Towards a Framework Driven by Use Cases in Data-heterogeneous Environments

Ginés Molina Abril[1,2][0000−0001−7877−0812], Oriol Caralt[2], Jose A. Martínez[3][0000−0003−2131−9101], and Javier Luis Cánovas Izquierdo[1][0000−0002−2326−1700]

[1] IN3 - UOC, Barcelona, Spain
{ginesmoli,jcanovasi}@uoc.edu
[2] Independent Consultant, Barcelona, Spain
oriol.caralt@gmail.com
[3] Universidad Politécnica de Cartagena, Spain
josean.martinez@upct.es

Abstract. Most organizations base their strategic decisions on the analysis of business performance data. With the emergence of artificial intelligence, this analysis also includes applying machine learning techniques, among others, which help to discover and predict patterns in data. Although there are a number of tools to perform data analysis, they require considerable effort to be adapted to each company's use case. Companies need to consider the cost associated with the infrastructure or the commitment to profiles responsible for building and maintaining these tools. Furthermore, the return on investment is hampered by the lack of skills, leadership or policies for using these tools. This paper proposes a framework to address this situation by facilitating the process of consuming and analysing data over time. Our proposal emphasizes the definition of data use cases, which drive the data enablement, consumption, discovery and storage phases. The proposed framework is being developed and put into practice through an industrial PhD within some companies evolving to be data-driven, thus allowing real-world validation.

Keywords: Data Engineering · Data Infrastructure · Data Governance

1 Introduction

The volume of data available to conduct business performance analysis has grown exponentially in recent years, resulting in a number of heterogeneous data sources that companies have to digest. Although technology has made it easier to obtain raw data, there may not be a single process to collect data that fit the company's use cases or questions. As far as companies are concerned, they need a specific data consumption and governance framework that differentiates them in a highly competitive environment [3], as most have information silos between different departments [6,2] and there is no single

source of truth. With the emergence of artificial intelligence, this framework may also include machine learning techniques to discover and predict data patterns to facilitate analysis and discover potential changes in the market.

Current approaches tackle all these problems by working on each part of the data lifecycle with different independent components. However, there is no comprehensive and affordable solution that allows any enterprise to implement a data governance model focused on information self-consumption, data quality and data consumption process performance. For example, TOKERN[4] is an Open-Source framework that facilitates monitoring sensitive data, auto-discovery, and management of data sources. There are also solutions such as AIRBYTE[5] that facilitate data ingestion, and APACHE AIRFLOW[6], which acts as an ETL workflow orchestrator. Furthermore, DBT[7] is a framework that allows the automation of data transformations, and DATAHUB[8] is used as a data catalog platform. There are other modern tools in workflow execution, such as MAGE[9], which offer a series of functionalities that make these tasks much more accessible to Business Analysts with minimal knowledge of SQL or Python. Although solutions are increasing, it is hard and time-consuming to adapt them to specific company's data use cases, resulting into inefficient processes [5] and the feeling of not getting the most out of the data.

This paper proposes a framework driven by use cases to facilitate heterogeneous data consumption from a standardized architecture to optimize business processes. Data use cases are described in a data governance definition, which covers the required configuration for data enablement, consumption, storage, and discovery. The main objectives are maximising the return on infrastructure investment, facilitating end-to-end data lifecycle traceability, providing reliable metrics, and facilitating data navigation. The proposed framework is being developed and put into practice by an industrial PhD, focused on business use cases, data engineering and data science applications. This paper shows the starting point, challenges and future work.

The remaining of this paper is organized as follows. Section 2 describes the approach and Section 3 presents the conclusions and future work.

2 Approach

Our approach is illustrated in Figure 1. It consists of three main phases: *Data Enablement* (A), *Data Consumption & Storing* (B), and *Data Discovery* (C). These phases are driven by a *Data Governance Core* that depends on *Business Strategy*, which includes a set of use cases describing the company's data analysis objectives in marketing (e.g., [1]). A data use case definition covers:

[4] `https://tokern.io`
[5] `https://airbyte.com`
[6] `https://airflow.apache.org`
[7] `https://www.getdbt.com`
[8] `https://datahubproject.io`
[9] `https://www.mage.ai`

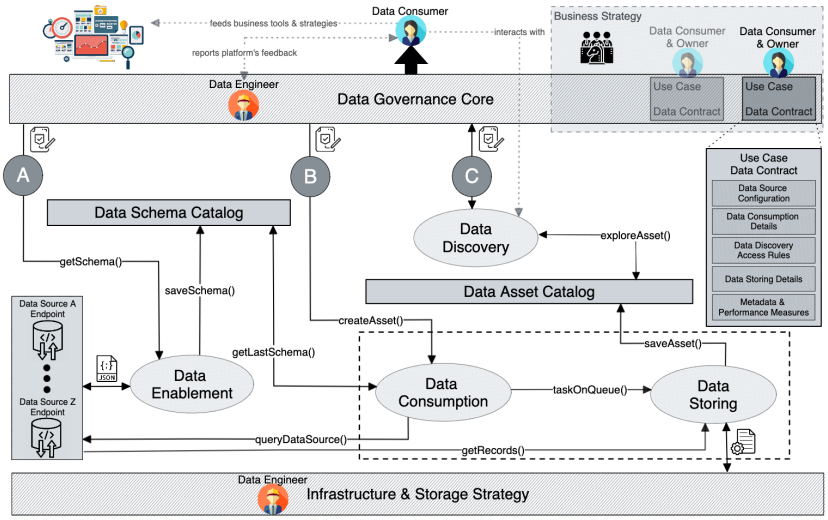

Fig. 1. Data Engineering Framework

(1) data source configuration, (2) consumption details, (3) access rules, (4) storing details and (5) metadata and performance measures. Furthermore, the approach also includes an *Infrastructure & Storage Strategy*, which defines the technical configuration to store collected data. The approach considers two roles for the stakeholders: (1) *Data Engineer*, who is responsible for the data governance definition and the infrastructure & storage strategy, and (2) *Data Consumer*, who may either own data assets based on a *Data Contract* or just need to explore data, whether internal employee or external person. In the following, we describe each phase and the challenges to address them.

Data Enablement. The objective of this phase is to track how the data schema of each data source changes over time. Once this phase is triggered (see *getSchema()*), it explores the schema using the configuration defined by the Data Engineer in the use case definition, and updates the schema catalog (see *saveSchema()*) according to the data configuration. The main challenge of this phase is the automation and pre-enablement of data given their heterogeneity.

Data Consumption and Storing. The objectives of this phase are threefold: (1) serve data consumer requests, (2) perform data ingestion and storage tasks, and (3) measure the tools and libraries performance. This phase is executed in two steps: (1) the data consumption process, which first recovers the schema information from the Data Schema Catalog (see *getLastSchema()*), then prepares the collection process (see *queryDataSource()*), and finally queues it (see *taskOnQueue()*); and (2) the data storing process, which prepares and executes the environment for each queued element, stores the data and reports the performance of the task (see *saveAsset()*). This phase is driven by the information regarding data consumption, discover and storage defined in the

use case definition. The main challenges of this phase are designing the infrastructure for pipeline execution, integrating different data storage solutions (and discovering potential data links), developing the performance measurement systems, and storing data following security rules and legal considerations (according to the *Infrastructure & Storage Strategy*).

Data Discovery. The objective of this phase is to allow the exploration, traceability, lineage, and version control of the data assets. In this case, the exploration task (see *exploreAsset()*) allows the user to perform a series of actions to consume data. This phase uses the information regarding the metadata and performance measures defined in the data governance definition. The main challenge of this phase is to enable different access roles in order to create or update data assets, redefine metrics, and share information.

3 Conclusion and Future Work

This paper proposes a framework driven by use cases to facilitate heterogeneous data consumption. Data use cases are described in a data governance definition, which drives the three phases of the framework: (1) data enablement, (2) consumption & storage, and (3) discovery. This approach follows some principles of the Data Mesh approach [4], which focuses on improving business processes by collaborating on defining business metrics, auto-consuming data through a flexible infrastructure, and implementing an effective data governance model. The proposed framework and the approach are being validated in a real-world scenario. We have presented the main challenges to address in each phase, which will be addressed in future work.

Acknowledgements This work has been supported by the TED2021-130331B-I00 project, funded by MCIN/AEI/10.13039/501100011033 and "NextGenerationEU"/PRTR.

References

1. Camilleri, M.A.: The use of data-driven technologies for customer-centric marketing. International Journal of Big Data Management **1**(1), 50–63 (2020)
2. Dell, R.K.: Breaking organizational silos: Removing barriers to exceptional performance. Journal AWWA **97**(6), 34–36 (2005)
3. Hilger, J., Wahl, Z.: Data Catalogs and Governance Tools, pp. 187–192. Springer International Publishing, Cham (2022)
4. Machado, I.A., Costa, C., Santos, M.Y.: Data mesh: Concepts and principles of a paradigm shift in data architectures. Procedia Computer Science **196**, 263–271 (2022)
5. Romero, O., Wrembel, R., Song, I.Y.: An alternative view on data processing pipelines from the dolap 2019 perspective. Information Systems **92**, 101489 (2020)
6. Walsh, M.J., McAvoy, J., Sammon, D.: Grounding data governance motivations: a review of the literature. Journal of Decision Systems **31**(sup1), 282–298 (2022)