Summary

Ferdinando Micco

June 2023

1 Motivation

In today's business landscape, Big Data refers to the vast amount of structured and unstructured data that inundates organizations on a daily basis. The ability to store, process, and analyze such data is crucial for extracting insights and making informed decisions. Companies generate and collect large volumes of data from various sources, including user interactions, enterprise resource planning systems, and other software applications. The goal is to leverage this data to gain business insights and derive value from it. However, without reliable and relevant datasets, business intelligence insights may be incomplete, inaccurate, or biased, leading to misguided strategic choices.

2 Problem

In the modern paradigm of data driven organizations, data engineers focus on building and managing all the infrastructure of data pipelines, while data analysts utilize advanced statistical and analytical techniques to derive insights from data. However, There is a misalignment between the technical expertise of Data engineer and the analytical skills of Data analyst may lead to difficulties in finding proper coordination and collaboration. To address this challenge, a new role is needed to bridge the gap. Analytics engineers serve as a connection between technology and business, with the goal of providing useful, clean, and accurate datasets for business needs.

3 Purpose and goal

dbt (Data Build Tool) is an open-source software developed by Fishtown Analytics, specifically designed to equip analytics engineers with robust tools for dataset creation and validation. The purpose of this thesis is to design and implement an architecture that embodies the new paradigm, wherein analytics engineers play a central role in the transformation phase. The goal of this thesis is to build a comprehensive dbt project from scratch, starting with a database, in order to showcase the key characteristics that facilitate the transition from

the old paradigm to the new one. Moreover, a proof-of-concept solution has been developed to seamlessly integrate the dbt project into a serverless data pipeline hosted on AWS. The deliverables include the thesis, the code and the documentation to explain the code.

4 Methodology

The methodology of this project can be considered as systemic research on dbt and AWS. The first step would be to figure out the components related to this project, and the interactions between them. The next step is to build a concept about how to realize the goals. The concept should be demonstrated whether it is feasible or not. In this case, it is to generate a design about how to run a serverless DBT project on AWS. Then, the design should be implemented to check its feasibility. According to the situation, minor changes could be applied to adjust the design in order to achieve a better result.

5 Project

The thesis project offers a practical use case where the new paradigm introduced by Analytics Engineering has been adopted and experimented with. Data engineers and data analysts can now focus on their primary tasks, relying on the expertise of analytics engineers for the creation of data models.

The final solution is a cloud-based data pipeline implemented on AWS. The data sources are stored as raw data in the Simple Storage Service (S3) and made accessible through Redshift Data Warehouse as external tables. The entire data transformation phase has been managed using dbt, which has been extensively tested and proven to be a comprehensive toolkit for analytics engineering. Within the dbt project, model creation, testing, validation, and documentation are fully supported and partially automated to ensure accurate and reliable datasets.

The dbt project has been deployed as a Docker image on an ECR instance and executed within a pre-configured virtual environment on AWS Fargate. To utilize the dbt command-line interface (CLI) and execute any type of command within the project, a cron schedule of ECS tasks has been established using EventBridge. Furthermore, an automated system for failure detection has been implemented through log monitoring with CloudWatch. Lambda functions have proven to be a reliable service for executing functions that respond appropriately to these failures.

6 Speculative analysis synthesis

In the traditional scenario challenges arise due to the Data analyst's primarily business-oriented skill set. The most critical step here is the Data collection and

preprocessing phase, which requires a deep understanding of the data infrastructure. It involves selecting the appropriate Data sources, ensuring data quality, handling data transformations, and addressing any inconsistencies. This can lead to potential errors or limitations in the resulting data model, affecting the quality and reliability of the insights derived from it.

In contrast, in the proposed scenario either a Data analyst or a Data engineer is involved in performing these operations. Data engineering team in this case have few understanding of the business needs and all the data collection process risk to be not accurate or not relevant. Data engineering teams usually provide data models that highlight only KPIs but are not able to provide real insight; furthermore, they are not experts in the discovery process of data analysis. Their job is to build and maintain the Data pipeline and manage its overall orchestration. This means that if a data analyst makes a request to data engineers, it will be added to the list of scheduled tasks to be completed. The lack of autonomy and the unpredictable waiting time significantly impair the efficiency of the analysts' work.

7 Conclusion

Analytics engineers are knowledge specialists: like librarians, they curate an organization's knowledge. They acquire, codify and make sure that all the knowledge is reliable and current. They represent the bridge between business and technology that fill this gap between existing roles and competences. These are the individuals who have been absent from our data projects, and this is the practice that we must refine if we are to address the fundamental dysfunction of data. Data engineers and data analysts can now focus on their primary tasks, relying on the expertise of analytics engineers for the creation of data models.

Moving forward, several potential improvements can be considered for future enhancements. Firstly, exploring the integration of additional data sources could broaden the scope and depth of insights derived from the pipeline. Secondly, enhancing the error handling and recovery mechanisms within the system would further improve the pipeline's reliability. Additionally, implementing real-time or near real-time data processing capabilities could enable more dynamic and timely analytics. Lastly, integrating a comprehensive data governance framework and security measures would ensure data privacy and compliance with regulations.

By addressing these future improvements and incorporating these elements, the data pipeline with dbt can continue to evolve and provide even more valuable insights for data engineers, data analysts, and other stakeholders involved in the analytics process.