POLITECNICO
MILANO 1863

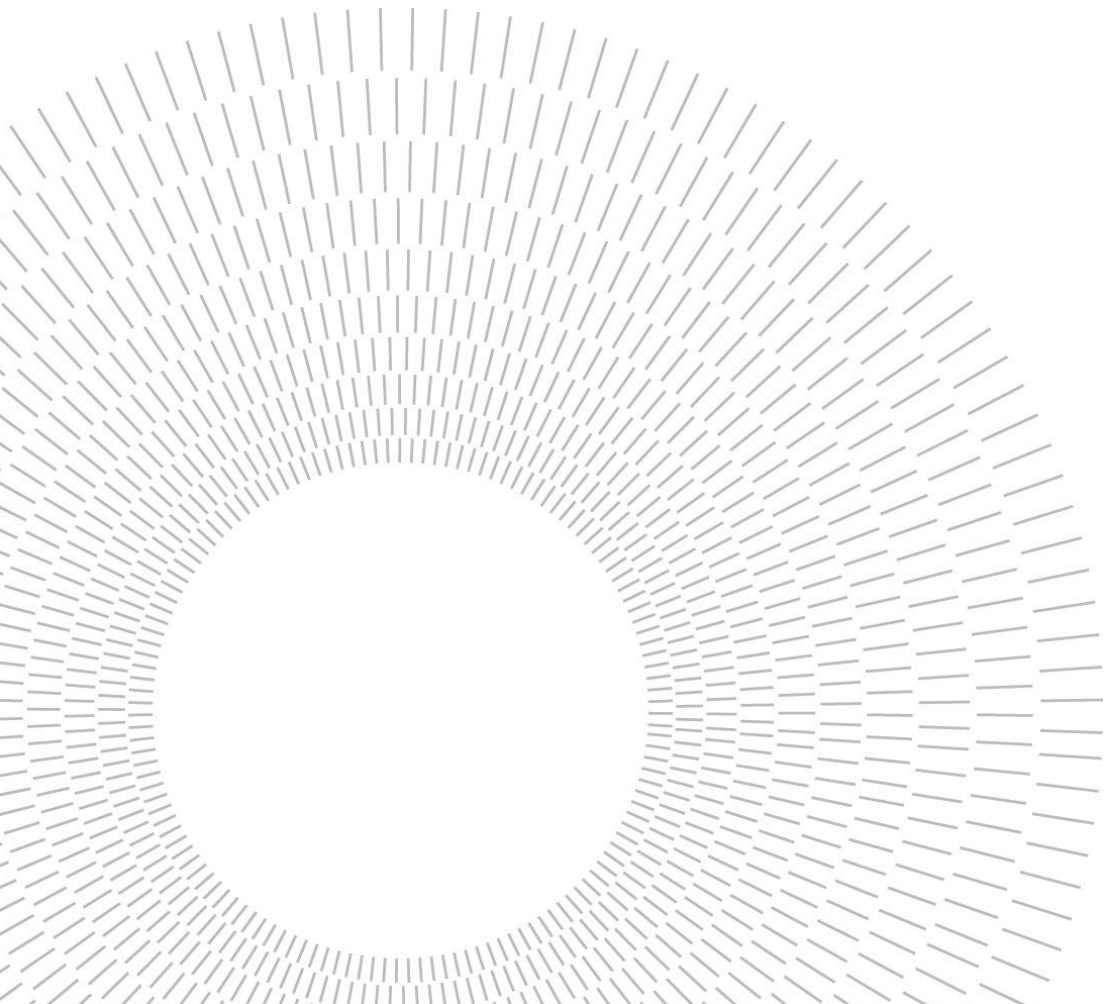
SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Data Management: analysis of market evolution over the last 10 years

TESI DI LAUREA MAGISTRALE IN
MANAGEMENT ENGINEERING-INGEGNERIA GESTIONALE

Author: **Simone Gentili**

Student ID:	10801668
Advisor:	Carlo Vercellis
Co-advisor:	Irene di Deo
Academic Year:	2021-2022



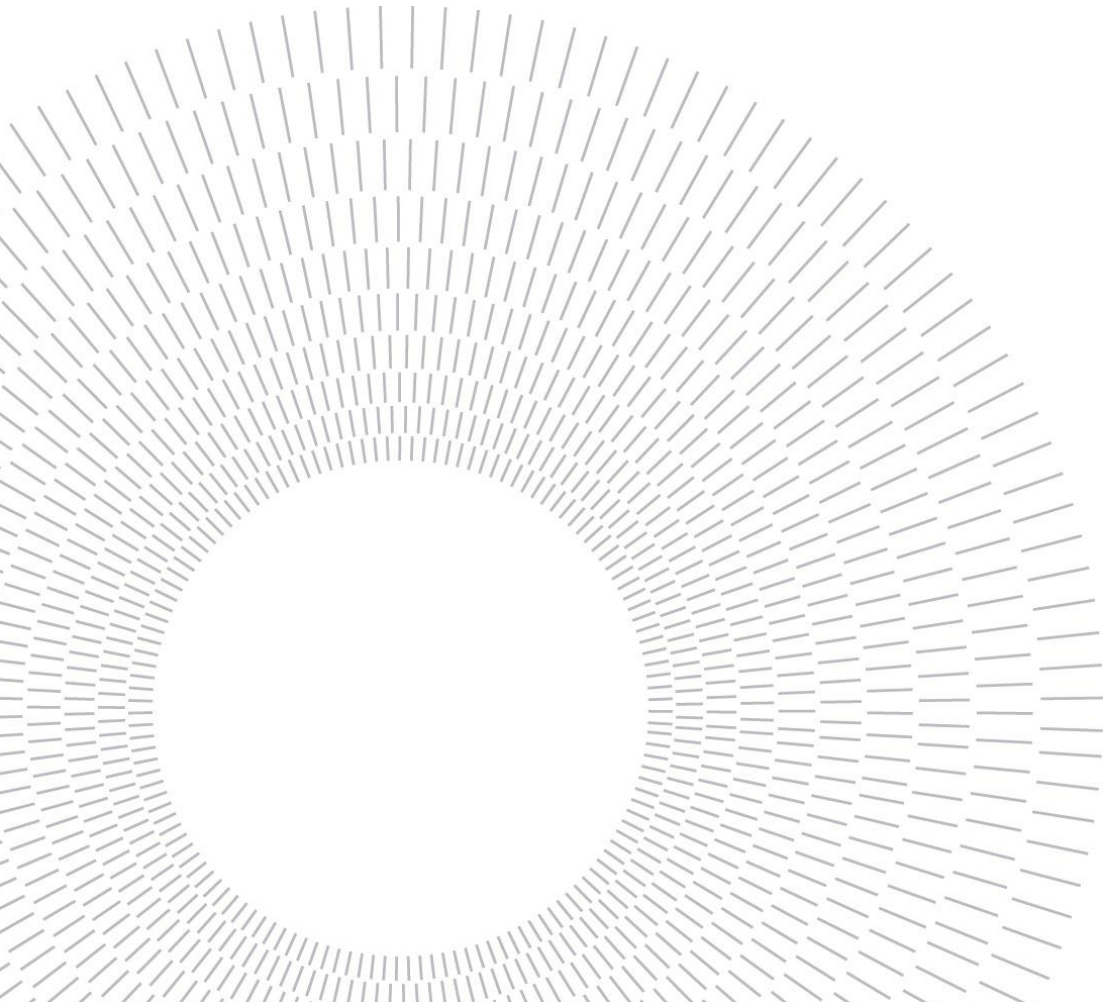
Alla mia famiglia, per aver compreso ogni mio silenzio e aver incondizionatamente supportato ogni mia scelta.

A tutti i miei amici, tra Terni, Bologna e Milano, per avermi accompagnato in questa avventura, colmandola di divertimento e stimolandomi a dare sempre il meglio di me.

Infine al Professor Carlo Vercellis, per aver riposto in me la sua fiducia e avermi dato la possibilità di lavorare a questa tesi.

Senza di voi non sarei dove sono ora, ma soprattutto non sarei la persona che sono ora, per questo vi ringrazio.

Simone



Abstract (English version)

Data Management has been a major topic in recent years. Data management encompasses those processes that allow data to be gathered, stored, and used in a secure and efficient manner. Extensively researched in the academic environment, its significance extends beyond the mere diffusion of technology advancements to developing new employment prospects and encouraging economic growth in general. For these reasons, an increasing number of organisations are specialising in various categories in order to create new technologies to speed up existing market processes and address unresolved difficulties.

The focus of this study is to examine the various Data Management technologies that have been offered in recent years. Through an analysis of enterprises, the study investigates changes in the sorts of technologies created and their geographical distribution.

This study's implementation was separated into three stages. First, a number of papers on Data management were examined, and a search for frameworks on the market that aid in Data management was conducted. Then there was the analysis of the DAMA-DMBOOK, a manual that thoroughly examines all aspects of Data Management. In the second step, a dataset was constructed for the analysis of the census data. The latter includes all companies that have managed data since 2012. The third and last process is the interpretation of the data via an analysis, emphasising the tendencies discovered in the sample analysed. At the end of the study, it can be stated that over the years the solutions and services of greatest interest are those related to Data Integration, however, enterprises are introducing new emergent typologies to fill market shortages.

Keywords: Data Integration, Data Management, DM-BOOK.

Abstract (Italian version)

La Data Management nel corso degli ultimi anni è stato un argomento di fondamentale importanza. Per Data Management si intendono tutte le pratiche che consentono di raccogliere, conservare e utilizzare i dati in modo sicuro ed efficiente. Ampiamente indagata nel panorama accademico, la sua rilevanza non si limita alla mera disseminazione di innovazioni tecnologiche, ma è legata anche alla nascita di nuove opportunità di lavoro e in generale alla stimolazione della crescita economica. Per queste ragioni sempre più aziende si stanno specializzando nelle diverse categorie, per sviluppare sempre nuove tecnologie che permettano di velocizzare quei processi già presenti nel mercato e di colmare le problematiche irrisolte.

Lo scopo di questa ricerca è l'analisi delle varie tecnologie proposte nel corso degli ultimi anni riguardanti la Data Management. La ricerca approfondisce, attraverso un censimento delle aziende, quali sono state le tendenze riguardanti le tipologie di tecnologie sviluppate e la loro distribuzione geografica.

La realizzazione di questo studio è stata articolata in tre diverse fasi Inizialmente sono stati analizzati alcuni articoli relativi alla Data Management ed è stata effettuata una ricerca dei Framework presenti nel mercato che aiutino la gestione del dato. Successivamente c'è stato lo studio del DAMA-DMBOOK, un manuale in cui vengono analizzate dettagliatamente tutte le aree della Data Management. In un secondo momento è stato costruito un dataset per l'analisi dei dati censiti. Quest'ultimo comprende tutte le aziende che dal 2012 gestiscono i dati.
La terza ed ultima fase, l'interpretazione dei risultati attraverso un'analisi di essi, ponendo in evidenza i trend riscontrati nel campione preso in esame. A conclusione dello studio si può affermare che nel corso degli anni le soluzioni e i servizi di maggiore interesse sono quelle relative alla Data Integration, nonostante le aziende stiano sviluppando nuove tipologie emergenti per soddisfare i divari presenti nel mercato.

Parole chiave: Data Integration, Data Management, DM-BOOK.

Table of Contents

List of Figure

List of Table

Chapter 1: Introduction

Data Management is an important part to deploy the IT systems that run business applications and provide analytical information to help drive operational decision-making and strategic planning by corporate executives, business managers and other end users. Data Management is the process of ingesting, storing, organizing, and maintaining the data created and collected by an organization.

The combination of these processes aims to ensure that the data in corporate systems is accurate, available, and accessible. Most of the required work is done by business users who participate in some parts of the process to ensure that the data meets their needs. These business users are called Data Management professionals, which are any person who works in any facet of Data Management, from technical management of data throughout its lifecycle to ensuring that data is properly utilized and leveraged, to meet strategic organizational goals. Data Management professionals fill numerous roles, from highly technical to strategic business.

Data management encompasses a wide range of processes. They range from database technical deployment and performance to the capacity to make trustworthy judgements about how to extract strategic value from data. As a result, Data management necessitates both technical and non-technical abilities. Information and data are assets in the sense that firms invest in them in order to improve their value in the future. To guarantee that an organisation has high-quality data that meets its strategic goals, management of data challenges must be shared across business and information technology responsibilities. The bulk of organisations relies on data and information to run their operations on a regular basis.

Whether or not an organization gets value from its analytics, it cannot even transact business without data. So, organizations have always needed to manage their data daily, but technological changes have expanded the scope of this management need as they have changed people's understanding of data. These changes have enabled

organizations to use data in new ways to create products, share information, create knowledge, and improve organizational success. But the rapid growth of technology and with-it human capacity to produce, capture, and mine data for meaning has intensified the need to manage data effectively.

Thus, everything is tied to data, yet effectively defining data is difficult. Long-standing definitions of data highlight its significance in expressing world realities. Another viewpoint from information technology is that data is also defined as information that has been saved in digital form. Another viewpoint holds that data reflect facts and are the sole way to talk clearly and meaningfully about the world. So, while there are several definitions of data, it is vital to remember that data without context is worthless.

Context can be thought of as data's representational system; such a system includes a common vocabulary and a set of relationships between components. If it is possible to know the conventions of such a system, it is possible to interpret the data. These typologies of data are called Metadata, which are data about data. So, it is essential to know how data can be used to reach organizational goals, which data are available for an organisation and what might be accomplished with it. There is a necessity to find a balance between strategic and operational needs. This balance can come following a set of principles that recognize salient features of Data Management and guide Data Management practices. Many different operations fall under Data management. They span the technical deployment and performance of databases to the ability to make reliable decisions about how to get strategic value from data. As a result, both technical and non-technical skills are required for Data management. Information and data are not only assets in the sense that businesses invest in them to increase their worth in the future. To ensure that an organisation has high-quality data that satisfies its strategic goals, the responsibility for managing data problems must be shared across business and information technology roles. Data and information are at the core of most businesses, making daily Data management a necessity. With technological advancements, organizations can now utilize data in new ways to enhance success. However, with the vast amount of data produced and captured, effective Data management has become increasingly important. Defining data is a complex task, with various viewpoints on its meaning. Nevertheless, data must be viewed within its context to be valuable. Context refers to a representational system that includes a shared vocabulary and relationships between components, which can be described as metadata.

Therefore, understanding available data and its potential use is crucial to achieving organizational goals, while balancing strategic and operational needs. Although striking a balance between these two is difficult, data considerations are necessary in order to help organisations achieve their objectives. One of them is that Data is an asset and, as such, differs materially from other assets in ways that have an impact on how it is handled. These facts have economic worth, which may and should be conveyed in that way. When the Data is referred to as a "asset," it is implied that the data is valuable financially. Organizations should develop standardised techniques for assessing that value if they want to make better data decisions. They should weigh both the disadvantages of low-quality data and its benefits. So, managing data is required to achieve this degree of high-quality data. Companies must make sure they are aware of the standards of quality that their stakeholders demand from them to manage quality data. The management of Metadata, as was also noted above, is an element that contributes to the achievement of these organisational objectives. Data used to organise and make use of data is known as metadata. Data cannot be manipulated or touched, therefore understanding what it is and how to utilise it requires definition and information in the form of metadata. The requirement to develop a planning map in order to coordinate operations is another aspect. Due to the frequent interchange of data across the various usage sites, data must be checked. Teams with specific expertise must handle these interactions. Both technical and non-technical skills, as well as the ability to work with others, are required for data management. However, it's important to additionally take the data's lifespan into account. Data management requires lifecycle management since it has an expiration date. The data lifecycle may be fairly intricate since data creates more data. As a result, different data types have various lifespan characteristics. Therefore, data management systems must be able to distinguish between these differences and be flexible enough to meet a range of data lifecycle requirements.

These are some of the principles that recognize salient features of Data Management and guide Data Management practices. But there is one that was not mentioned regarding the intrinsic risks associated with data. Data, in addition to being an asset, also poses a risk to a business. Data can be misplaced, stolen, or abused. Organizations must think about the ethical consequences of their data use. Risks associated with data must be controlled as part of the data lifecycle. Also, low-quality data, whether

erroneous, partial, or out-of-date, poses a danger since the information is incorrect. However, data is problematic since it may be misinterpreted and abused. The finest quality data provides the most value to organisations.

Data is evaluated using a variety of criteria, including availability, relevance, completeness, accuracy, consistency, timeliness, usability, meaningfulness, and comprehension. However, data might exhibit these features while still containing asymmetric information about what is achievable and what you need to know to make an informed decision. Corporate liabilities such as information gaps can have a substantial influence on operational performance and profitability. Organizations that realise the benefits of high-quality data can take steps to improve data and information quality and usability within a regulated framework.

Because of the increasing importance of information as an organisational asset across all industries, authorities are concentrating more on potential information uses and abuses. Similarly, when customers become more aware of how their data is being utilised, they demand more data security and efficiency, as well as respect for their privacy. As a result, Data Management jobs are usually larger than they were previously. As a result, creating an appropriate strategy for managing and utilising this data is crucial. A strategy is a set of decisions and activities that, when combined, provide a high-level route of action for achieving high-level goals.

A strategic plan is a high-level plan for achieving high-level goals. Business plans must be included in a data strategy. In this approach, information may be used to gain a competitive edge and support corporate goals. Data strategy stems from an awareness of what our business plan requires: which data firms require, how these data must be handled, how companies will utilise these data, and how they will keep these data to ensure dependability and security over time. A robust data strategy is essential for organizations to effectively harness the potential of big data, and it typically includes a data management program strategy. The Data management program strategy is a plan that outlines the processes and policies for preserving and enhancing data integrity, quality, access, and security, while also minimizing known and inferred risks. The program strategy typically includes activities such as Data governance, Data quality management, Data security, and metadata management, all of which are crucial for ensuring that Big data is trustworthy and accurate. The concept of Big Data has gained

widespread attention due to the exponential growth of data generated by various sources such as social media, internet usage, mobile devices, and sensors. The increasing availability of Big data offers immense potential for organizations to gain valuable insights, identify new business opportunities, and improve decision-making processes. However, managing such massive amounts of data poses significant challenges, especially for traditional data management techniques.

Data management involves a set of processes that ensure that data is accurate, reliable, accessible, and secure. Effective data management includes activities such as data quality, data integration, data governance, and metadata management. Data quality is essential for ensuring that data is accurate, consistent, and complete. Data integration involves combining data from various sources into a single, unified view. Data governance refers to the processes and policies for managing data, including data privacy, security, and compliance. Metadata management involves managing information about data, such as its origin, format, and meaning.

Big data presents several challenges for data management, such as data volume, velocity, and variety. The sheer volume of data generated by Big data sources makes it difficult to store and process the data efficiently. The velocity at which data is generated requires real-time processing capabilities to enable quick decision-making. The variety of data sources and formats require integration and governance processes to ensure consistency and accuracy across the data.

Effective data management is crucial for gaining insights from Big data. By ensuring that data is of high quality, accessible, and secure, organizations can maximize the value of Big data. Proper data management can also reduce the risk of errors and inconsistencies in data analysis, leading to better decision-making. So Big data offers immense potential for organizations to gain valuable insights, but it requires effective data management to make sense of it. Organizations need to invest in robust data management strategies to ensure that big data is accurate, consistent, and trustworthy, leading to better decision-making and improved business outcomes. By implementing a Data management program strategy, organizations can effectively manage big data, reduce the risk of errors, and gain valuable insights that can inform better decision-making processes. A comprehensive data management program strategy is, therefore, a key component of a successful data strategy.

The purpose of this analysis is to follow new market developments. To keep track of these new technologies, an analysis will be conducted in which enterprises will be classified geographically and according to the services they provide. This analysis is meant not simply to maintain track of these businesses, but also to track trends over the last 5 years. This study will allow us to understand how technologies have changed and where firms have received more investment. With firms were considered all startups which are gaining more importance over the years. In addition, well-known companies were not taken into account because they would have an impact on the analysis and consequently the results would not be reliable.

Chapter 2: Literature Review

The scope of this section is to provide a general review of the Data Management literature. In this sense, there was initial research on the different frameworks present in the market. This analysis points out that there are two important frameworks: DCAM and DAMA framework. There will be presented the two frameworks and put in evidence main differences between these two. After that, there will be a definition of the different areas of the DAMA Wheel. Lastly, there is a short presentation of emerging Data management trends which are continuously growing in the market

2.1 Data Management Framework

Data has emerged as a fundamental component of company operations. It is increasingly recognised as a critical input component in the complete range of business and organisational activities. The development and administration of a data control environment is a common theme for organisations that are efficient in their use of data. The scope is to decrease operational costs, automate manual procedures, combine redundant systems, eliminate reconciliation, and boost commercial potential. Data must be defined consistently since it reflects an object. Without concrete and correct data, all other procedures derived from them are useless. The importance and the complexity of Data are evolved quickly over the years. There are an increasing number of items that must be kept under control. Due to these difficulties, researchers found that was necessary to create a framework which makes it simpler to completely comprehend data management and recognise the links between its components. Some models were built over the years to better understand Data. The most known models are DCAM, also known as Data Management Capability Assessment Model, and the DAMA-DMBOOK framework, which is into three different sections DAMA Wheel, Environmental Factors Hexagon, and The Knowledge Area Context Diagram.

The first one is the industry standard framework for data management. DCAM specifies the skills needed to build, enable, and sustain a mature Data Management discipline. It tackles the strategies, organisational structures, technology, and operational best practices required to successfully drive Data Management across business and ensure

data can support digital transformation, advanced analytics like AI and ML, and data ethics.

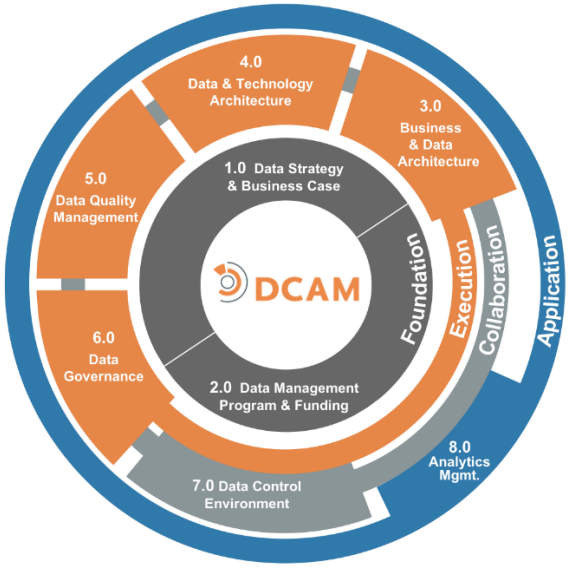


Figure 1 DCAM – the Data Management Capability Assessment Model – framework[1].

Data management is critical for obtaining information from analysis, feeding our models with certainty, improving client service, and capitalising on new business prospects. DCAM gives the assistance required to analyse your data programme's present state and provide the desired state's objectives. From the picture is possible to see that the DCAM is divided into eight major components.

1. **Data Management Strategy & Business Case** explains the components of a good data strategy, why they are essential, and how the organisation should be structured to accomplish them.
2. **Data Management Program & Funding:** tackles the development of the business case, the funding model that goes with it, and the necessity of including top executives and important stakeholders in the approval process.
3. **Data Architecture:** focuses on the core concepts of "data meaning" – how data is defined, described and related.
4. **Technology Architecture:** focuses on the relationship of data with the physical IT infrastructure needed for operational deployment.

[1]Data Management Capability Assessment Model (DCAM), EDM COUINCIL. https://dgpo.org/wp-content/uploads/2016/06/EDMC_DCAM_-_WORKING_DRAFT_VERSION_0.7.pdf

5. **Data Quality:** refers to the concept of fit-for-purpose data and the processes associated with the establishment of both data control and data supply chain management.
6. **Data Governance:** defines the operating model and the importance of policies, procedures and standards as the mechanism for alignment among (and compliance by) stakeholders.
7. **Data Management Program:** discusses what's organizationally needed to stand up a sustainable Data Management Program.
8. **Data Operations:** defines the data lifecycle process and how data content management is integrated into the overall organizational ecosystem.

The components are structured into 35 capabilities and 109 sub-capabilities. These capabilities and sub-capabilities are the essences of the DCAM. They define the goals of Data Management at a practical level and establish the operational requirements that are needed for sustainable Data Management.

The second which was analysed is the DAMA-DMBOOK framework, which is into three different sections DAMA Wheel, Environmental Factors Hexagon, and The Knowledge Area Context Diagram. This framework is the most used in the market. The scope of this framework is to analyse more in-depth the Knowledge areas which represent the overall scope of Data Management. As stated before, the framework is divided into three parts:

1. **DAMA Wheel**

The Data Management Knowledge Areas are defined by the DAMA Wheel. It places data governance at the centre of Data Management operations since governance is necessary for function consistency and balance. The remaining Knowledge Areas are evenly spread around the Wheel. They are all necessary components of a mature Data Management function, although they may be deployed at different times depending on the needs of the company.

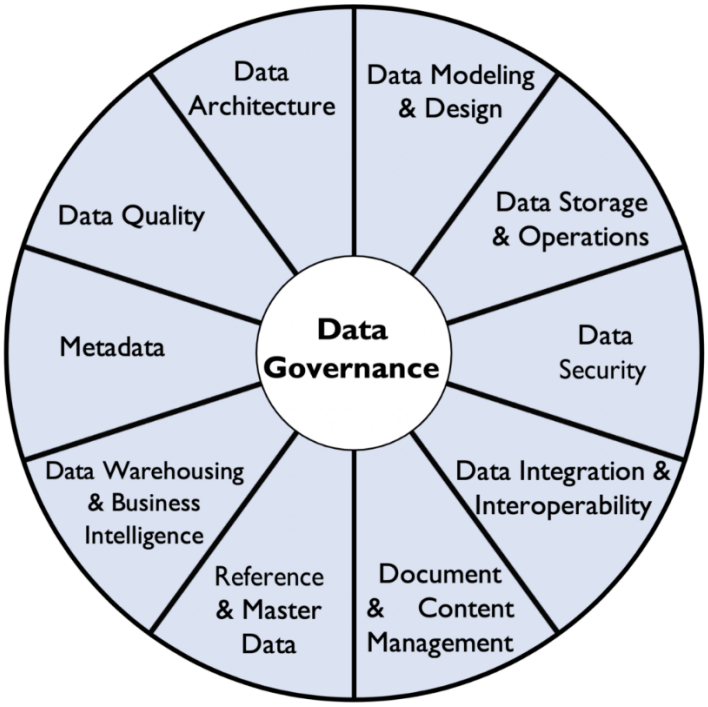


Figure 2 DAMA Wheel[2]

2. Environmental Factor Hexagon

The Environmental Factors hexagon depicts the interaction of people, processes, and technology and serves as a key to understanding the DMBOK context diagrams. It prioritises goals and principles because they guide how individuals should carry out tasks and effectively employ the instruments necessary for successful Data Management.

[2] DAMA BOOK, https://www.dama.org/cpages/dmbok-2-wheel-images

Figure 3 Environmental Factors Hexagon[3]

3. Knowledge Area Context Diagram

The Knowledge Area Context Diagrams illustrate the specifics of the Knowledge Areas, such as people, processes, and technology. They are based on a SIPOC diagram, commonly used in product management (Suppliers, Inputs, Processes, Outputs, and Consumers). Context Diagrams prioritise actions because they provide outputs that fulfil the needs of stakeholders. Each context diagram begins with the description and goals of the Knowledge Area. Plan (P), Develop (D), Operate (O), and Control (C) are the four phases of activities that drive the goals (centre) (C). The Inputs and Suppliers are on the left side. Consumers and deliverables are on the right. Below the Activities is a list of participants. The Inputs and Suppliers are on the left side. Consumers and deliverables are on the right. Below the Activities is a list of participants. Tools, Techniques, and Metrics that impact parts of the Knowledge Area are listed at the bottom. The lists in the context diagram are illustrative rather than complete. Various items will apply to different organisations. The high-level role listings only comprise the most critical roles. This design may be customised to meet the demands of each business.

[3] DAMA Book, https://www.dama.org/cpages/dmbok-2-wheel-images

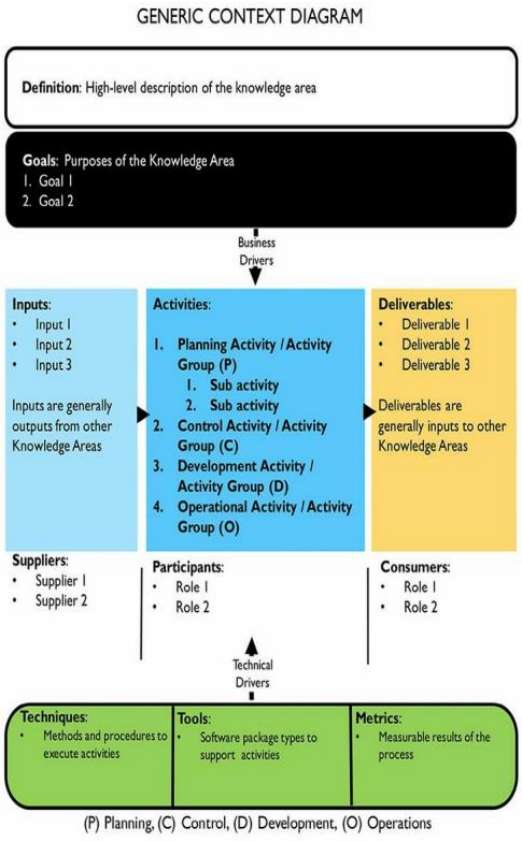


Figure 4 Knowledge Area Context Diagram generic[4]

These three different sections represent different parts of the framework. The DAMA Wheel provides the collection of Knowledge Areas at a high level, the Hexagon detects components of Knowledge Area structure, and the Context Diagrams present information inside each Knowledge Area. None of the existing DAMA Data Management framework components clarifies the link between the various Knowledge Areas. Over the years a lot of different researchers provide new interpretations of the DAMA framework using these three different sections. One example is Peter Aiken's framework which uses the DMBOK functional areas to describe the situation in which many organizations find themselves. An organization can use it to define a way forward to a state where they have reliable data and processes to support strategic business goals. But the most recent is the implementation of the DAMA Wheel by Sue Geuens. In this new vision, Sue recognizes that all other data management tasks are dependent on Business Intelligence and Analytical Operations. They are reliant on data warehouse and Master Data management programmes directly. However, those in turn are reliant on applications and feeding protocols. Trustworthy systems and applications are built

[4] DAMA Book, https://www.dama.org/cpages/dmbok-2-wheel-images

on techniques for reliable Data Quality, Data Design, and Data interoperability. Furthermore, data governance, which in this paradigm comprises reference data management, Metadata Management, Data Security, and Data Architecture, offers a framework on which all other tasks are based. Therefore, taking a cue from Sue's new framework, which focused on the dependencies between the various Knowledge Areas, we arrived at an evolution of the Checkers Wheel. This is the resulting framework.

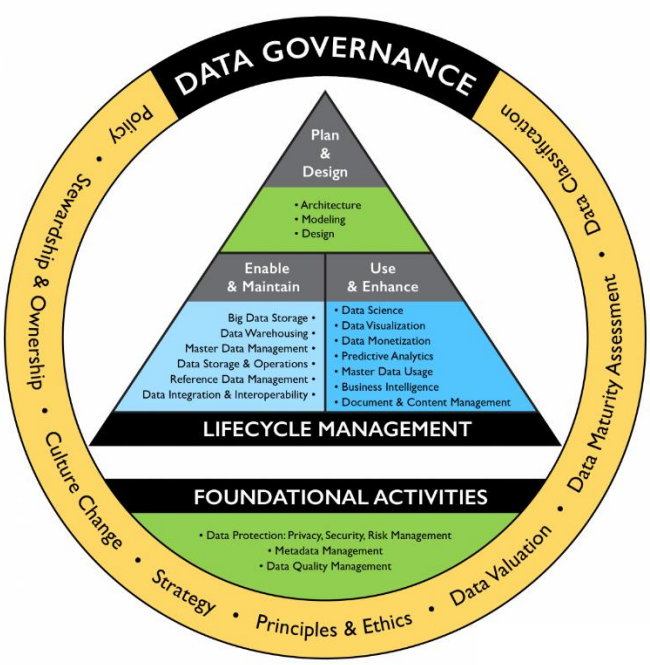


Figure 5 DAMA Wheel Evolved[5]

The core activities, such as metadata management, data quality management, and data structure development, are at the heart of this new interpretation of the DAMA Wheel. The intention in presenting different visual depictions of the DAMA Data Management Framework is to provide additional perspective and to open discussion about how to apply the concepts presented in DMBOK. As the importance of data management grows, such frameworks become useful communications tools both within the data management community and between the data management community and our stakeholders.

So, to decide which is the best framework will be presented a comparison between these two to put evidence of the differences. Looking at the models it is possible to evidence two main differences. Firstly, the nature of the building blocks of data management

[5] DAMA Book, https://www.dama.org/cpages/dmbok-2-wheel-images

models. In fact, according to the DAMA-DMBOK model, data management is a business function, and thanks to different Knowledge Areas the scope is made up. But it is impossible to find any definition of business function. So, the Knowledge Areas can be considered all the assumptions made concerning the relationships between key components, such as people or processes. This depends on the assumptions made. While DCAM bases its model on the concept of capability, but also in this case there isn't a definition of what capability is and which components constitute a capability. It is possible to assume that a business capability delineates what a business does without explaining how, why, or where it uses the capability. So, the DCAM 'capability' clearly does not fit the definition and a capability definition remains a mystery. So, it is possible to conclude that DAMA-DMBOK and DCAM base their data management models on the concepts of Knowledge Area and Capability correspondingly. These concepts are not clearly defined and are not compatible with each other.

The second one is regarding the relationship between data management and IT function. DCAM spreads the perspective of data management experts. The model outlines the duties of data management specialists. The data control environment includes the IT function. Data management experts should make sure to work along with IT experts. While DAMA-DMBOK looks at Data Management from the viewpoint of the enterprise. Thus, there are at least two perspectives on data management that depend on the relationship between data management and the IT function. The first is a broad viewpoint, looking at the lifetime of data as it moves around an organisation. This is how the DAMA-DMBOK model operates. The second viewpoint, from the point of view of the duties that data management specialists must do, is the most limited. This strategy is used in the DCAM model.

The best framework between DCAM and DAMA depends on the specific use case and requirements. It is not possible to determine which one is better in general as they serve different purposes. But to sum up, is possible to say that DCAM is used to automate IT infrastructure and operations within a data centre, is a memory architecture for deep neural networks, enables parallel processing of memory access operations, and improves the speed and efficiency of deep learning models, while DAMA is a professional association focused on data management and refers to a set of architectural principles for designing data management systems, it focuses on efficient data storage and retrieval, and supports the integration of multiple data sources. In this analysis,

there will be selected subcategories of Knowledge Areas and new Trends related to Data Management. The selected topics are Master Data, Data Profiling, Data Visualization, Data Observability, DataOps, Data Discovery, Data Pipeline, Metadata, Data Lineage, Data Catalog, Data Integration, Data Governance, and Data Quality.

2.2 Data Governance

Data Governance (DG) is defined as the exercise of authority and control (planning, monitoring, and enforcement) over the management of data assets. All organizations make decisions about data, regardless of whether they have a formal Data Governance function. Those that establish a formal Data Governance program exercise authority and control with greater intentionality (Seiner, 2014).

Data Governance represents the centre of the DAMA Wheel. In fact, without coordinated Data Governance is impossible to run a correct business. All other Data Management functions are guided by the Data Governance function. The goal is to guarantee that data is correctly handled by rules and best practices. While the goal of Data Management is to guarantee that an organisation derives value from its data, Data Governance focuses on how data choices are made and how people and processes are expected to act about data.

The main function of Data governance is to help companies to create policies and procedures. At the same time, it is used to demonstrate the benefits of better data management and the behaviours needed to reach these goals. But it is not simple to take use of all the advantages that data offers. Without an inclination to change the team and also the possibilities in terms of money and background, in the sense of the company's willingness to change, even the strongest data strategy, data governance, and data management strategies will fail. Data governance does not exist for its own sake. It must be perfectly in line with corporate strategy. People are more willing to alter their behaviour and adopt governance methods when it is evident how it helps address organisational challenges. So, Data Governance changes over time. In fact, due to the continuous changing in data management is important to monitor constantly the strategy. For doing this is necessary to create a Data Governance team. This will be a line organisation with certain tasks or a virtual organisation. To be effective, the roles and activities within data governance need to be well understood. They should be built around an operating framework that functions well in the organization. A Data Governance program should consider distinctive organizational and cultural issues and the specific Data Management challenges and opportunities within the organization.

Companies need to change their perspective. They must alter how it transforms strategy into action to become data-centric. Data are no longer seen as an afterthought of

operations and programmes. Effective Data Management becomes a primary responsibility as organisations work to base choices on the information learned from the analysis. Consequently, Data quality assurance is one of the goals of business operations.

So, Data governance actions overcome organisational and system constraints to allow for an integrated view of the data. A clear understanding of who and what is being handled is crucial for successful Data Governance. Also, perspective is important. If Data Governance is seen as a project that is just focused on a single functional area is not that good. It is important to consider it as a company-wide initiative. Establishing the enterprise idea is typically required to specify the scope of Data Governance inside an organisation. Thus, a strategy needs to be defined. It is important to establish the strategy in terms of how it relates to IT and data management plans as well as the broader company strategy. It is not a consequential process, but an iterative one. It is necessary to implement it daily to develop and approve new parts. But is not so easy to define this strategy. There is no technical solution to this problem since the focus of Data Governance is organisational behaviour. However, several tools help with the entire procedure. A Data Governance programme must handle its tasks and data appropriately. Tools can be useful for both these tasks and the measurements that go along with them. Before choosing a tool for a particular function, such as a business glossary solution, an organisation should define its basic governance goals and requirements. A crucial Data Governance tool is a business glossary. It includes accepted definitions of business jargon and its data connections. There are numerous business glossary tools available, some as standalone programmes and others as parts of larger ERP systems, data integration tools, or metadata management tools.

2.3 Data Integration & Interoperability

Data Interoperability will not be examined in this section; only Data Integration will. The process of merging data from several source systems to produce uniform sets of information for both operational and analytical usage is known as Data Integration. One of the key components of the overall Data Management process is integration, whose main goal is to create consolidated data sets that are accurate and consistent and satisfy the information requirements of various end users within an organisation. Data after the Integration is used in different areas to improve the business. For example, can be supplied into transaction process systems to power business applications, or put into Data Warehouses to provide analytics or corporate reporting. To do this there are a lot of techniques. The reason why the importance of Data Integration is constantly growing is due to the difficulty to manage data between the different databases. The movement of data between different databases is called Data Mobility. The latter is one of the driving drivers behind Data Integration. Due to the increasing number of databases, stores, and databases, every information technology organisation is now in charge of managing the processes for moving data between the data stores inside the company and to and from other organisations. Therefore, it's crucial to manage the data-moving process correctly. The most known technique that Data Integration utilizes is ETL. ETL refers to the process of extracting, transforming, and loading data. There are several procedures to perform this action, including digitally, physically, in batches, or in real-time. ETL can be carried out as a regularly planned event and is called batch ETL, or anytime fresh or updated data is available, depending on Data Integration and this one is called real-time or event-driven ETL. Now will be examined each stage of the ETL process.

1. **Extract**:

The extraction procedure entails choosing the necessary data and removing it from its source. The extracted data is then staged, either in memory or on physical data storage on a disc. The source data store, the destination data store, or both may be put close by if the staging data store is physically staged on a disc. If this process operates on an operational system, it should use as minimal resources as possible to prevent having a negative influence on the operational operations.

2. Transform:

The transformed data is made compatible with the destination data store's structure through the transformation procedure. The transformation includes instances where data is replicated to several destinations, where data is used to trigger events but is not persistent, and when data is removed from the source as it goes to the target. It is possible to do the transformation in batches or in real-time, and then either physically store the results in a staging area or virtually store the converted data in memory until it is time to proceed to the loading stage. The transformed data should be prepared to be integrated with the target structure's data.

3. Load:

The load stage of an ETL process involves physically presenting or storing the transformed data in the destination system. Depending on the transformations carried out for the intended use of the target system and the planned application, the data may need extra processing to be merged with other data or it may be in a finished condition, ready to be given to clients.

So, after that, the ETL process is finished. Another technique that is equivalent to the ETL is extracted, loaded, and transformed (ELT). When the destination system or an intermediary application system cannot modify data as effectively as the source system, this is frequently used. This is done so that data may be converted using this process typology after being loaded into the target system. Usually, is preferred this technique due to the huge amounts of data. Data warehouses need to process different types of data and there will come problems. Using an ETL technique to manage millions of records in these new formats may be time-consuming and expensive. There are several advantages to ELT, including management simplification: ETL isolates the various tasks by minimising the relationships between these processes, hence lowering risks. Alternatively, ETL may be built for usage in Data Lake and Data Warehousing systems. Regarding the difference between these two techniques, the main one is the order of operations. ELT copies or exports all the data from the source location, but it loads the raw data directly to the target data store without any transformation required, so these must be handled after the loading phase, as the name explains. Even if, both processes

leverage a variety of data repositories, such as databases, data warehouses, and data lakes, each process has its advantages and disadvantages.

For example, ELT is particularly useful for high-volume, unstructured datasets as loading can occur directly from the source. Another strength is the fact that ETL doesn't need much upfront planning for data extraction and storage, and this reason is ideal for big Data Management. While regarding the ETL process, on the other hand, there is a necessity to provide more definitions. Specific data points need to be identified for extraction along with any potential "keys" to integrate across disparate source systems. So consequently, these are the processes to modify data that will be used for companies. But Data Integration is not only ETL or ELT. Data Integration is meant also the process of delivering information when, and in the format that is required. So, to have a good Data Integration, it is critical to have a well-organized plan. Data Integration implements two important processes: Data Discovery and Data profiling.

Data discovery is the process of navigating or applying advanced analytics to data to detect informative patterns that could not have been discovered otherwise the goal of Data Discovery is to identify potential sources of data for the data integration effort. So, Data Discovery is a mix of reviewing existing documents, verifying information gathered against the actual data through data profiling or other analysis, and interviewing subject matter experts Data Discovery helps the company to identify where data might be acquired and where it might be integrated.

Data Profiling refers to the procedure of inspecting, analysing, evaluating, and summarising data sets to gather knowledge about the data's quality. As with high-level Data Discovery, Data Profiling includes verifying assumptions about the data against the actual data. Capture results of Data Profiling in a Metadata repository for use on later projects and use what is learned from the process to improve the accuracy of existing Metadata.

2.4 Reference and Master Data

While transactional data is at the centre of data-driven organisational goals, the availability and quality of Reference and Master Data are critical for making such transactional data work effectively. The data used to describe and classify other data are known as Reference Data. While Master Data is information about company entities like customers and goods. The context required for business interactions is provided by Master Data. Improving Reference and Master Data availability and quality impacts overall data quality and business confidence in data. These solutions provide benefits such as increased efficiency and productivity, and the ability to improve customer experience. Master Data can be seen as an aggregation of Reference Data, Enterprise structure data, and Transaction structure data. The first object of Master Data is entity resolution. This refers to the discerning and managing associations between data from different systems and processes. But, to clarify the differences between these three is possible to have an example. All data that are used to characterise other data in an organisation is referred to as Reference Data. Enterprise Structure Data, on the other hand, refers to all of the frameworks that allow company operations to be recorded by corporate responsibility. Then, Transaction Structure Data encompasses all data that describes the pieces that must be present for a transaction to take place.

Chisholm's definition distinguishes Master Data from transaction activity data records details about transactions, and transaction audit data describes the state of transactions, as well as Metadata, which describes other data (Chisholm, 2008). Various definitions have been presented over the years. Another definition made by David Loshin in 2008 of Master Data objects believes them to be essential business objects utilised in many applications within a company, together with their related Metadata, attributes, definitions, roles, linkages, and taxonomies. Now will be presented in two different sections more in-depth Reference and Master Data.

Reference Data

Reference Data is any data used to characterize or classify other data, or to relate data to information external to an organization (Chisholm, 2001). The complexity of Reference Data depends on the data itself. For example, a basic one can consist only of a code with a description. While, a complex, can incorporate a map and hierarchy.

Reference Data are different depending on the company's needs. But there are some techniques which are standard for all of them. These are for example code tables in relational databases, linked via foreign keys to other tables to maintain referential integrity, or systems which maintain business entities, allowed, future-state, or deprecated values. An example of Reference Data can be a list. The management of these Data is called Reference Data Management. In this management can be found all parts related to the control and maintenance of established domain values, definitions, and connections inside and across domain values. The purpose is to guarantee that values are consistent and up-to-date across functions and that the data is available to the company. Nowadays, data changes every day, so the relevance of data is greater, but it is also critical to keep data that is consistent with our company's current condition. The volatility of data must also be addressed.

Master Data

Master Data are information about corporate entities that provide context for transactions and analysis and should comprise the most up-to-date and correct information about critical business organisations. When Master Data values are managed correctly, they may be trusted and used with confidence. Business standards are typically used to establish the Master Data structure and permitted value ranges. Master Data Management is defined as a technology-enabled discipline in which business and IT collaborate to assure the enterprise's official shared Master Data assets' uniformity, correctness, stewardship, semantic consistency, and accountability. One problem that this topic faces is the different ways how people represent similar concepts. A concept can be explained in different ways, but a dataset needs to be coherent and precise. Furthermore, data sometimes change over time and is not always easy to fulfil the technological know-how to account for these changes. The objective is to minimise the possibility of confusing identifiers while ensuring the availability of correct and current information. As a result, each business has unique Master Data management motivations and challenges. This is determined by the system's typology, the business processes it supports, and how data is used for both transactions and analytics. These factors can sometimes assist the organisation in identifying new possibilities to improve customer service and operational efficiency.

But to do all these things Reference and Master Data Management need tools. Master Data Management needs tools made expressly to support identity management. Data integration tools, data remediation tools, operational data stores, data sharing hubs, or specific apps can all be used to achieve Master Data Management. There are also solutions made by several vendors in the market. These are available and can address one or more of these topic areas. But there are also other companies which encourage customers to build their Master Data solutions. By incorporating these services, businesses may address unique objectives while utilising best-of-breed solutions that are integrated into their whole company architecture.

2.5 Metadata

The most common definition of Metadata is “data about data,”. Metadata can refer to a wide range of various sorts of information. Metadata encompasses data rules and constraints, logical and physical data structures, and technical and economic operations. It describes both the data and the concepts it represents, as well as the links that exist between the two. Metadata may help an organisation better understand its data, systems, and workflows. It allows for the evaluation of data quality and is essential for managing databases and other applications. It facilitates the processing, maintenance, integration, security, auditing, and control of other data. An organisation cannot manage its data as an asset without Metadata. A company might not be able to manage its data at all without Metadata. Consequently, Metadata should be viewed as data. This line is conceptually related to the level of abstraction provided by data representations. This, however, might be arbitrary.

For these reasons different kinds of Metadata exist. Now will be presented the most known. Business metadata comprises data governance information and focuses on the content and status of the data. Inside this category, there are a lot of concepts. For example, it includes attribute data types and other attribute properties, or calculations. While, Technical Metadata provides information about the technical details of data, the systems that store data, and the processes that move it within and between systems. Examples of Technical Metadata include Physical database tables and column names Column properties Database object properties. Operational Metadata describes details of the processing and accessing of data. For example, Logs of job execution for batch programs, History of extracts and results, and Schedule anomalies.

There are also other typologies of Metadata. But to consider Data as Metadata is important to clarify another difference. Metadata needs to be registered. There can be registered as Structured or Unstructured. The first one is related to a set of principles which Metadata needs to respect to be in this category. Consequently, if the conditions are not met they are called Unstructured. But Unstructured Metadata is not only all the Metadata which are not Structured. Unstructured data is any data that is not in a database or data file, such as documents or other media.

Most businesses, however, do not manage Metadata properly at the application level, because Metadata is frequently generated as a by-product of application activity rather

than as a product. Metadata, like other types of data, requires a significant amount of preparation before it can be merged. Metadata strategy may be summarised as how an organisation wants to manage them and how it will transition from current to future state practices. The strategy defines the organization's future state enterprise Metadata architecture as well as the phases of execution necessary to fulfil strategic objectives. It is important also to consider the quality of Metadata. Low-quality Metadata is useless for companies. And the quality is constructed with highly relatable Metadata. To have high-quality Metadata is necessary to plan everything. Starting from integration procedures collect and combine Metadata from across the company, including Metadata collected from outside sources.

2.6 Data Quality

Data quality is something which sometimes is underestimated. But nowadays poor-quality data are useless. Data Quality is something that is inside Data Management. The capacity to create data for applications, store and access it securely, distribute it correctly, learn from it, and make sure it satisfies business needs are all parts of Data Management. All of these contribute high-quality data. The purpose of all Data Management disciplines should be to produce high-quality data that helps the company since all data management disciplines contribute to the quality of data. Therefore, Data Quality must be used in more than one field. Data quality is connected to everything. But how to manage data correctly depends on the organization. The latter should be aware of the difficulties of building a correct plan and procedures. But Data Quality depends on all who interact with data. Consequently, if there are perfect procedures, a well-managed plan, and secure and correct access to the data, but there are people who don't know how to use these data in the correct way all these measures are useless. Thus, companies that do not properly manage Data Quality will figure out more problems than those that manage them.

So is necessary to create a specific team which manages all these problems. The Data Quality team oversees enlisting both technical and business Data Management experts and leading the effort to apply quality management approaches to data to guarantee that data is suitable for consumption for a range of reasons. The group will probably work on several projects that will allow them to create procedures and best practices while solving crucial data-related problems. Is important to clarify that Data Quality depends on the context in which the company uses data, but also on the request of the data consumer. The fact that expectations about the quality are not always understood presents one of the obstacles in maintaining the quality of data. These expectations can be measured through some "Dimension". A Data Quality dimension is a measurable feature or characteristic of data. Dimension refers to the relationship between dimensions in physical measurement. Data quality dimensions provide a vocabulary for defining data quality requirements. But as stated before these dimensions depend on heavily context or on subjective. So is possible to define what is Data Quality. ISO 8000 which is the international standard for data quality, provides a definition: "The ability to create, collect, store, maintain, transfer, process and present data to support business processes in a timely and cost-effective manner require both an understanding of the

characteristics of the data that determine its quality and an ability to measure, manage and report on data quality." The purpose of ISO 8000 is to assist organisations in defining what constitutes quality data, enabling them to request it using standard procedures, and confirming that they have received it using the same procedures.

The benefits of Data Quality are different. Financially speaking, preserving high data quality standards helps businesses to lower the cost of finding and resolving inaccurate data in their systems. Additionally, businesses can prevent operational blunders and disruptions in company processes, which can raise operating costs and lower revenues. Another benefit is that efficient data quality management allows data management teams to concentrate on more beneficial duties rather than cleaning up data sets. This led to a more consciousness regarding the data and consequently to better business decision-making which led to better results. Consequently, these benefits help companies to reduce the waste of time and resources. For example, there could be a situation where the dataset of the customers is incomplete, so this can lead to a wrong marketing campaign due to the possibility to send promotions to someone who is not interested in that topic. Or in another case, with an incomplete dataset is difficult also to work on it due to missing values. These last two problems are a few examples of the issues that Data Quality need to manage. Another issue is related to the relationship between the data inside a database. Precisely to lack of referential integrity. Referential integrity is referred to all data records that have a referencing counterpart. So, these lacks are all the situations where there is for example an ID but without the corresponding data. So all the processes which manage Data Quality are important. They may have negative consequences if employed without awareness of the data. Members of the Data Quality team should collaborate with development teams to guarantee that threats to data quality are handled and that the organisation makes the most of the ways that efficient modelling and data processing may enable higher-quality data.

2.7 Data Management Trends

In the modern era of technology, businesses are required to find effective means to acquire, store, and manage their data. The advancement in remote and hybrid business models has led to the necessity to shift towards cloud storage. This has necessitated the need for solutions that are scalable and flexible enough to cater to the varying needs of businesses. However, with the continuous evolution of technology, new trends are constantly emerging, and they are increasingly becoming essential for businesses to stay relevant and competitive. Therefore, it is crucial to be aware of these trends, which will now be presented.

2.7.1 Data Catalog

Data Catalog have swiftly established themselves as essential to contemporary data management. Successful data catalogue implementations have a tremendous impact on the quality and efficiency of data analysis as well as the engagement and excitement of those who need to conduct it. It helps data experts and business users in finding pertinent data for analytical purposes. A Data Catalog consists of a software application that creates an inventory of an organization's data assets. It is driven by Metadata which helps the company to provide contextual information about data assets. But not only Metadata drives Data Catalog. These are combined with data management and search tools to help users find what they are looking for in the catalog. Data Catalog usage is increasing as firms increasingly rely on data analytics to drive business strategy and operations. Catalogs are currently an essential component of many data management systems. Consequently, Data Catalogs help companies manage data more easily and make it possible to discover all relevant data for analysis applications. Another benefit which brings a good Data Catalog is productivity. If users can find data easily and precisely for their request, time to work is reduced. In addition, duplicates are deleted, so that when someone selects the data, they are the correct ones.

2.7.2 DataOps

Another new trend is DataOps. It is also known as Data operations, it is a Data Management technique used throughout an organisation to promote communication, integration, and data flow automation among data administrators and consumers. The scope is always to minimize the time and increase the production of analytical datasets. Is important to take into consideration the fact that thanks to DataOps, Agile development is introduced to manage data in alignment with business goals. Based on ongoing user feedback, the team may review its priorities and more readily adjust to changing requirements. So, the environment is constantly changing, and companies need to always be ready to change their data and strategy. The growing importance of this can be seen also in the increasing volume of data which companies need to manage. Building a correct DataOps strategy has become critical. The first step in DataOps is to clean raw data and provide an infrastructure that makes it easily available for use. So, DataOps enhances data quality by automating formerly manual and error-prone tasks including data cleansing, transformation, and enrichment. Companies may make better decisions faster and with greater confidence as a consequence of enhanced data quality. As a result, also communication and cooperation across teams are implemented. Because they share the same DataOps architecture and approach, various teams may interact.

2.7.3 Data Observability

Another trend that is constantly growing over the years is Data Observability. This one can be seen as the capacity to analyse, diagnose, and manage data health across numerous IT technologies across the data lifecycle. Data management teams can utilize data observability approaches to monitor the quality, dependability, and delivery of data and highlight issues that need to be solved. The importance of this topic is grown due to the difficulties of Data Quality. The latter use tools to identify problems in data sets, but they find only specific problems, they don't control the entire health of the dataset. While Data Observability is intended to aid in the maintenance of healthy data pipelines, high data dependability, and timely data delivery, as well as to minimize any data downtime caused by mistakes, missing values, or other issues that render data sets worthless. This, help companies to maintain dataset more efficiently, thanks to the possibility to check issues and an easier way to find root causes. Data Observability also helps teams to keep a reliable data environment. It also allows for more effective problem-solving and provides contextual information for planning and prioritising repair operations. Consequently, to these two benefits, data downtime is minimized. Data teams can notice and analyse data outage events in real-time and take rapid remedial action. But to have good Data Observability is necessary to have DataOps and the tendency of the data professionals to learn and improve their methodologies. So, it is important to take into consideration also the challenges. One example could be the difficulty to find organizational support. Even if the team wants to adopt these new practices, they need support to implement the required elements. On the other hand, there could be a case where there is correct organizational support, and teams are motivated to implement Data Observability, but sometimes happen the data systems are not fully integrated. Connecting all an organization's systems may not be simple or feasible. If this is the case, there will be missing values so data pipelines could be incorrect or incomplete.

2.7.4 Data Pipeline

Data Pipeline is a mechanism which takes raw data from different data sources, sometimes processes these data, and after that, this tool transfers it to a data repository. The transformation could be the aggregation of two different data, or masking and filtering, in this way the tool assures proper data integration and uniformity inside the data repository. Sometimes Data Pipeline can be confused with ETL. But processes are different. An ETL extract data from a system after some transformations are made and lastly loads data into a repository. While a Data Pipeline carries out the same activities as an ETL but with advanced processes. Combines operational and business logic. Nowadays, the importance of Data Pipeline is growing. This is because the amount of significant data inside a data repository is constantly growing. Data sometimes are dispersed into different repositories and systems. This is a problem. Data Pipeline tries to solve that by trying to connect these systems to enable more in-depth analysis. By doing this, Pipeline facilitates human operations and permits the automatic flow of data from one step to the next. Therefore, these processes can lead to some benefits. For example, Data Pipeline helps companies to maintain high data quality. Thanks to transformations, data are always checked and modified if it is necessary before going inside repositories. Consequently, if there is high data quality, it is easier for the systems to put in some new data and provide some analysis. This is because all teams know how to use these pipelines and they can implement all the services with new data.

2.7.5 Data Visualization

The last new trend that is analyzed is Data Visualization. Gartner provides a definition: "Data visualization is a way to represent information graphically, highlighting patterns and trends in data and helping the reader to achieve quick insights." The main goal is to present information easier to read to identify trends and outliers in large data sets. Data Visualization is important for every company. It can be used to display trends over the last years and compare the different competitors with their products. Can be used also to visualize the outputs of predictive analytics or some machine learning algorithms. Thus, Data visualization is universal. It is easy to handle and permits to visualize through trends which factors affect customer behaviour or in other cases customer satisfaction. Thanks to this easy way to utilize Data Visualization there are some benefits linked to this topic. One of these is the advantage to choose a better strategy thanks to the visualization of the data. With Data Visualization there is the possibility to highlight problems but also new possibilities. Consequently, Data visualization leads to a more comprehensive and deep analysis of trends. Customer satisfaction can be checked and analyzed through this chart. But sometimes all these benefits cannot be used. This is because Data Visualization uses a dataset that companies give to a different tool. So, if the dataset has simplified data there is a situation where charts provided by the tool are useless for the company. So there is the necessity to check the meaningless of data before putting them inside Data Visualization tools. Another problem is that nowadays companies rely a lot on these charts. But if the analytical tools used to analyze this situation don't follow the technological advances can be problematic.

2.7.6. Data Lineage

Data lineage is a crucial aspect of data management that refers to the tracking of data from its origin to its final destination. It is the process of understanding and documenting the movement of data throughout its lifecycle. Data lineage provides a complete view of the data flow, including its source, transformation, and consumption. It is a critical component in data governance, compliance, and regulatory requirements. Data lineage helps organizations to track and understand how data moves through the system. It provides a comprehensive understanding of the data, including its origins, transformations, and final destinations. This understanding is essential for data governance, as it helps organizations to ensure that the data is accurate, consistent, and compliant with regulatory requirements. The data lineage process typically starts with the identification of the data sources. These sources could be internal or external to the organization, structured or unstructured, and stored in different formats. The next step is to identify the data processing steps that transform the data into a usable format. These steps could include data cleaning, data transformation, and data enrichment. Once the data is transformed, it is stored in a data repository, such as a data warehouse, data lake, or a database. The final step is to identify the data consumers, who use the data for analysis or decision-making. There are several benefits of data lineage. Firstly, it helps organizations to maintain data quality and accuracy. By tracking the movement of data, organizations can identify any inconsistencies or errors in the data and take corrective actions. Secondly, it helps organizations to comply with regulatory requirements. Regulatory bodies such as GDPR, HIPAA, and SOX require organizations to track the movement of data to ensure data privacy and security. Thirdly, it helps organizations to make better decisions by providing a complete view of the data flow. Organizations can use this information to identify areas of improvement, optimize processes, and reduce costs. There are several challenges associated with data lineage. Firstly, it can be complex and time-consuming, particularly in large organizations with multiple data sources and processing steps. Secondly, it requires a high level of collaboration between different teams, including data analysts, data scientists, and IT teams. Thirdly, it requires the use of specialized tools and technologies to track the movement of data. To overcome these challenges, organizations can use automated tools and technologies to track data lineage. These tools can automate the data lineage process, making it easier and more efficient. They

can also provide a centralized view of the data flow, making it easier to collaborate and share information across different teams. In conclusion, data lineage is a critical component in data management, providing a complete view of the data flow from its origin to its final destination. It helps organizations to maintain data quality, comply with regulatory requirements, and make better decisions. While data lineage can be complex and challenging, organizations can use automated tools and technologies to make the process more efficient and effective.

2.7.7 Data discovery

Data discovery is a vital process that helps organizations identify and analyze their data assets to gain insights into their business processes and gain a competitive edge. It is a challenging process that requires expertise, specialized tools, and technologies.To begin the data discovery process, an organization must first identify the business need for data. This involves understanding the specific problem or opportunity that the data will be used to address. Once the business need is established, data sources need to be identified. These sources can include internal data such as databases, data warehouses, and data lakes, as well as external sources like social media, market research, and government data. Next, it is essential to assess data quality. This step involves examining the data to ensure that it is complete, accurate, consistent, and timely. Issues with data quality can negatively impact the accuracy and reliability of insights generated from data analysis, so it is important to address them before proceeding. Once data quality is assured, the data can be analyzed using statistical analysis, machine learning, and other techniques. These techniques help identify patterns, relationships, and trends in the data, as well as identifying anomalies and outliers that could point to potential issues or opportunities. The insights generated from data analysis can be presented using data visualization tools, making them easy to understand and act upon. Finally, the insights gained from data analysis must be shared with stakeholders. This can involve creating reports, dashboards, or other visualizations that communicate insights in a clear and actionable way. Effective communication is crucial to ensure that stakeholders can use the insights to drive business value. Automated data discovery tools and technologies can make the process of data discovery more efficient and effective. These tools help automate the identification and analysis of data, providing a centralized view of an organization's data assets. By reducing the time and effort required for data discovery, organizations can gain insights more quickly and make informed decisions based on data-driven insights. Data discovery is a critical part of data management that enables organizations to leverage their data assets for a competitive advantage. The process involves identifying the business need for data, identifying data sources, assessing data quality, analyzing data, and sharing insights with stakeholders. Automated data discovery tools and technologies can help organizations overcome challenges and make the process more efficient and effective.

2.7.8 Data Profiling

Data profiling is an important component of data management that involves examining and analyzing data to gain a better understanding of its characteristics and quality. The goal of data profiling is to assess data quality, identify data issues, and understand the relationships between data elements. The first step in data profiling is to identify the data to be profiled. This involves understanding the business need for the data and identifying the data sources that contain the data of interest. Once the data sources have been identified, the next step is to collect the data and prepare it for analysis. Data profiling typically involves a range of techniques and methods, including statistical analysis, data visualization, and machine learning. Statistical analysis techniques can be used to identify patterns, relationships, and trends in the data. Data visualization tools can be used to present the data in a clear and understandable way, making it easier to identify patterns and outliers. Machine learning techniques can be used to automate the process of data profiling, making it faster and more efficient. During the data profiling process, data quality issues are identified and documented. These issues may include missing data, inconsistent data, or incorrect data. The goal is to identify any data quality issues that could impact the accuracy and reliability of insights generated from data analysis. Data profiling also involves understanding the relationships between data elements. This includes identifying dependencies between data elements and understanding how they are related to each other. This information is critical for developing accurate data models and ensuring that data is properly integrated into business processes. Finally, the results of the data profiling process are documented and communicated to stakeholders. This documentation typically includes a summary of data quality issues, data relationships, and any other relevant insights generated from data analysis. Effective communication is critical to ensuring that stakeholders can use the insights gained from data profiling to make informed decisions and improve business processes. In conclusion, data profiling is a critical component of data management that enables organizations to gain a better understanding of their data assets. It involves a range of techniques and methods, including statistical analysis, data visualization, and machine learning. The goal of data profiling is to assess data quality, identify data issues, and understand the relationships between data elements. By documenting and communicating the results of the data profiling process, organizations can use the insights gained to make informed decisions and improve business processes.

2.8. Evolution of Data Management and emerging roles

The evolution of technology in the management of data, particularly with the advent of cloud computing, has led to the emergence of new competencies in the field of data management. Cloud computing allows users to store their data remotely on servers owned and managed by third-party companies, rather than on physical hardware located on-premises. This means that businesses no longer need to invest in expensive physical hardware to store their data. Instead, they can rent storage space on cloud servers, which can be easily scaled up or down as their needs change. These competencies are essential for businesses to make sense of the vast amounts of data that they generate and to leverage it for strategic decision-making. One of the most significant new competencies to emerge is data analytics. Data analytics involves the use of statistical and computational methods to extract insights from large datasets. Data analysts are responsible for analyzing data to identify trends, patterns, and relationships that can inform business decisions. They use a variety of tools, including statistical software and machine learning algorithms, to analyze data and generate insights. Another important competence that has emerged is data engineering. Data engineering involves the design, building, and maintenance of the infrastructure required to store process, and analyze large amounts of data. Data engineers work with data scientists and analysts to ensure that data is collected, stored, and processed efficiently and securely. They also ensure that data is stored in a way that allows for easy and fast access by other stakeholders in the organization. However, cloud computing also poses some challenges in terms of security and data privacy. Businesses need to ensure that their data is stored and managed securely and that they comply with relevant regulations and standards such as the General Data Protection Regulation (GDPR) and the Health Insurance Portability and Accountability Act (HIPAA). The emergence of these new competences has also led to the creation of new roles in data management. Some of these roles include Data Engineer, Data Analyst, Data Architect, and Data Scientist. These new roles in data management reflect the growing importance of data in organizations and the need for specialized professionals to manage this critical resource.

2.8.1 Data Engineer

A Data Engineer is responsible for designing, building, maintaining and troubleshooting the infrastructure needed for optimal extraction, transformation, and

loading of data from a variety of sources into the data analytics systems. The goal is to ensure the availability, efficiency, and reliability of data for the Data Scientists and Analysts to work with. Key responsibilities include data ingestion, data storage, data processing, and data governance. To succeed as a Data Engineer, you need a strong technical background, including proficiency in programming languages such as Python or Java, as well as experience with big data technologies and data storage solutions. In addition, you need excellent problem-solving skills, attention to detail, and the ability to work with cross-functional teams. A Data Engineer plays a critical role in ensuring that data is properly managed and processed so that it can be effectively analyzed and utilized to drive business value

2.8.2 Data Analyst

A Data Analyst is a professional responsible for using data to gain insights and support decision-making in an organization. They analyze large and complex data sets to uncover patterns, trends, and relationships that can inform business decisions. Data analysts conduct exploratory data analysis to uncover insights and identify areas for further investigation, in other cases, they apply statistical methods to data to gain insights and make predictions. This figure utilizes also Data visualization to predict and present data in a clear and understandable format. Consequently, it interfaces with businesses in order to communicate the results of data analysis to stakeholders and make recommendations based on the findings. To succeed as a Data Analyst, you need strong analytical and statistical skills, as well as proficiency in data visualization and data manipulation tools. Additionally, you need excellent communication and interpersonal skills, as well as the ability to translate complex data findings into actionable insights for stakeholders. So a Data Analyst plays a crucial role in helping organizations make informed decisions based on data and contributes to the success of data-driven decision-making in the organization.

2.8.3 Data Architect

A Data Architect is a professional who is responsible for designing and implementing the overall data architecture for an organization. The goal of a Data Architect is to ensure that the organization's data assets are properly aligned with the business strategy and that they support the organization's data needs. Data Architect is responsible for a lot of things. One of these is for example the design and implementation of efficient

and scalable data warehouses that support the organization’s data needs. Consequently, It ensures the performance and scalability of the data architecture. Another important task which needs to handle is the definition of the overall data strategy for the organization and ensuring it aligns with the business strategy. So it represents a sort of bridge between the business and the organization. To succeed as a Data Architect, you need a strong technical background, including expertise in data warehousing, data modeling, and data integration. Additionally, you need excellent leadership skills, as well as the ability to work with cross-functional teams and stakeholders to ensure that the data architecture supports the needs of the organization. A Data Architect plays a critical role in ensuring that an organization's data assets are properly aligned with its business strategy and that they support the organization's data needs.

2.8.4 Data Scientist

A data scientist is a professional who works with data to extract insights, make predictions, and drive decision-making. They have a diverse set of skills, including knowledge of statistics, programming, data visualization, and machine learning, as well as business acumen and strong communication skills. The role of a data scientist is to take raw data and turn it into actionable information that can be used to inform business decisions. This typically involves collecting, cleaning, and organizing data, developing models and algorithms to analyze data, and presenting findings in a clear and concise manner to stakeholders. Data scientists work in a variety of industries, including technology, finance, healthcare, and retail, and they may be involved in a wide range of tasks, from identifying trends and patterns in data to developing predictive models to automating data-driven processes. In order to be successful as a data scientist, it's important to stay up-to-date with new technologies and trends in the field and to have strong problem-solving and critical thinking skills. A strong understanding of statistics and programming languages, such as Python and R, is also essential.

Data Scientists and Data Analysts both play important roles in the field of data analytics, but their duties and skill sets differ. Data analysts deal mostly with structured data that has already been collected and stored in databases. They clean and analyse data using statistical and analytical tools such as SQL, Excel, and BI tools, and then build reports and dashboards to convey the results to stakeholders. They are concerned with comprehending historical performance, finding opportunities for development,

and making data-driven decisions. Data Scientists, on the other hand, deal with massive, complicated datasets, which are frequently unstructured or semi-structured, and process and analyse the data using computer languages such as Python and R, as well as big data platforms such as Hadoop and Spark. Their major focus is on building predictive models and algorithms that can aid in the resolution of difficult business challenges, the optimization of business processes, and the identification of new prospects. They employ machine learning algorithms to create prediction models and offer suggestions based on data analysis, and their mission is to deliver future insights.

Data Scientists must also be proficient in complex statistical techniques like as linear regression, time series forecasting, and predictive modelling, as well as be able to communicate their results to stakeholders in a clear and straightforward manner. They frequently collaborate with other departments, such as marketing, finance, and operations, to provide solutions that may propel corporate development and success. Hence, while both Data Scientists and Data Analysts work with data, their roles, skill sets, and objectives are distinct. Data Analysts analyse historical performance and find areas for improvement, whereas Data Scientists use prediction models and algorithms to address complex business challenges and drive corporate development. Both jobs are critical in data analytics and contribute to organisational performance in various ways.

These new roles bring fresh perspectives and unique skill sets to the field of Data Management, and their contributions can lead to increased efficiency, improved decision-making, and better overall performance. the evolution of technology in data management has created new roles and competencies that are essential for organizations to manage data efficiently and effectively. These new roles bring unique skill sets and fresh perspectives to the field of data management, and their contributions can lead to increased efficiency, improved decision-making, and better overall performance. As organizations continue to collect and store vast amounts of data, it is crucial to have professionals with the necessary skills to manage and analyze this data. The emergence of new roles such as Data Analysts, Data Scientists, and Data Engineers shows how data management has become a specialized field, and the demand for these professionals will continue to grow as technology continues to evolve. These professionals are tasked with developing and implementing data strategies, ensuring data quality, analyzing data to derive insights, and maintaining data privacy and security. In summary, the creation of new roles and competencies in data management

reflects the growing importance of data as a strategic asset for organizations, and the need for specialized skills to manage it effectively.

Chapter 3: Methodology

The question was given on how firms have grown technologically through time and what new sorts of services they have introduced. To do this, a subset of the market's firms was examined. They were chosen after careful consideration. This chapter will go through the processes that were taken to arrive at the final analysed dataset.

The initial step was to research the themes for the literature. This investigation began by exploring alternative frameworks other than the DM-BOOK. The DCAM was examined in several articles. Following that, a search for the various categories in the DAMA Wheel was conducted. The emphasis was mostly on papers involving Data Governance since it is the central pillar of the DAMA Wheel. This is due to the latter serving as the focal point for Data Management efforts. This area is required for uniformity and balance throughout the Knowledge Areas. The other Areas (data architecture, data modelling, etc.) are balanced around the wheel. They are all necessary parts of a mature Data Management function but can be implemented at different times, depending on the requirements of the organisation.

As a result, the remaining Knowledge Areas were examined. Other articles, in addition to the DAMA book, were utilised to build on their definitions. Following that, the Hexagon framework was examined, which recognises the components of the structure of Knowledge Areas, while the Context Diagrams offer the specifics inside each Knowledge Area. After that, were chosen a subset of the Knowledge Areas to analyse. The subset selected are Data Governance, Data Integration, Reference and Master Data, Metadata, Data Quality, DataOps and Data Observability.

Following then, more considerations were made about the status of companies. With this term are considered all the start-ups which respects the selected characteristics. First, it was determined whether the firms were still active in the market, whether their operations were profit-driven, and, ultimately, whether companies purchased by other corporations were deleted. The dataset was complete after these modifications. Inside

this file, there was a lot of information. Here there is a table with all the information inside the Excel file:

Attribute	Description
Organization Name	Name of the organization
Operating Status	If they work right now
Company Type	If they are for profit or not
Founded Date	The found date
Founded Date Precision	If the found date is date month or year
Industries	Some keywords of the industry
Full Description	The full description provided by the company
Last Funding Amount	Amount of money of the last fund
Last Funding Amount Currency	Currency of the last fund
Last Funding Amount in USD	Amount of money of the last fund in USD
Description	A small description of the company
Number of Investors	Number of total investors
Total Funding Amount	The total amount of money
Total Funding Amount Currency	Currency of the total fund
Total Funding Amount in USD	The total amount of money in USD
Headquarters Location	Description of the headquarters
Headquarters Regions	Description of the state
Website	Link for the website
Number of Funding Rounds	Number of funding receives
Last Funding Date	Date of the last fund

Table 1 Description of the Attributes

Only a subset of these attributes were selected to analyze the companies.

Attribute	Description
Organization Name	Name of the organization
Website	Link for the website
Typology	Different categories of services provided by the company
Founded Date	Year of foundation
Field	Sector in which the company works
Total Funding Amount in USD	The total amount of money received in USD

Last funding Amount in USD	Amount of money of the last fund in USD
Last Funding Date	Date of the last fund
Headquarters Location	Description of the headquarters
State	Description of the state
Headquarters Regions	Description of the region

Table 2 Final Attribute used in the Dataset

The other one were used to be better understand companies context and to select the correct one for the study. In fact, all the companies website were analyzed. There was a check between what is written in the website and what we found out on Cruchbase. Thanks to that, their websites revealed that these organisations offer a variety of services in addition to those identified by the keyword inside Crunchbase. So consequently to that other columns were added to be more precise in the study. As a result, these were the kinds' of final columns:

Typology	Description	Value
Data Quality	Data quality is the measure of how well-suited a data set is to serve its specific purpose.	1 if they provide it 0 if not
Metadata	Metadata is "data that provides information about other data", but not the content of the data, such as the text of a message or the image itself.	1 if they provide it 0 if not
Data Governance	Data governance refers to the collection of practices, policies, and roles related to the effective acquisition, management, and utilization of data—ensuring that the data provide as much value as it can within an organization.	1 if they provide it 0 if not
Data Lineage	Data lineage is the process of understanding, recording, and	1 if they provide it 0 if not

	visualizing data as it flows from data sources to consumption.	
Data Discovery	Data discovery involves the collection and evaluation of data from various sources and is often used to understand trends and patterns in the data.	1 if they provide it 0 if not
Data Integration	Data integration is the process of bringing data from disparate sources together to provide users with a unified view.	1 if they provide it 0 if not
Data Catalog	A Data Catalog is a collection of Metadata	1 if they provide it 0 if not
Master data	Master data is the core data that is essential for running operations within a business enterprise or unit.	1 if they provide it 0 if not
DataOps	DataOps is a collaborative Data Management practice focused on improving the communication, integration and automation of data flows between data managers and data consumers across an organization.	1 if they provide it 0 if not
Data profiling	Data profiling is a technology for discovering and investigating data quality issues, such as duplication, lack of consistency, and lack of accuracy and completeness.	1 if they provide it 0 if not
Data Visualization	Data visualization is a way to represent information graphically, highlighting patterns and trends in data and helping the reader to achieve quick insights.	1 if they provide it 0 if not

Data Observability	Data observability is the ability to understand, diagnose, and manage data health across multiple IT tools throughout the data lifecycle.	1 if they provide it 0 if not
Data Pipeline	A data pipeline is an end-to-end sequence of digital processes used to collect, modify, and deliver data.	1 if they provide it 0 if not

Table 3 Description of the typologies

Once the dataset had been expanded with the new additions, a rigorous analysis was performed to ensure the reliability and accuracy of the results. To achieve this, a stringent filter was applied to the total investments received by each company in order to select only those with a solid foundation and established structure, precisely more than $30.000. This approach ensured that the study focused on companies that were well-positioned to provide meaningful insights and data on the topic at hand. Companies that were still in the early stages of development and were waiting for additional funding rounds to fully realize their ideas were not considered, as their data would not have provided a complete picture of the market. By implementing these measures, the study was able to provide comprehensive and meaningful results that can be trusted by industry experts and stakeholders alike.

Chapter 4: Results of the study

This chapter will detail the different typologies of startups and how these are selected to run the analysis. After the study of the literature, this work consisted of a two-step procedure. In the first instance, there is a focus on finding a reliable dataset that is constructed thanks to research made on Crunchbase and focused primarily on the most important topics for the observatory. Then, there was an analysis of the startup found on the website and different properties were set to select a subset of startups.

4.1 Construction of the dataset

In this work, a subset of categories was selected because there are many companies in the market working on very different topics. So, according to the topic that the observatory research, was selected a subset of the topics in order to reduce the dimension of the analysis. The selected topics were: Data Governance, Data Integration, Reference and Master Data, Metadata, Data Quality, DataOps and Data Observability. The first search made on Crunchbase came out with 317 companies. This number of companies came from a preselection made with these parameters:

- The company is active right now.
- The company wasn't acquired by another company.
- The company needs to be for-profit.
- Different hashtags were used to select whether the companies are linked to the correct section, for this search the hashtags mentioned earlier were used.

Regarding the founding rounds, weren't set a specific amount of money. After that, there was another analysis. And it was noticed that some of them had only the information and they weren't received any found rounds, so these companies were removed. Then, some companies were removed from the lists. These were removed because they didn't receive an investment in the last 5 years and to be considered was set a minimum level of 25k in USD.

4.2 Explanation of the typologies

After these first changes, there was another check on the website if the hashtag was correct or not. So, after reading the information on the website of these companies and

there is control regarding the typology of service they provide. Came out that the majority of these firms don't provide a single service but more than one. So, the final number of businesses is not the algebraical sum of these typologies. So, the final number of companies is 138. In this table (Table 4) are presented all these typologies with the number of companies which provides these services. As stated before, due to the possibility of providing different typologies of services the sum of the firms in this table is higher than 138.

Table 4 Distribution of companies over the typologies

TYPOLOGY	NUMBER OF COMPANIES WHICH IMPLEMENT THIS TOOL
DATA CATALOG	18
DATA INTEGRATION	90
DATA GOVERNANCE	31
DATA LINEAGE	14
METADATA	19
DATA PIPELINE	51
DATA DISCOVERY	9
DATAOPS	13
DATA OBSERVABILITY	13
DATA VISUALIZATION	11
DATA PROFILING	3
MASTER DATA	3
DATA QUALITY	33

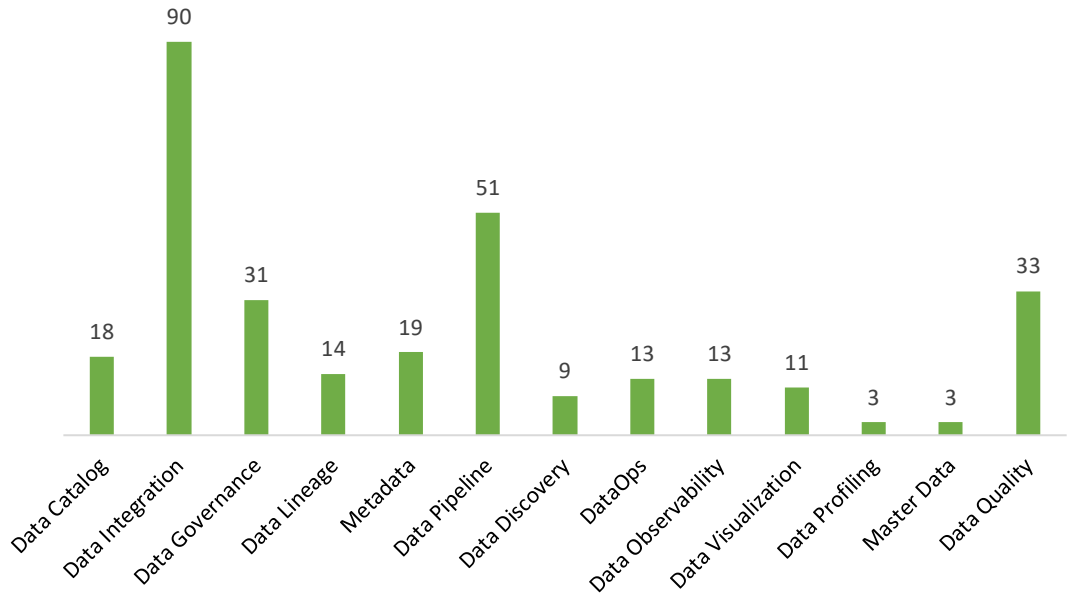


Figure 6 Representation of the distribution of companies over the typologies

From this graph can be seen that the majority of companies manage Data Integration. Secondly, also by the fact that they are connected is Data Pipeline which manages data

like Data Integration. The third typology is Data Quality. It represents an important part of Data analysis. Without correct data analysis are useless. Is important to point out that Data Pipeline, a new trend, is managed by many companies. Considering the total amount of companies which is 138 it represents more or less the half of the total.

4.3 Analysis of the dataset

Following these changes, the data were analysed. The firms were examined in two ways: from a geographical perspective, and from a typology perspective. Finally, the three firms that got the greatest funding in this study were analysed.

Geographical analysis

After that, there was another change in the dataset. Were checked the different headquarters regions in order to track where these companies come from. So, in order to do that, the world was divided into 4 regions which are: APAC, Nord America, Europe, and the United Kingdom. As a result, a summary table will be shown with a division by region, as well as the value of the total funding sum and the average total funding value:

REGION	NUMBER OF COMPANIES	SUM OF TOTAL FUNDING ($)	AVERAGE OF TOTAL FUNDING ($)
APAC	12	153.146.280	12.762.190
EUROPE	20	310.376.185	15.518.809
NORD AMERICA	93	3.951.626.656	42.490.609
UNITED KINGDOM	13	57.352.079	4.411.698
TOTAL	**138**	**4.472.501.200**	**32.409.429**

Table 5 Division of the companies over the region

REGION	SUM OF TOTAL FUNDING ($)	Δ % OF THE AVERAGE
APAC	3,42%	-60,62%
EUROPE	6,94%	-52,12%
NORD AMERICA	88,35%	31,11%
UNITED KINGDOM	1,28%	-86,39%
TOTAL	**100,00%**	

Table 6 Percentage of the sum and the average total funding rounds

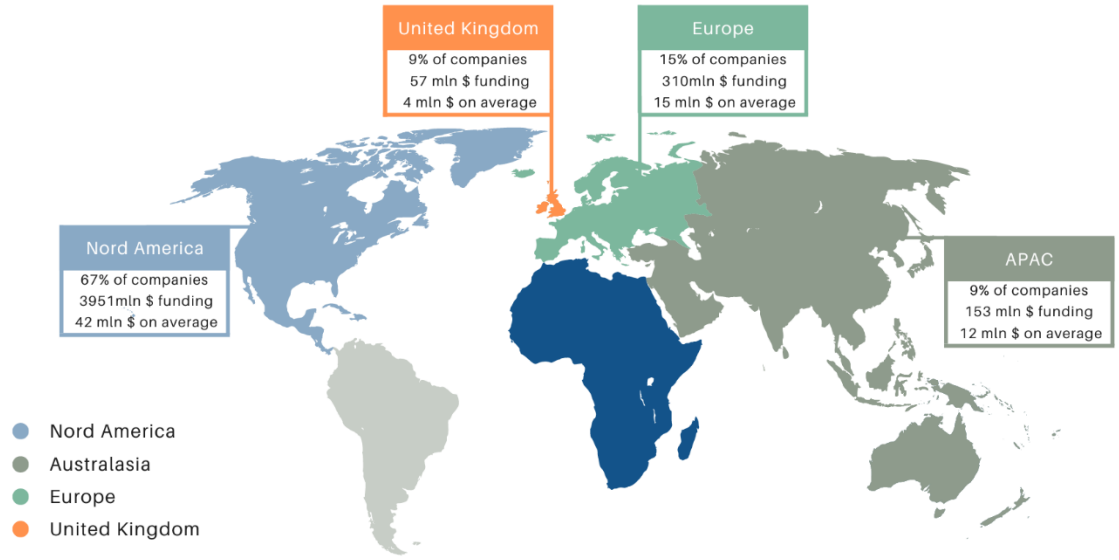


Figure 7 Distribution of the companies

It's easy to see that most companies are in Nord America, which accounts for 67% of the analysis. This number of firms can be related to the easiest way to get economic possibilities in these countries, and better possibilities to receive infrastructure which can be used to run their business. Europe and the United Kingdom represent the second power in the market with 24%. Asia and Australia were put together as APAC to have a better result and are the third power in the market.

The power of Nord America can be noticed also in the funding rounds. In fact, on average a Nord American firm receives 42mln $ compared to a firm in Europa which receives only 15mln $ which is 1/3, and on average Nord America receives 29% more than the other. But from these results, it's important to put attention to APAC. The latter accounts for 9% of the total, which is the same value as the United Kingdom, but APAC companies receive on average three times more found compared to the UK. This difference is more evident looking at the total funding amount and at the number of organizations. In fact, APAC receives 153mln $ with 12 companies, while the United Kingdom only has 57mln $ with 13 companies. As a result, APAC firms receive a very high number of financial resources compared to the United Kingdom. These results can be noticed also in the percentage of the average found. APAC firms on average receive -61% compared to the average value of total funding. This value can be not so high but compared to other values, it is possible to notice that on average all of them went under average. Only in Nord America, there is a higher value compared to the average of all

companies (+29%). After that was made a zoom-in of Nord America which is the most influential region of this analysis. This is the resulting table:

REGION	NUMBER OF COMPANIES	SUM OF TOTAL FUNDING ($)	AVERAGE OF TOTAL FUNDING ($)
CALIFORNIA	47	3.005.520.926	63.947.254
CANADA	7	54.768.255	7.824.036
COLUMBIA	1	46.000.000	46.000.000
DELAWARE	3	1.445.000	481.667
GEORGIA	2	9.300.000	4.650.000
ILLINOIS	1	1.770.000	1.770.000
LOUISIANA	1	200.000	200.000
MAINE	1	4.975.000	4.975.000
MASSACHUSETTS	3	27.550.000	9.183.333
NEW JERSEY	2	8.050.000	4.025.000
NEW YORK	8	402.774.969	50.346.871
NORTH CAROLINA	1	160.000.000	160.000.000
OREGON	2	18.767.507	9.383.754
SOUTH CAROLINA	1	25.000	25.000
SOUTH DAKOTA	1	3.250.000	3.250.000
TENNESSEE	1	460.000	460.000
TEXAS	2	44.700.000	22.350.000
UTAH	2	6.499.999	3.250.000
VIRGINIA	1	160.000	160.000
WASHINGTON	3	39.400.000	13.133.333
WEST COAST	3	116.010.000	38.670.000
TOTAL	93	3.951.626.656	42.490.609

Table 7 Division of the companies over Nord America

This is the division by states of Nord America. Canada was considered a single state. California has the majority of the companies, it has also a higher amount of money. It represents 76% of the total amount. This is because California is well-known for giving excellent funding options to IT startups and companies, particularly those located in Silicon Valley, which is a place filled with prospects for individuals seeking to establish themselves in the information technology field. However, California takes a more focused approach to invest, and even with the first venture cash, IT businesses may struggle to develop owing to the severe competition. The second is New York, which is, after California, the most profitable area for the creation of new companies. This is thanks to different factors, for example, there are tax-free zones or the facility to reach infrastructure and knowledge which is difficult to achieve in other cities. Another factor is the facility to reach a loan.

Typologies analysis

Upon this analysis, was made another analysis of the typology with the region. Here, if one company provides more than one service, the monetary value of the total funding is not split on the base of the number of services that provide. This can also refer to the number of companies linked to the region depending on the typology. These are the resulting table:

TYPOLOGY	APAC	EUROPE	NORD AMERICA	UNITED KINGDOM
DATA CATALOG	83.050.000	68.417.583	131.317.301	0
DATA DISCOVERY	69.000.000	12.000.000	271.958.019	20.208.704
DATA GOVERNANCE	4.211.963	13.174.950	833.512.495	15.323.696
DATA INTEGRATION	12.813.295	198.240.724	2.757.251.252	40.806.126
DATA LINEAGE	72.200.000	17.941.337	249.262.787	1.269.475
DATA OBSERVABILITY	0	59.298.159	197.145.002	16.087.508
DATA PIPELINE	67.619.317	124.675.885	2.388.471.493	51.200.383
DATA PROFILING	0	6.021.913	69.600.000	0
DATA QUALITY	13.082.985	42.056.708	1.291.317.910	3.523.696
DATA VISUALIZATION	49.000.000	34.189.849	211.125.000	525.000
DATAOPS	5.667.612	45.696.031	1.026.639.000	10.300.000
MASTERDATA	0	0	6.210.994	0
METADATA	120.913.668	13.833.918	572.070.282	13.823.696

Table 8 Division of typology and region of the total funding in $

TYPOLOGY	APAC	EUROPE	NORD AMERICA	UNITED KINGDOM
DATA CATALOG	3	5	10	0
DATA DISCOVERY	1	1	5	3
DATA GOVERNANCE	3	2	20	5
DATA INTEGRATION	5	13	65	8
DATA LINEAGE	3	2	8	1
DATA OBSERVABILITY	0	4	8	1
DATA PIPELINE	3	8	34	8
DATA PROFILING	0	1	2	0
DATA QUALITY	4	5	22	2
DATA VISUALIZATION	1	2	7	1
DATAOPS	2	2	10	1
MASTERDATA	0	0	3	0
METADATA	3	2	11	3

Table 9 Division of typology and region of the number of companies

From these tables can be noticed that the firms which manage Metadata the majority come from Nord America. Where Metadata is data that describes other data, providing a structured reference that helps to sort and identify attributes of the information it describes. This trend for Nord America can be also seen in all other trends thanks to the high number of companies present in Nord America. More in-depth, Data Catalog is a collection of Metadata, combined with Data Management and search tools, that helps analysts and other data users to find the data that they need.

Another fact that can be highlighted is the amount of total funding that Data Pipeline has in Nord America. With 34 firms they have collected 2.388mln $ while looking at Data Quality which has only 22 fewer firms, they have collected 1.291mln $. Data Pipeline is a way for ingesting raw data from numerous data sources and then transferring it to a Datastore, such as a data lake or data warehouse, for analysis. Before data flows into a data repository, it is often processed. This high amount of money is also thanks to the presence of Astronomer and Fivetran. These two companies will be presented in the next part. Data integration is the most prevalent typology in the market. But looking at APAC in the total funding part it is possible to notice that Data Integration is the majority in number, but not in monetary terms. Data Lineage, Data Discovery and Metadata have higher amounts of money invested compared to Data Integration. Where, Data Lineage is a type of data life cycle that covers the sources of the data as well as where it flows over time. This phrase may also be used to describe what happens to data as it passes through various processes. Data Discovery, on the other hand, is the act of browsing or applying sophisticated analytics to data to uncover informative patterns that would not have been discovered otherwise. This trend was monitored also by KPMG in a report. The company stated that APAC tech companies are booming at an unprecedented rate. Venture Capital (VC) deals in the region reached an all-time high of USD 152.68 billion in 2018 before seeing a drop in 2019-2020 and hitting the USD 116.91 billion mark in 2021. Meanwhile, Q1 2022 has already seen USD 32.62 billion worth of deals. And they are expanding their growth also in other markets like fintech, biotech, Software-as-a-Service (SaaS), blockchain, health tech, and artificial intelligence (AI). So, APAC is attempting to invest in new Data Management trends. This may be done by inspecting the data. This is also evident when the overall amount of money invested in Data Integration is considered. Five organisations provide this sort of service, but the monetary value is significantly smaller than that of Data Visualisation, which has just one company.

After that was made a zoom-in for Nord America again. In this case were used only Canada, New York, and California because they were significant for the analysis. These are the resulting tables.

TYPOLOGY	CALIFORNIA	CANADA	NEW YORK
DATA CATALOG	118.903.019	9.614.282	0
DATA DISCOVERY	270.188.019	0	0
DATA GOVERNANCE	452.265.000	10.942.495	288.030.000
DATA INTEGRATION	2.224.868.014	42.325.760	68.744.971
DATA LINEAGE	246.871.000	2.271.787	0
DATA OBSERVABILITY	117.025.000	7.342.495	0
DATA PIPELINE	1.737.847.748	26.798.747	386.529.998
DATA PROFILING	0	3.600.000	0
DATA QUALITY	811.979.164	30.398.747	272.030.000
DATA VISUALIZATION	128.125.000	0	17.000.000
DATAOPS	208.156.000	0	52.500.000
MASTERDATA	1.710.995	1.500.000	0
METADATA	522.456.000	9.614.282	0

Table 10 Division of typology of Nord America and the total funding in $

TYPOLOGY	CALIFORNIA	CANADA	NEW YORK
DATA CATALOG	7	2	0
DATA DISCOVERY	4	0	0
DATA GOVERNANCE	9	2	3
DATA INTEGRATION	31	4	5
DATA LINEAGE	6	1	0
DATA OBSERVABILITY	3	1	0
DATA PIPELINE	17	1	4
DATA PROFILING	0	1	0
DATA QUALITY	11	2	1
DATA VISUALIZATION	3	0	3
DATAOPS	4	0	1
MASTERDATA	1	1	0
METADATA	8	2	0

Table 11 Division of typology of Nord America and the number of companies

From these tables can be noticed that New York focuses the investment principally on three areas, which are Data Governance, Data Pipeline, and Data Quality. Looking at the number of companies present there the amount of money received is very high. This can be associated with the factors previously mentioned. Compared to Canada which has the same number of companies more or less, they received a lower amount of money in all these three categories. While regarding California is important to highlight Metadata. With only 8 companies which manage this typology, they received 522mln$ which is higher compared to Data Governance which received 452mln$n with 9 companies.

After these first analyses, there was the curiosity to check how the investment grew over the years for the new trends. For doing that, was considered only the last investment received by the companies in the last 5 years.

TYPOLOGY	2017	2018	2019	2020	2021	2022
DATA CATALOG		12.000.000		25.726.413	37.564.115	207.494.356
DATA DISCOVERY		12.000.000	644.718		135.563.986	224.958.019
DATA GOVERNANCE		16.113.618	10.644.718	85.690.790	426.008.978	327.765.000
DATA INTEGRATION	2.515.000	6.199.950	3.805.794	199.417.213	1.669.573.963	1.127.599.477
DATA LINEAGE		12.000.000	1.269.475	45.300.000	48.291.787	233.812.337
DATA OBSERVABILITY				13.626.413	127.214.836	131.689.420
DATA PIPELINE		25.000	13.302.551	96.304.694	1.625.399.107	896.935.726
DATA PROFILING					69.600.000	6.021.913
DATA QUALITY		9.088.618	65.232.881	185.385.000	461.461.306	628.813.494
DATA VISUALIZATION			10.000.000	2.500.000	100.314.849	182.025.000
DATAOPS				40.552.989	27.975.000	291.666.654
MASTERDATA	1.500.000		2.999.999		1.710.995	
METADATA		14.913.668	644.718	14.476.413	62.900.765	627.706.000
TOTAL	**4.015.000**	**82.340.854**	**108.544.854**	**708.979.925**	**4.793.579.687**	**4.886.487.396**

Table 12 Comparison over the years of the total funding amount in $ divided by typology ($)

TYPOLOGY	2017	2018	2019	2020	2021	2022
DATA CATALOG	0	1	0	4	4	9
DATA DISCOVERY	0	1	1	0	4	3
DATA GOVERNANCE	0	4	2	7	12	6
DATA INTEGRATION	1	3	3	10	38	35
DATA LINEAGE	0	1	1	2	4	6
DATA OBSERVABILITY	0	0	0	3	6	4
DATA PIPELINE	0	1	3	4	20	23
DATA PROFILING	0	0	0	0	2	1
DATA QUALITY	0	3	3	4	10	13
DATA VISUALIZATION	0	0	1	1	4	5
DATAOPS	0	0	0	2	2	8
MASTERDATA	1	0	1	0	1	0
METADATA	0	2	1	3	5	8

Table 13 Comparison over the years of the number of companies divided by typology

These are the resulting values for all different Knowledge Areas analyzed. Through this table can be seen an higher amount of money on average received in the last years. Contrary to the COVID period companies received investment. Obviously in the 2021 after the pandemic period companies start again to invest in new technologies. In this case cannot be made a precise confront between 2021 and 2022 due to the fact that the sample use data from 06/12/2017 to 06/12/2022 so wasn't considered the last period of December where could be the possibility to invest in these companies. But this does not impact the total much. In fact more funding was received in 2022. Especially we can see this in Data Lineage one of the new trends in the market that is gaining momentum. Or another trend that has risen is DataOps, which will be analyzed in the next section. Another trend to note is Data Integration but from a negative point of view. This factor could be derived from the new business possibilities that are being created in the last few Years. Just think of all the new positions previously analyzed that allow to study all the different facets of data and not dwell only on Data Integration. The biggest

exploit we find in Metadata where going from 5 to 8 companies that invested in this area, the total investment went up by $600,000. This can be related to the fact that Metadata are becoming increasingly important due to the vast amount of data being generated and the need for effective data management. Metadata provides information about data, including its structure, format, and context, making it easier to understand, locate, and use. With the rise of big data, metadata plays a crucial role in data governance, helping organizations maintain data quality and compliance with regulations. Additionally, metadata enhances search functionality, making it easier to find relevant data quickly. As data continues to grow in complexity and volume, metadata will continue to be a vital component in managing and extracting value from data. Now will be analyzed more in depth the changes for the new trends.

TYPOLOGY	2018	2019	2020	2021	2022
DATA OBSERVABILITY			13.626.413	127.214.836	131.689.420
DATA PIPELINE	25.000	13.302.551	96.304.694	1.625.399.107	896.935.726
DATA VISUALIZATION		10.000.000	2.500.000	100.314.849	182.025.000
DATAOPS			40.552.989	27.975.000	291.666.654

Table 14 Comparison over the years of the total funding amount in $ for New Trends

TYPOLOGY	2018	2019	2020	2021	2022	TOTAL
DATA OBSERVABILITY	0	0	3	6	4	13
DATA PIPELINE	1	3	4	20	23	51
DATA VISUALIZATION	0	1	1	4	5	11
DATAOPS	0	0	2	2	8	12

Table 15 Comparison over the years of the number of companies for New Trend

Looking at the new trend that has been growing over the last years, the increasing tendency to invest in these topics can be seen. IBM defined DataOps as the "practices that bring speed and agility to end-to-end data pipelines process, from collection to delivery. “So the increasing importance of this topic can be associated firstly with the increasing number of data that companies need to manage. The amount of data granularity to which data teams have access will only increase. Consequently, another topic that DataOps wants to manage is the reduction of processes which are useless for the company or processes which are run without thinking about the costs. The latter is the cost of mismanaging data, which affects not just money but also teams' ability to make better judgements. Regarding the Data Observability can be seen that 2021 received a higher amount of investment, but the value is higher compared to

2020. Data Observability has intended the capacity to manage data health across different IT tools. The final object is to check the data lifecycle. The growing importance of this topic can be related to the increased importance of the transition from physical to cloud systems. Cloud architecture enables businesses to store and use data in the cloud, while data visualisation tools help teams make sense of that data. But the necessity to implement Data Observability is to prevent data problems. By making it simpler to manage their data architecture and minimise excessive over-provisioning, businesses may increase productivity, accelerate innovation, and potentially cut IT expenditures. While a data pipeline can be seen as a set of processes that move data from one place to another. It involves extracting data from various sources, transforming it into the desired format, and loading it into a destination, such as a database or a data lake. The data pipeline is an automated process that runs on a regular schedule to ensure that data is up-to-date and available for analysis and decision-making. Data pipelines are important for organizations to manage their data efficiently, as they help to automate the tedious and time-consuming task of manual data transfers and ensure data consistency and integrity. Looking at the data can be seen important investment in 2021. This is because as previously some of the most important companies which impact the analysis received investment this year. Examples are mParticle or Fivetran. Looking more at these two the first one is a customer data platform (CDP) that assists businesses in unifying and managing customer data across numerous touchpoints and channels. To gather and organise client data, the firm offers a single data layer that interfaces with numerous data sources such as mobile applications, websites, and connected devices. This information may then be utilised to establish a unified consumer perspective and generate tailored experiences across numerous channels. mParticle also offers a set of data management capabilities, including data governance, data quality management, and data privacy controls. The platform developed by the firm is adaptable and scalable, and it can be utilised by organisations of all kinds, from startups to major companies. While Fivetran received a higher amount of money in the market and will be analysed in the next part. The last trend might be defined as the process of displaying facts graphically or pictorially. It aids in the visual representation and analysis of data in order to get insights and make educated decisions. It entails developing charts, graphs, maps, and other visual representations of data in order to

communicate information and patterns efficiently. The purpose of data visualisation is to make complicated data clear and actionable. In this case as Data Pipeline here there is another important company called Dataloop AI. The DataLoop data management and annotation platform simplify the preparation of visual data for machine and deep learning. Dataloop is used to create and deploy robust computer vision pipelines. They provide also other typologies of services. For example, their platform helps organizations to leverage AI and ML techniques to improve their operations and make data-driven decisions. The company's offerings may include features such as predictive maintenance: using machine learning algorithms to analyze data and predict when equipment is likely to fail, allowing organizations to proactively address potential problems. The goal of the company is to help organizations to increase efficiency, reduce costs, and improve overall performance through the use of AI and ML.

Analysis of the 10 Most Funded companies.

After, there was the decision to take a closer look at the best-funded businesses to understand which areas of innovation are attracting the most interest. To contain the research were selected only the most financed companies were. This is the summarized table:

TYPOLOGY	TOTAL FUNDING ($)
FIVETRAN	728.108.000
ALATION	314.950.000
ASTRONOMER	282.900.000
MPARTICLE	272.030.000
AIRBYTE	181.200.000
CDATA SOFTWARE	160.000.000
CRUX	115.935.000
OBSERVE	112.000.000
STRIIM	108.500.000
UNRAVEL DATA	107.156.000

Table 16 Top ten companies in terms of total amount of funding received

Let’s have a look more in-depth at to first three. Fivetran is a technology company that provides an automated data integration platform for organizations. The platform helps businesses to connect and centralize their data from various sources, including cloud and on-premise systems, into a single source of truth. Fivetran's technology is designed to be fast, reliable, and scalable, making it easy for organizations to connect and analyze their data without the need for custom code. The platform provides pre-built connectors for popular data sources and can be easily configured for other sources. The second one

is Alation. Alation is a technology company that provides a data catalog and governance platform for organizations. The platform enables businesses to discover, understand, trust, and effectively use their data assets. Alation's technology helps organizations improve the quality and efficiency of their data operations by providing a centralized location for metadata and data assets, as well as tools for data collaboration, stewardship, and governance. The third one is Astronomer which is a technology company that provides a cloud-based data engineering platform for organizations. The platform helps businesses to streamline their data pipeline and make data easily accessible to decision-makers. Astronomer's technology provides an end-to-end solution for data engineers and analysts to manage their data workflows and infrastructure, including tools for data ingestion, transformation, and analysis. The platform is designed to be scalable and modular, making it a good fit for organizations of all sizes. Another company which manage pipeline is Striim. Striim is a technology company that provides a real-time data integration and streaming platform for organizations. The platform enables businesses to collect, process, and analyze data in real time from various sources, including cloud and on-premise systems. Striim's technology helps organizations make informed decisions and take action in real time by providing a single view of data from disparate sources. With this, customers through dashboards can easily monitor the state of their data flows. Another fact is that these 10 companies account for 53.28% of the total found in the sample and all come from North America.

Chapter 5: Conclusion and Further Improvements

The objective of this study was to analyse the different companies in the market which provide Data Management services and try to understand the trend over the last year and how these changed over the years. To do this, these companies were analysed in monetary and geographical terms to keep track of where and how much they invested in the different technologies.

All of these firms are related to a subset of Knowledge Areas present in the DAMA Wheel. This comes from the most important book which talks about Data Management. To be in line with the Observatory and the research selected Data Governance, Data Integration, Reference and Master Data, Metadata, Data Quality, DataOps and Data Observability regarding the Knowledge Areas. While to track the new trends present in the market were selected also DataOps, Data Observability, Data Catalog, Data Visualization, and Data Pipeline.

The intention of this analysis is to monitor which services companies provide and which are the topic more important in this period. This historical period is characterized by a very fast technological innovation that has allowed companies not to stop their activity. This can be related to the increasing necessity of new technologies to protect data, to manage in a more complex and coordinated way the data, but also to the more requests of the consumers in terms of computational velocity and security. Hence, the investments did not stop, and new topics are discovered over the years. Can be seen also in the new roles that are emerging in these latest years like Data Scientist or Data Engineer.

The analyzed context constitutes a small amount compared to the entire market. Were chosen only companies which provide the mentioned areas. These were filtered by typologies of service which they provide after that were checked the geographical location of the foundation looking before the region and after that the corresponding state. After this preliminary part was made an analysis over the years.

In the first analysis can be seen that Nord America receives the highest amount of money. This can be related to the typology of the market that was analysed. In fact, in these last years, APAC which is Asia-Pacific the part of the world near the western Pacific Ocean has been an important growth market regarding new technologies like Web3 and Blockchain, and regarding the new Data Management trends.

This dominant position of Nord America can be seen in several companies present in this sample. Indeed, in the APAC area, the number is fewer than in Europe and North America (9% vs. 15% and 67%, respectively), and companies in this region receive the same amount of money on average.

While the United Kingdom had invested a small amount of money in these companies. This trend cannot be explained with this analysis maybe it is reflected in the difficulty to reach investment. It is important also to clarify the fact that only the United Kingdom was considered a single region. This was an assumption made after the coming of Brexit. So, these results can be distorted.

To summarise, the market is now primarily focused on providing Data Integration solutions. In terms of geography, North America has received most of the investments. However, this tendency has begun to reverse in recent years. Indeed, we can see this in new regions such as APAC, where most investment is in Data Pipeline and Data Lineage, which are new concepts that have gained traction in recent years. This new tendency might be attributed to the necessity to develop new data-management services. Clearly, there is a shift toward new technologies since, as this research shows, the market for certain services is adequately serviced while others have a lot of room for growth.

Further analysis can be done considering also other fields not only the selected ones. This can enlarge the sample and can highlight the difference between the countries. There could be the possibility to enhance regions like Africa or LATAM, which is Latin America, that wasn’t present in this analysis since no companies with the prerequisites were found.

Lastly could be select a subset of tools provided by these companies and analyse the different capabilities which can implement. After that apply them to a real sample of data to check and monitor the different methodologies and typologies of

implementation and treatment of data. This resulting analysis could be merged with a comparison of these new technologies with the big provider present in the market. This can be done to discover if the new providers are trying to implement new functionality or only upgrade the present one. The big provider can be checked through Gartner Magic Quant. Magic Quadrants compare vendors based on Gartner's standard criteria and methodology. With this comparison can be discovered necessities which are not covered by the new technologies and the existing ones.

References

- H. Daki, A. El Hannani, A. Aqqal, A. Haidine and A. Dahbi; *Big Data Management in smart grid: concepts, requirements and implementation;* 2017
- C. Stendman, J. Vaughan; *What is Data Management and why is it important?*; 2019
- Arun Kumar, Matthias Boehm, Jun Yang; *Data Management in Machine Learning: Challenges, Techniques, and Systems*; 2017
- E. Deelman, A. Chervenak; *Data Management Challenges of Data-Intensive Scientific Workflows*; 2008
- A. Fernando, C. Catalin; *The main challenges and issues of big Data Management*; 2012
- L. zanotti; *Big Data Management significa capire meglio la domanda per dare una risposta in tempo reale;* 2015
- C. Harvey; *What Is Big Data Management?*; 2017
- D. Agrawal, P. Bernstein, E. Bertino, S. Davidson, U. Dayal, M. Franklin, J. Gehrke, L. Haas, A. Halevy, J. Han, H. V. Jagadish, A. Labrinidis, S. Madden, Y. Papakonstantinou, J. M. Patel, R. Ramakrishnan, K. Ross, C. Shahabi, D. Suciu, S. Vaithyanathan, and J. Widom; *Challenges and Opportunities with Big Data*; 2012
- H. E. Pence; *What is Big Data and Why is it Important?*; 2014
- F. Nargesian, E. Zhu, R. J. Miller, Ken Q. Pu, Patricia C. Arocena; *Data lake management: challenges and opportunities;* 2019, pp 1986–1989
- D. McGilvray; *Ten Steps to Quality Data and Trusted Information;* 2009
- R. S. Seiner; *A Review of Malcolm Chisholm's Book: Definitions in Information Management: A Guide to the Fundamental Semantic Metadata;* 2010

- M. Janssen, P. Brous, E. Estevez, Luis S.Barbosa, T. Janowski; *Data governance: Organizing data for trustworthy Artificial Intelligence*
- O. Benfeldt, J. S. Persson and S. Madsen*; Data governance as a collective action problem. Information Systems Frontiers*, 2020, pp 299–231
- I. Alhassan, D. Sammon, M. Daly; *Data governance activities: an analysis of the literature;* 2016
- C. Stedman, J. Vaughan; *What is data governance and why does it matter?; 2022*
- A. Morris; *Data Discovery: What Is It & Why Is It Important?; 2021*
- T. Olavsrud*; What is data architecture? A framework for managing data; 2020*
- S. Roddewig*; Master Data: What Is It & Why Does It Matter for Businesses?; 2022*
- D. Agrawal, P. Bernstein, E. Bertino, S. Davidson, U. Dayal, M. Franklin, J. Gehrke, L. Haas, A. Halevy, J. Han, H. V. Jagadish, A. Labrinidis, S. Madden, Y. Papakonstantinou, J. M. Patel, R. Ramakrishnan, K. Ross, C. Shahabi, D. Suciu, S. Vaithyanathan, and J. Widom, *Challenges and Opportunities with Big Data – A community white paper developed by leading researchers across the United States, 2012*
- Guetat, S.B.A. and Dakhli, S.B.D., *The architecture facet of information governance: the case of urbanized information systems,* 2015.
- H.J. Watson, C. Fuller, T. Ariyachandra; *Data warehouse governance: best practices at blue cross and blue shield of North Carolina", Decision Support Systems*; pp. 435-450
- P. Vassiliadis, A. Simitsis, P. Georgantas, M. Terrovitis, and S. Skiadopoulos*; A generic and customizable framework for the design of ETL scenarios; 2005 pp 492-525*
- D. Loshin; *Data Governance for Master Data Management and Beyond: a white paper; 2013*

- B. Otto; *Organizing Data Governance: Findings from the Telecommunications Industry and Consequences for Large Service Providers;* 2011 pp. 45-66
- A. McAfee, E. Brynjolfsson; *Big Data: The Management Revolution. Harvard Business Review*; 2012 pp. 60-68.
- D. Myers; *The Value of Using the Dimensions of Data Quality*; 2013
- Y. Noguchi.; *The Search for Analysts to Make Sense of Big Data;* 2011
- A. Immonen, P. Pääkkönen, E. Ovaska ;*Evaluating the Quality of Social Media Data in Big Data Architecture*, 2015
- Osservatorio Digital Innovation. *Le 5V dei Big Data: dal Volume al Valore*. 2019; Retrieved from Osservatorio Digital Innovation: https://blog.osservatori.net/it_it/le5v-dei-big-data.
- Osservatorio Startup Intelligence. *Startup Intelligence Data Monetization. Milano: Politecnico di Milano*. 2021
- R. Abraham, J. Schneider, J. Brocke; *Data governance: A conceptual framework, structured review, and research agenda;* 2019
- R. Atan, R. Abdullah, M. Azrifah and A. Murad; *Security Framework of Cloud Data Storage Based on Multi Agent System Architecture: Semantic Literature Review;* 2010
- N. Elgendy, A. Elragal; *Big Data Analytics: A Literature Review Paper. ICDM - Industrial Conference on Data Mining; 2014* pp. 214-227
- M. H. Ofner, K. Straub, B. Otto and H. Oesterle; *Management of the master data lifecycle: a framework for analysis;* 2011
- Lahiru K. W. Fernando and Prasanna S. Haddela; *Hybrid framework for master Data Management; 2017*
- R. Silvola, Olli Jaaskelainen, H. Kropsu-Vehkapera and H. Haapasalo; *Managing one master data – challenges and preconditions;* 2010

- J. F. Roddick, L. A1-Jadir, L. Bertossi, M. Dumas, F. Estrella, H. Gregersen, K. Hornsby, J. Lufter, F. Mandreoli, T. Mainnisto, E. Mayol, L.Wedemeijer; *Evolution and change in data management — issues and directions;* 2000
- R. Ande, B.Adebisi , M. Hammoudeh , J. Saleem; *Internet of Things: Evolution and technologies from a security perspective;* 2019
- A. Tahsin; Md. M. Hasan *Big Data & Data Science: A Descriptive Research on Big Data Evolution and a Proposed Combined Platform by Integrating R and Python on Hadoop for Big Data Analytics and Visualization;* 2020
- R. Abraham, J. Brocke, J. Schneider*; Data Governance: A conceptual framework, structured review, and research agenda. International Journal of Information Management;* 2019
- *Asia-Pacific's tech startups scene is hot: KPMG report;* https://coingeek.com/asia-pacific-tech-startups-scene-is-hot-kpmg-report/
- *Evolution of Data;* https://www.ewsolutions.com/evolution-of-data-and-data-management/