-ibm-watson/

Cavanell, Z., 2023. *AI-powered Data Classification | Microsoft Purview*. [Online] Available at: https://techcommunity.microsoft.com/blog/microsoftmechanicsblog/ai-powered-data-classification--microsoft-purview/3919206

Chen, M., 2022. *What Is a Data Lakehouse?*. [Online] Available at: https://www.oracle.com/si/big-data/what-is-data-lakehouse/

Chiarello, D., 2020. *Datawarehouse implementation and development of Advanced Analytics methods to support business strategies*. s.l.:s.n.

Chu, D., 2025. *Secoda*. [Online] Available at: https://www.secoda.co/blog/data-governance-metrics

Chukwurah, N., 2024. Frameworks for effective data governance: best practices, challenges, and implementation strategies across industries. *Computer Science & IT Research Journal*.

Compendium, 2024. *Integrare la Business Intelligence nelle Operazioni Quotidiane: Guida Pratica per CEO e COO*. [Online] Available at: https://www.it-compendium.com/it/blog/integrare-la-business-intelligence-nelle-operazioni-quotidiane-guida-pratica-ceo-e-coo

Cultur-e, 2018. *IBM, la storia*. [Online] Available at: https://www.fastweb.it/fastweb-plus/digital-magazine/ibm-la-storia/

D.Foote, K., 2024. *Data Governance Metrics: How to Measure Success*. [Online] Available at: https://www.dataversity.net/data-governance-metrics-how-to-measure-success/

Databricks, n.d. *Data Lakehouse*. [Online] Available at: https://www.databricks.com/glossary/data-lakehouse

DataCat, 2013. *Top 10 SQL Server Integration Services Best Practices*. [Online] Available at: https://techcommunity.microsoft.com/blog/sqlserver/top-10-sql-server-integration-services-best-practices/305163

Datamaze, n.d. *La Business Intelligence... dalla BI alla Z!*. [Online] Available at: https://www.datamaze.it/risorse/tutto-sulla-business-intelligence

Devart, n.d. *SSIS Tutorial — from Basics to*. [Online] Available at: https://www.devart.com/ssis/what-is-ssis.html

Digital4, 2021. *Data-driven: cosa significa e perché un approccio basato sui dati è importante in azienda*. [Online] Available at: https://www.digital4.biz/marketing/big-data-e-analytics/sei-regole-d-oro-per-un-data-driven-marketing-di-successo/

Dresse, L., 2023. *Data Governance Metrics: 5 Best Practices for Measuring the Effectiveness of Your Program*. [Online]
Available at: https://www.cdomagazine.tech/branded-content/data-governance-metrics-5-best-practices-for-measuring-the-effectiveness-of-your-program

Eliasy, N., 2024. *Analisi di un progetto di integrazione della generative AI ai processi ETL aziendali*. s.l.:s.n.

Evren Eryurek, U. G., 2019. *Principles and best practices for data governance in the cloud*. [Online]
Available at: https://cloud.google.com/blog/products/data-analytics/principles-and-best-practices-for-data-governance-in-the-cloud?utm_source=chatgpt.com

EWSolution, 2024. *Data Governance Metrics: Performance for Governance and Data Stewardship*. [Online]
Available at: https://www.ewsolutions.com/performance-metrics-data-governance-data-stewardship/

Ferrari, G., 2024. *Data warehouse, 5 motivi per usare AI e machine learning*. [Online]
Available at: https://www.zerounoweb.it/analytics/data-warehouse-5-motivi-per-usare-ai-e-machine-learning/

Filippo La Noce, L. D., 2008. *Data Warehousing - dal dato all'informazione*. s.l.:Franco Angeli.

Gade, K. R., 2024. Data Governance in the Cloud: Challenges and Opportunities. *MZ Computing Journal*.

Gandini, A., 2024. *L'importanza della Data Culture nell'era dei Business Data-Driven*. [Online]
Available at: https://blog.besharp.it/it/limportanza-della-data-culture-nellera-dei-business-data-driven/

Gartner, 2020. *Software as a Service (SaaS)*. [Online]
Available at: https://www.gartner.com/en/information-technology/glossary/software-as-a-service-saas

Girardo, A., 2024. *Data driven: perché occorre partire da una base culturale*. [Online]
Available at: https://www.zerounoweb.it/big-data/data-driven-perche-occorre-partire-da-una-base-culturale/

Google Cloud, 2023. *Che cos'è Platform as a Service (PaaS)?*. [Online]
Available at: https://cloud.google.com/learn/what-is-paas?hl=it

Greco, A., 2018. *E-Commerce monitoring solution for product allocation and marketing planning forecasting*. s.l.:s.n.

Grossi, M., 2020. *IBM Cloud: La data governance con IBM Watson Knowledge Catalog*. s.l.:s.n.

Hewlett Packard, n.d. *What is a Data Lakehouse?*. [Online] Available at: https://www.hpe.com/emea_europe/en/what-is/data-lakehouse.html

IBM, 2020. *IBM Knowledge Catalog*. [Online] Available at: https://www.ibm.com/it-it/products/knowledge-catalog

IBM, n.d. *Cos'è la business intelligence (BI)?*. [Online] Available at: https://www.ibm.com/it-it/topics/business-intelligence

IBM, n.d. *IBM Cloud Pak for Data*. [Online] Available at: https://wcs-it-ibmshowcase-itsinformation.mydmportal.com/IBMCloudPakforData

IBM, n.d. *IBM Products*. [Online] Available at: https://www.ibm.com/products/watson-ai

IBM, n.d. *Risorse*. [Online] Available at: https://www.ibm.com/it-it/products/knowledge-catalog/resources

IBM, n.d. *What is a data lakehouse?*. [Online] Available at: https://www.ibm.com/think/topics/data-lakehouse

Immon, B., 2016. *Data Lake Architecture: Designing the Data Lake and Avoiding the Garbage Dump*. s.l.:s.n.

Islam, A., 2024. *DATA GOVERNANCE AND COMPLIANCE IN CLOUD-BASED BIG DATA ANALYTICS: A DATABASE-CENTRIC REVIEW*. [Online] Available at: https://www.semanticscholar.org/paper/DATA-GOVERNANCE-AND-COMPLIANCE-IN-CLOUD-BASED-BIG-A-Islam/82d961529aa1a3144c1e1bd7c0de283abffae8f6

Jolas, X., 2023. *The Importance of Data Ownership in Data Governance*. [Online] Available at: https://dainstudios.com/insights/importance-data-ownership-data-governance/?utm_source=chatgpt.com

Kalibbala, J., 2023. *IBM Watson: A Cheat Sheet*. [Online] Available at: https://www.techrepublic.com/article/ibm-watson-the-smart-persons-guide/

Karol Bliznak, M. M. A. P., 2024. A Systematic Review of Recent Literature on Data Governance (2017-2023). *IEEE Access*.

Kazlow, D., 2024. *The Data Maturity Model: A Comprehensive Guide*. [Online] Available at: https://thedatagovernance.com/data-maturity-model/?utm_source=chatgpt.com

Koppichetti, R. K., 2023. *Framework of Hub and Spoke Data Governance Model for Cloud Computing*. [Online] Available at: https://www.semanticscholar.org/paper/Framework-of-Hub-and-Spoke-Data-Governance-Model-Koppichetti/869ebf1f5992dfabb5c7210beca61f8cc0d20a50

Ladley, J., 2012. *Data Governance How to Design, Deploy, and Sustain an Effective Data Governance Program*. s.l.:s.n.

Lashch, I., 2023. *Data Governance for Cloud-Based Environments: Ensuring Security and Compliance*. [Online] Available at: https://lightpointglobal.com/blog/data-governance-for-cloud-based-environments?utm_source=chatgpt.com

Lavecchia, V., 2023. *Differenza tra Normalizzazione e Denormalizzazione nei database*. [Online] Available at: https://vitolavecchia.altervista.org/differenza-tra-normalizzazione-e-denormalizzazione-nei-database/

Lin, S., 2024. *Semarchy*. [Online] Available at: https://www.semarchy.com/blog/how-to-measure-success-with-data-governance-metrics/

Lorusso, M. M., 2021. *IBM Cloud Pak for Data cos'è e perché è il motore dell'AI in azienda*. [Online] Available at: https://www.impresacity.it/news/26407/ibm-cloud-pak-for-data-cos-e-e-perche-e-il-motore-dell-ai-in-azienda.html

Maffeo, L., 2023. *Designing Data Governance from the Ground Up*. s.l.:s.n.

Mathew, A., 2024. *Cloud Data Sovereignty Governance and Risk Implications of Cross-Border Cloud Storage*. [Online] Available at: https://www.isaca.org/resources/news-and-trends/industry-news/2024/cloud-data-sovereignty-governance-and-risk-implications-of-cross-border-cloud-storage?utm_source=chatgpt.com

Mazzi, G., 2021. *Cloud native: approaches and benefits*. [Online] Available at: https://edalab.it/en/cloud-native-vantaggi/

Melody Chien, J. M., 2024. *Magic Quadrant for Augmented Data Quality Solutions*. [Online] Available at: https://www.gartner.com/doc/reprints?id=1-2GUODPN3&ct=240307&st=sb

Microsoft, 2024. *Data Flow Performance Features*. [Online] Available at: https://learn.microsoft.com/en-us/sql/integration-services/data-flow/data-flow-performance-features?view=sql-server-ver16&redirectedfrom=MSDN

Microsoft, n.d. *Data warehouse moderni per piccole e medie imprese*. [Online] Available at: https://learn.microsoft.com/it-it/azure/architecture/example-scenario/data/small-medium-data-warehouse

MJV, 2023. *Cinque passi per implementare una cultura dei dati nella tua azienda*. [Online]
Available at: https://www.mjvinnovation.com/it/blog/cultura-dei-dati/

Mosley, M., 2008. *DAMA-DMBOK Functional Framework*. [Online] Available at: https://governance.foundation/assets/frameworks/dama/DAMA-DMBOK_Functional_Framework_v3_02_20080910.pdf

New Gen Apps, 2018. *IBM Watson and its Key Features*. [Online] Available at: https://www.newgenapps.com/en/blogs/ibm-watson-and-its-key-features

Ocean, D., n.d. *Digital Ocean*. [Online] Available at: https://www.digitalocean.com/resources/articles/cloud-governance

Ofori, A. Y., 2024. *Data Security and Governance in Multi-Cloud Computing Environment*. [Online]
Available at: https://www.semanticscholar.org/paper/Data-Security-and-Governance-in-Multi-Cloud-Yeboah-Ofori-Jafar/f5719f6904ead6dae3ecb0cc4b00e51a3fb757cc

Onyinye Obioha Val, O. S.-A. T. M. K. M. O. G. O. O. O. O. O., 2024. *Real-Time Data Governance and Compliance in Cloud-Native Robotics Systems*. [Online] Available at: https://www.semanticscholar.org/paper/Real-Time-Data-Governance-and-Compliance-in-Systems-Val-Selesi-Aina/63add02352515e6468abd3277c67199ae410a2b6

OptimizeMRO, 2024. *DAMA-DMBOK: A Comprehensive Framework for Data Management*. [Online]
Available at: https://www.optimizemro.com/blog/dama-dmbok-a-comprehensive-framework-for-data-management/

Osservatorio Cloud Transformation, 2024. *Cloud Computing, cos'è*. [Online] Available at: https://blog.osservatori.net/it_it/cloud-computing-significato-vantaggi

Ot, A., 2023. *What is a Data Lakehouse? Definition, Benefits and Features*. [Online] Available at: https://www.datamation.com/big-data/what-is-a-data-lakehouse/

Paypro Global, 2024. *Cos'è il monitoraggio cloud?*. [Online] Available at: https://payproglobal.com/it/risposte/cose-il-monitoraggio-cloud/

Pegdwendé Sawadogo, J. D., 2021. On data lake architectures and metadata management. *Cornell University*.

Pereira, D., 2024. *Modello aziendale IBM*. [Online] Available at: https://businessmodelanalyst.com/it/ibm-business-model/

Perel, N., 2023. *Using AI for Complex Data Discovery*. [Online] Available at: https://www.rubrik.com/blog/technology/23/12/guide-to-ai-driven-data-discovery-and-classification

Petricca, R., 2024. *Le 10 applicazioni cloud più competitive del momento*. [Online] Available at: https://www.agendadigitale.eu/industry-4-0/le-10-applicazioni-cloud-piu-competitive-del-momento/

PureStorage, n.d. *Che cos'è la normalizzazione dei dati?*. [Online] Available at: https://www.purestorage.com/it/knowledge/what-is-data-normalization.html

Pwint Phyu Khine, Z. S. W., 2018. *Data lake: a new ideology in big data era*. [Online] Available at: https://www.itm-conferences.org/articles/itmconf/abs/2018/02/itmconf_wcsn2018_03025/itmconf_wcsn2018_03025.html

Ratman, K. V., 2024. *AUTOMATING CLOUD SECURITY AND DATA GOVERNANCE CHALLENGES IN MULTI-CLOUD ENVIRONMENTS*. [Online] Available at: https://www.semanticscholar.org/paper/AUTOMATING-CLOUD-SECURITY-AND-DATA-GOVERNANCE-IN-Ratnam/5a89c83180b0ae9713729a540002268442c19826

Red Hat, 2023. *Servizi cloud: Confronto tra IaaS, PaaS e SaaS*. [Online] Available at: https://www.redhat.com/it/topics/cloud-computing/iaas-vs-paas-vs-saas

Regesta Lab, 2024. *Come trasformare l'azienda in una data driven company*. [Online] Available at: https://www.regestaitalia.eu/regesta-lab/come-trasformare-lazienda-in-una-data-driven-company/

Rezzani, A., 2012. *Business Intelligence - Porcessi, metodi, utilizzo in azienda*. s.l.:Feltrinelli.

Rubocki, B., 2018. *Performance Techniques for SSIS in Azure Data Factory*. [Online] Available at: https://pragmaticworks.com/blog/performance-techniques-for-ssis-in-azure-data-factory

Secureframe, n.d. *Secureframe*. [Online] Available at: https://secureframe.com/hub/grc/data-governance-metrics

Seidor, n.d. *Analítica avanzada con IBM*. [Online] Available at: https://www.seidor.com/it-it/data/analitica-avanzata/ibm

Siatec, 2024. *I 6 step per una cultura aziendale Data-Driven*. [Online] Available at: https://siatec.it/i-6-step-per-una-cultura-aziendale-data-driven/

Smallcombe, M., 2023. *ETL vs ELT:*. [Online] Available at: https://www.integrate.io/blog/etl-vs-elt/

SmallNet Consulting, 2025. *Leveraging IBM Knowledge Catalog for Effective Data Governance and AI Readiness*. [Online] Available at: https://www.smallnetconsulting.co.uk/leveraging-ibm-knowledge-catalog-for-effective-data-governance-and-ai-readiness/

Spiller, A., 2021. *Il ruolo della Data Integration nei processi si digitalizzazione delle imprese manufatturiere*. s.l.:s.n.

Stedman, C., 2023. *Definition Data Lakehouse*. [Online] Available at: https://www.techtarget.com/searchdatamanagement/definition/data-lakehouse

Talarico, A., 2024. *Dddm: così i dati diventano leva strategica per le aziende*. [Online] Available at: https://www.agendadigitale.eu/cittadinanza-digitale/data-management/dddm-cosi-i-dati-diventano-leva-strategica-per-le-aziende/

Talend, n.d. *Che cos'è la Business Intelligence (BI)?*. [Online] Available at: https://www.talend.com/it/resources/what-is-business-intelligence/

Ticong, L., 2024. *Mastering AI Data Classification: Ultimate Guide*. [Online] Available at: https://www.datamation.com/big-data/ai-data-classification/

Tot, R., 2024. *What is IBM Watson? Complete Guide to IBM's Advanced AI Solutions*. [Online]
Available at: https://articlesbase.com/tech/emerging-technologies/artificial-intelligence/ai-tools-and-software/what-is-ibm-watson-complete-guide-to-ibms-advanced-ai-solutions/

Vidette Poe, P. K. a. S. B., 1998. *Building a Data Warehouse for Decision Support*. s.l.:Prentice-Hall.

Vierrath, M. W., n.d. *IBM Cloud Pak for Data*. [Online] Available at: https://www.valantic.com/en/business-analytics/ibm-cloud-pak-for-data/

Vincent, J., 2024. *Top Best Practices for Data Governance in the Cloud: Security, Compliance, and Accessibility*. [Online] Available at: https://www.cloudsecureplatform.com/top-best-practices-data-governance/?utm_source=chatgpt.com

Wikipedia, 2025. *IBM*. [Online] Available at: https://en.wikipedia.org/wiki/IBM