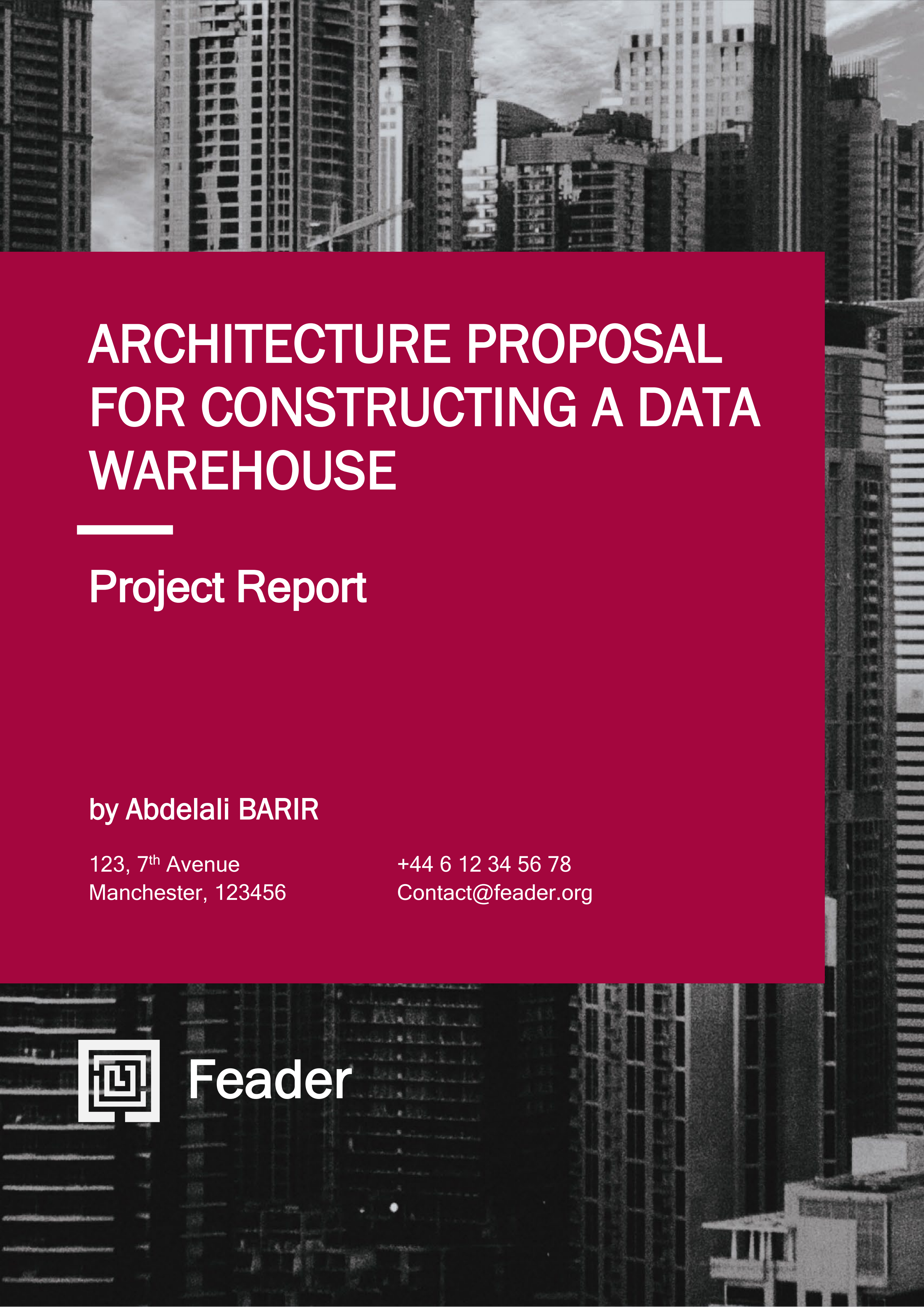


ARCHITECTURE PROPOSAL FOR CONSTRUCTING A DATA WAREHOUSE

Project Report

by Abdelali BARIR

123, 7th Avenue
Manchester, 123456

+44 6 12 34 56 78
Contact@feader.org

Feader

Table of contents

List of Abbreviations

BI : stands for Business Intelligence, refers to the procedural and technical infrastructure that collects, stores, and analyses the data produced by an organization's activities (Frankenfield, 2022).

KPIs : stands for Key Performance Indicators, refer to a set of quantifiable measurements used to gauge an organization's overall long-term performance (Twin, 2023).

ETL : stands for Extract, transform, and load, is a data process that combines data from multiple sources into one single data storage unit, which is then loaded into a data warehouse or similar data system (Frankenfield, 2022).

CSP : stands for Cloud Service Provider an Information Technology (IT) company that provides its customers with computing resources over the internet and delivers them on-demand. (Rouse, 2022)

AWS : stands for Amazon Web Services, the cloud platform offered by Amazon in 245 countries. It's made up of many different cloud computing products and services in 81 availability zones in which its servers are located. These products and services include storage, networking, remote computing, email, mobile development, and security (Page, 2022).

List of Figures

List of Tables

1. Introduction

Feader, is a UK-based and registered non-store online retail company. The company mainly sells unique all-occasion gifts. Started to operate in the beginning of 2020, the company started gain attention from around the globe. The founders decided to adopt Business Intelligence (BI) to better understand and leverage their data. This report is intended to outline the implementation of a corporate data warehouse to provide strategic insights to decision-makers within the company. The project's primary goal was to offer the company a centralized repository of data that is clean, reliable, and trustworthy. This data warehouse is designed to be scalable to accommodate the growing volume of data as the business expands, while also adhering to data protection and privacy regulations to safeguard sensitive information. Additionally, the project is set to make the data readily accessible for analytics, enhancing the organization's ability to derive actionable insights. After gathering the stakeholders and gauging their readiness through a suitable assessment, a close description to the potential source systems within the organization is provided as they form the basis of our data warehouse, followed by an implementation strategy. This strategy considers the optimal choice between on-premises infrastructure and the cloud, exploring their advantages and disadvantages in terms of costs, performance, and availability. The selection of the most suitable architecture for the business needs is then discussed, followed by its implementation in the preferred infrastructure. As an illustrative example, a Key Performance Indicator (KPI) is introduced, providing a concrete example of the necessary Extract, Transform, and Load (ETL) process, shedding light on the steps involved, as well as the advantages and potential challenges.

2. Stakeholders

Supported by the founders, the organization is interested on integrating 3 departments in this data warehouse including marketing, finance, and sales with the latter one being a priority. For this matter, the data scientists, and analysts along with their head of departments were invited to take part of the meetings to define business requirements for the project. The IT department is held responsible for providing a dedicated project team, including a competent team manager. Table 1 define the stakeholders for this project. It’s worth noting that selecting a project manager with prior expertise in data warehousing projects is crucial, as these systems require unique challenges distinct from typical software development projects (Ponniah, 2010).

3. Assessment of readiness

Before carrying project requirement meetings, the project manager delivered a formal presentation to the stakeholders. This presentation served to showcase the project manager's existing knowledge of crucial data warehousing concepts. Subsequently, a two-week training program is organized with the assistance of the Lead Trainer. This training is followed by a one-on-one readiness assessment conducted by the project manager. The assessment aimed to measure the stakeholders' dedication and understanding, showing notably high levels of positivity.

IT team	Sales	Finance	Marketing
Project manager	From each department, their Head of Department, a Data scientist, and a Data analyst.		
Lead architect			
Data warehouse administrator			
Infrastructure specialist			
Business analyst			
Data modeler			
User liaison manager			
Data transformation specialist			
Quality assurance analyst			
Development programmer			
Testing coordinator			
End-user apps specialist			
Lead trainer			

Table 1: Stakeholders of the project. Source: Own representation based on Ponniah (p.89, 2010)

4. Business requirements

Ponniah (2010) emphasizes that it is crucial to prioritize business requirements over technical ones. Effective development of a data warehouse requires a full perception of business needs focusing on the information users truly seek rather than adding excessive features. Regular meetings, applying the agile project management approach, really facilitated the maintenance of a close feedback loop among stakeholders. As a result, the final requirements emerged as follows:

- The project must demonstrate rapid results.
- The sales department is a priority.
- Ability to identify historical best-selling products and their specific sales locations.
- Capability to track and analyse monthly revenue on a per-city or per-country basis.
- The system should be optimized to support data analytics and predictive modelling.
- Implementation without upfront costs and a commitment to continuous maintenance.

5. Designing the data warehouse

While many architectural approaches exist for building a data warehouse, they typically contain the same fundamental layers as represented in Figure 1. These essential layers include:

- Source layer : refers to the source of the data.
- Staging area : refers to where data is temporarily stored before loading process.
- Storage layer : refers to where processed data is stored.
- Presentation layer : refers to the utility of BI tools in reporting and analysis.

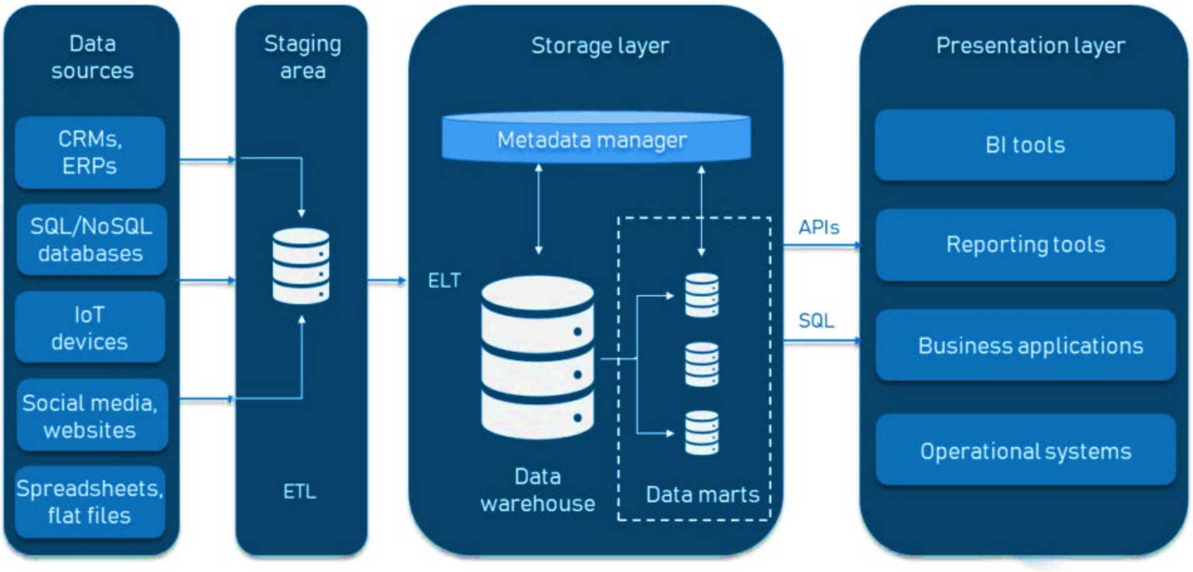


Figure 1: Basic layers of data warehouse. *Source: Altexsoft.com.*

5.1 Data sources

The fundamental success of a data warehouse is based upon the acquisition and proficient use of data. To accomplish this goal, comprehensive identification of the necessary data, its sources, structure, and the hardware related to its storage is given in table 2.

Data Sources	Data Structures	Hardware	Historical data
ERP System	- Comma separated values files (.csv)	*Server infrastructure:* - Amazon EC2 m7gd.4xlarge. - 1TB NVMe SSD. - 64 GB DDR5 RAM. *Database server:* - Amazon RDS for Oracle. *Network*: - Up to 10 Gbps for both upload and download speeds. *Operating system*: - Amazon Linux.	Available
CRM System			
HR System			Unavailable
Online click-stream data	- JavaScript Objects Notation files (.json)		Available

Table 2: Overview of data sources. *Source: Own representation.*

In our project, the sources responsible for supplying data to our data warehouse include:

- ***Enterprise Resource Planning*** (ERP) provide a centralized database that integrates multiple functions and departments within an organization, including finance, inventory management, manufacturing, and sales.

- ***Customer Relationship Management*** (CRM) allows businesses to store and manage customer-related data, including contact information, purchase history, preferences, interactions, and support tickets in one place.
- ***Human Resources System*** are designed to manage employee information such as personal details, payroll, benefits, performance evaluations, and training records.
- ***Online clickstream data*** refer to the collection, measurement, analysis, and reporting of various data related to how users interact with a website.

5.2 Data architecture

Many considerations need to be met to build an organization's data warehouse. Some of these considerations relates to the approach, are we intending to build individual business-specific data marts and later integrate them in a central data warehouse or to build the latter one first and derive data marts from it. A distinction is set to be made here between the terms since they're used interchangeably. Table 3 illustrates the major differences.

Features	Data Warehouse	Data Mart
Type	Enterprise-wide	Department-specific
Business processes	Multiple business processes	Single business process
Data source	Staging area	Data warehouse/staging area
Data modelling	ER model	Facts & dimensions
View of the data	Corporate	Departmental

Table 3: Data warehouse vs. data mart. Source: Own representation based on Ponniah (2010).

5.2.1 Kimball methodology

Known as the bottom-up approach by Ralph Kimball. In this method, the corporate data warehouse is a collection of departmental data marts, created one at time based on a pre-established schema determining the order of creation based on business needs. Table 4 lists both advantages and disadvantages of this methodology.

Advantages	Disadvantages
Faster and easier implementation.	Data redundancy.
Less risk of failure.	Data inconsistency.
Capability to schedule important data marts first.	Each data mart has its own narrow view of data.

Table 4: Advantages and disadvantages of Bottom-up approach. Source: Own representation based on Ponniah (2010).

The data marts are made for specific business cases using a special way of organizing data. Some of the architectures depending on this methodology are:

- ***Individual data marts****:* the data warehouse is a collection of unconnected data marts, each serving a specific department.
- ***Federated data marts****:* the data warehouse is a collection of federated data mart i.e., users are allowed to access and utilize shared data elements.
- ***Data-mart bus***: all data marts form one single data warehouse because the business dimensions and measured facts are conformed and linked among the data marts. These data marts may serve the entire enterprise, not just single departments.

5.2.2 Inmon methodology

On the other hand, Bill Inmon introduced the top-down approach. The data warehouse in this approach is the central repository for the entire enterprise, in where the data is stored at the lowest level of granularity based on a normalized data model. Table 5 lists both advantages and disadvantages of this methodology.

Advantages	Disadvantages
An enterprise view of data.	Time consuming to build.
Single, central storage of data.	High exposure to risk of failure.
Centralized rules and control.	High outlay without proof of concept.

Table 5: Advantages and disadvantages of Top-down approach. Source: Own representation based on Ponniah (2010).

In this approach, the aim is to build the overall, large, integrated, enterprise-wide data warehouse. Architectures adopting this methodology are:

- ***Centralized***: no data marts, only a normalized central data warehouse forming one source of truth.
- ***Hub and spoke***: in addition to the data warehouse, dependant data marts exist and depend on the data warehouse for retrieving data.

5.2.3 The adopted methodology

In compliance with the business requirements, our team opted to embrace the Kimball methodology due to its capacity to yield immediate results particularly crucial for our startup's urgent business-specific processes. In comparison, the Inmon approach was deemed time-consuming and prone to errors, lacking the immediacy we required.

We picked the data-mart bus architecture, this decision was wise as it allows for a harmonious blend of both methodologies in the long term. Beginning with data marts enables us to gradually progress towards constructing a comprehensive central data warehouse. This approach facilitates a strategic evolution, combining the strengths of both methodologies to achieve our overarching goals effectively over time.

6. Implementation

6.1 Reasons for cloud adoption

We had a choice to make between on-premises infrastructure or cloud-based infrastructures. Table 6 includes a comprehensive comparison of the two infrastructures. Opting for an on-premises infrastructure isn't suitable for the company, considering its status as startup with limited human and financial resources. Investing in and maintaining on-premises infrastructure could burden the IT department with additional tasks and cause the company significant costs. On the other hand, cloud infrastructure appears to be a fitting solution by offering high availability rates, built-in multi-layered security solutions, and continuous hardware maintenance. Allowing users to pay only for the services utilized.

Features	On-premises	Cloud
Scalability	Manually up/downgrade of software/hardware is required.	Available instantly.
Availability	Depends on the software/hardware quality and the competencies of your IT-Team.	Up to 99,99% with leading service providers.
Security	Depends on the competencies of your IT-team.	Ensured by the Cloud Service Provider (CSP)
Performance	Excellent query performance (measured in milliseconds).	Excellent query performance in multiple geographic locations (measured in seconds).
Costs	Requires significant initial investments (hardware, IT team, training, etc.)	Pay only for the used services.

Table 6: Features comparison of on-premises and cloud solutions. Source: Own representation based on Scnsoft.com.

6.2 Prominent cloud service providers

There are many CSPs available in the market, the well-known are Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform. The lead architect and the infrastructure specialist of our team ruled in the favour of utilizing AWS, due to major benefits like the massive collection of services available, cost-effectiveness, security, superior performance, and scalability options to scale vertically and horizontally to manage the growing workloads.

6.3 Cloud implementation strategies

An overview of the collection of AWS Services implemented is illustrated in figure 2, while a summary of how these services function is described in the following:

- ***Amazon Simple Storage Service or S3***, is a highly scalable and durable storage solution for storing and managing large volumes of structured/unstructured data.
- ***AWS Glue*** is an ETL service that automates the process of preparing and transforming data, making it easier to extract data from various sources, apply transformations, and load it into a data warehouse or analytics platform.
- ***AWS Glue Data Catalog*** is a centralized metadata repository meant to store and organize metadata about various data sources, tables, and their schemas.
- ***Amazon Kinesis Data Streams*** eases the real-time ingestion of streaming large volumes of data, allowing users to collect and process continuous streams of data from multiple sources (Website clickstreams, application logs, IoT devices, …).
- ***Amazon Redshift*** is a cloud-based data warehousing service for analysing large datasets using SQL queries, offering high performance via columnar storage, parallel processing, scalability, and seamless integration with other AWS services.
- ***Amazon QuickSight*** is a cloud-based business intelligence service that helps users easily analyse and visualize data from various sources through interactive dashboards and visualizations.
- ***Amazon SageMaker*** is a managed machine learning service, simplifying the creation, training, and deployment of machine learning models at scale. It offers flexibility, scalability, cost-effectiveness, and built-in security features for efficient model development and deployment.

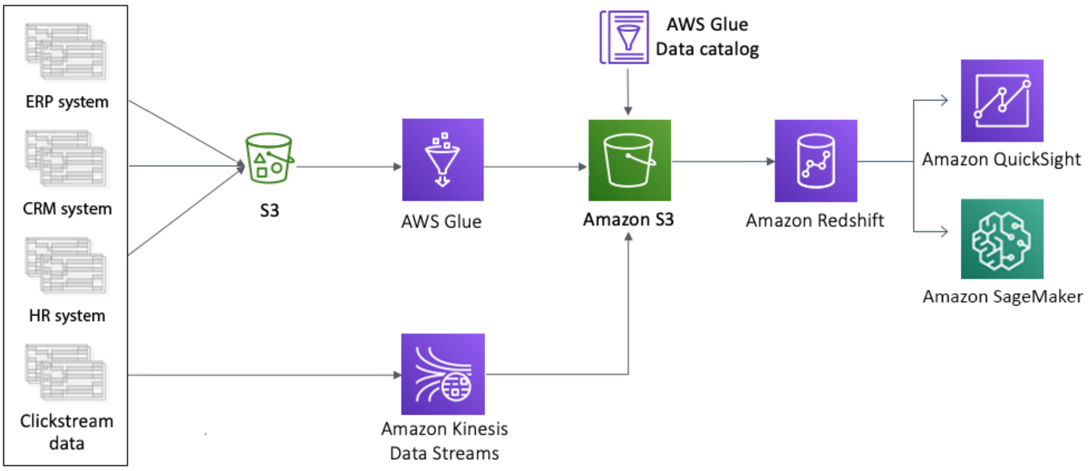
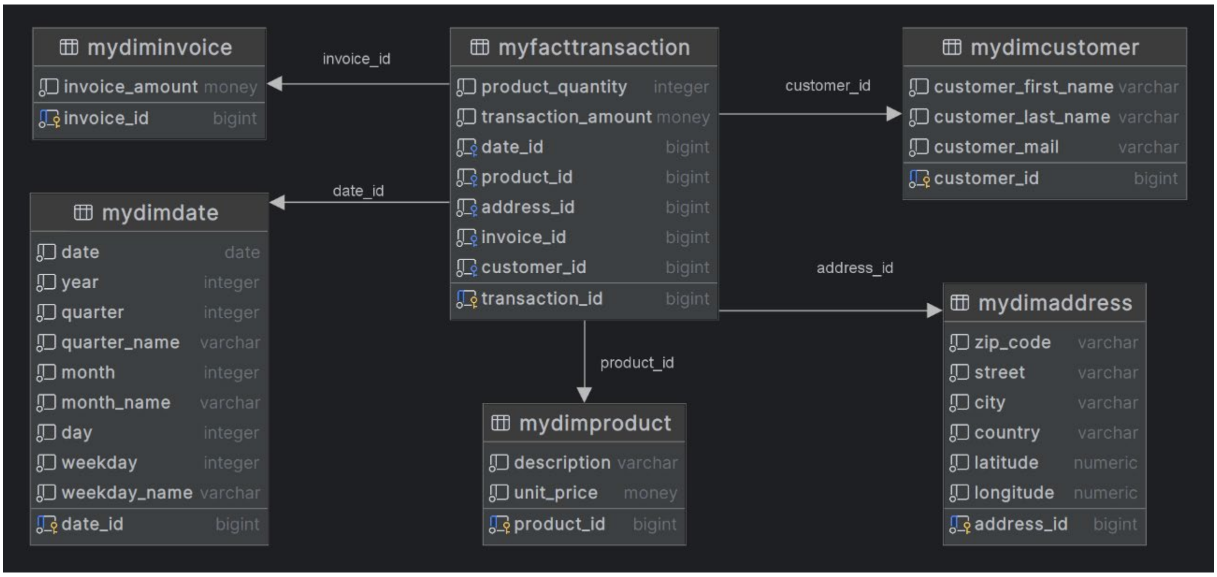


Figure 2: Data warehouse architecture in AWS cloud. *Source: Own representation.*

6.4 Example of KPI: Sales

The foremost among the KPIs, sales serve as the primary metric for evaluating the overall success of the company across various dimensions such as locations and timeframes. The necessary data for this metric is extracted from the raw formats of ERP and CRM systems and associated into an Amazon S3 bucket, functioning as a staging area. AWS Glue manages the data transformation procedures to ensure compliance with data quality standards. Simultaneously, the AWS Glue Data Catalog records and preserves metadata for each step of the process. S3 is employed to store the processed data and facilitate its integration with Amazon Redshift for executing parallel data loading. Subsequently, Amazon Redshift organizes and distributes data swiftly according to the schema depicted in Figure 3. At this point, the data is prepared for analysis within Amazon QuickSight or can serve as training datasets for machine learning models within Amazon SageMaker



mydiminvoice
invoice_amount money
invoice_id bigint

myfacttransaction
product_quantity integer
transaction_amount money
date_id bigint
product_id bigint
address_id bigint
invoice_id bigint
customer_id bigint
transaction_id bigint

mydimcustomer
customer_first_name varchar
customer_last_name varchar
customer_mail varchar
customer_id bigint

mydimdate
date date
year integer
quarter integer
quarter_name varchar
month integer
month_name varchar
day integer
weekday integer
weekday_name varchar
date_id bigint

mydimproduct
description varchar
unit_price money
product_id bigint

mydimaddress
zip_code varchar
street varchar
city varchar
country varchar
latitude numeric
longitude numeric
address_id bigint

Figure 3: The Star schema of the sales data mart. *Source: Own representation.*

Displayed in Figure 4 is the sales dashboard representing the company's performance over the past eighteen months. An illustrative demonstration of the effectiveness of BI can be found in this example. The map illustrates the countries contributing the most to sales, indicating the UK as the company's largest market. However, the bubble plot reveals that the primary cities driving sales are Amsterdam, Barcelona, and Porto. Without visualizing the bubble plot, one might have assumed that cities in the UK like Manchester, London, or Birmingham would dominate. This information supports the decision to develop new products associated with Amsterdam, Barcelona, and Porto. This example underscores the importance of uncovering such details. There are undoubtedly more insights awaiting discovery, and these revelations could further enhance the company's profits.

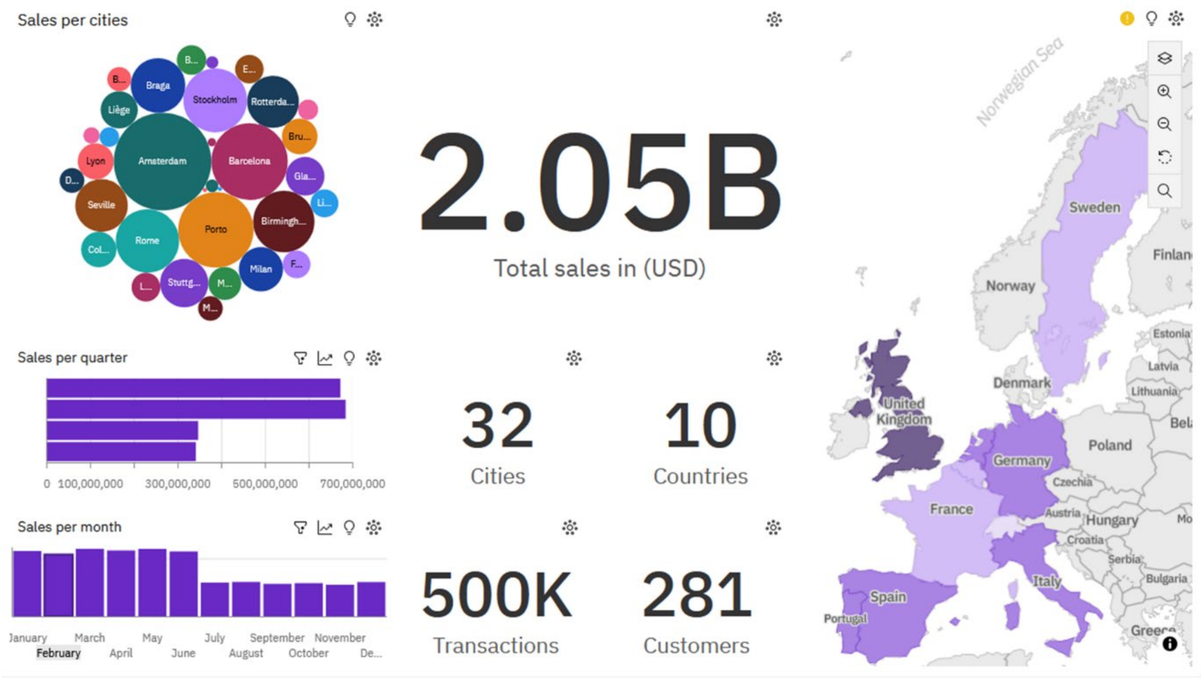


Figure 4: Sales dashboard. *Source: Own representation.*

7. Project schedule

As shown in table 7, the project was delivered on the span of 10 weeks. Management of the project was done via Microsoft Project, enabling precise allocation of tasks, their respective team members, and deadlines. Permitting team collaboration and facilitated the tracking of task progress.

Events	Duration	Start date	End date
Formal presentation	1 day	01/09/2023	
Stakeholders Training	5 days	04/09/2023	08/09/2023
Assessment of Readiness	1 day	08/09/2023	
Requirements definition	2 days	11/09/2023	12/09/2023
Data Source identification	3 days	13/09/2023	15/09/2023
Data Modelling	3 days	18/09/2023	20/09/2023
Designing Cloud Architecture	7 days	21/09/2023	29/09/2023
Implementation	2 days	02/10/2023	03/10/2023
Data warehouse testing	3 days	04/10/2023	06/10/2023
Initial loads	3 days	09/10/2023	11/10/2023
Ongoing loads configuration	2 days	12/10/2023	13/10/2023
Monitoring the data warehouse	15 days	16/10/2023	03/11/2023

Table 7: The project timeline. *Source: Own representation.*

8. Risk analysis

Delivering such projects often involve a tremendous focus on the financial resources. However, the exclusive attention on cost without recognizing the value gained and achieving a fine balance between these aspects is not recommended. Therefore, both financial considerations and the added value were precisely addressed into the project's risk assessment.

8.1 Potential benefits

Several benefits that could be derived from the data warehouse can't be accessible if the latter one didn't take place. The major benefits are:

- ***Centralized source of truth***: A data warehouse provides a centralized location for storing structured and often transformed data from various sources. Without it, data might remain scattered across disparate systems.
- ***Improved Decision Making***: Data warehouses enables quick access to organized data, enabling faster and informed decision-making processes. The lack of it may result in delays or inaccuracies in decision-making due to fragmented or incomplete data.
- ***Enhanced BI and Analytics***: Data warehouses are crucial for robust BI and analytics initiatives, empowering the use of robust tools and technologies for data analysis, visualization, and reporting, leading to actionable insights. Without it, performing these types of analytics becomes more challenging.
- ***Data Consistency and Quality***: Data warehouses often incorporate processes for data cleaning, transformation, and standardization, ensuring data consistency and quality. Without these processes in place, data might be inconsistent.
- ***Predictive and Prescriptive Analytics***: Data warehouses are involved in performing advanced analytics, including predictive and prescriptive analytics. The lack of a data warehouse limits the ability to leverage advanced analytical techniques effectively.

8.2 Required funds

The expenses for this project were notably budget-friendly in comparison to other projects that would typically endure monthly bills amounting to our annual costs of ***15,436.80 USD***. Furthermore, the service is designated for the "eu-west-2" region in London, thereby relieving concerns regarding data protection regulations. Table 8 describes cost estimations of the provided services and their respective configurations.

Service	Configuration summary	Monthly cost
Amazon Simple Storage Service (S3)	- 2 TB per month of S3 Standard storage. - 100 GB per month of data returned by S3 Select. - 1000 requests PUT, COPY, POST, LIST. - 1000 GET, and all other requests from S3 Standard. - 100 GB per month of data scanned by S3 Select.	47.38 USD

AWS Glue	- 10 DPUs for Apache Spark job. - 0.0625 DPUs for Python Shell job. - 1 million objects stored per month. - 1 million access requests per month.	22.06 USD
Amazon Kinesis Data Streams	- 24 hours of data retention. - 100 records per second as baseline number. - 1000 records per second as peak number.	25.58 USD
Amazon Redshift	- 1 Node. - Instance type: ra3.xlplus.	792.78 USD
Amazon SageMaker	- Instance name: ml.c5.12xlarge. - 3 data scientists - 10 instances/data scientist of Studio Notebook. - 5 hours of Studio Notebook per day. - 1 day Studio Notebook per month.	367.20 USD
Amazon QuickSight	- 30 working days per month. - 10 GB SPICE capacity. - 1 author. - 5 readers.	31.40 USD
Total monthly cost: 1,286.40 USD		
Total 12 months cost: 15,436.80 USD		

Table 8: Cost estimation of the services. *Source: Own representation.*

9. Conclusion

We've been profoundly engaged with this project from its beginning because we recognized its significant value for the company. Knowing that half of such initiatives often fade, turning into data repositories lacking quality and accessibility (Ponniah, 2010), we're immensely grateful to the founders for their solid support throughout, rescuing this project from multiple setbacks. One common cause of these failures lies with the users of data warehouses, but our situation differs as our users primarily consist of experienced data scientists, analysts, and department heads with strong technical backgrounds. However, we acknowledge the necessity of continual training for new users as our user base expands. While the system requires time to showcase its worth, it's crucial to consistently consider enhancements, particularly when anticipating a transition to a data lake house, revealing hidden patterns and significant opportunities within unstructured data of varied formats. This move aligns well with our need for increased data storage, as data lakes demand more capacity, accommodating additional terabytes of information. The compliance to project deadlines was clear to everyone. Instead of burning the midnight oil to meet these deadlines, the team showcased exceptional efficiency by establishing clear task definitions from the beginning during the planning stages. Particularly admirable was the cost-effectiveness in managing financial resources, as we utilized only the necessary services, resulting in a significant reduction in expenses on our bills.

Bibiliography

Enterprise Data Warehouse (EDW) Full Guide. (n.d.). AltexSoft. Retrieved 26 November 2023, from https://www.altexsoft.com/blog/enterprise-data-warehouse-concepts/

Frankenfield, J. (2022a, September 14). *What Is a Data Warehouse? Warehousing Data, Data Mining Explained*. Investopedia. https://www.investopedia.com/terms/d/data-warehousing.asp

Frankenfield, J. (2022b, October 30). *What Is Business Intelligence (BI)? Types, Benefits, and Examples*. Investopedia. https://www.investopedia.com/terms/b/business-intelligence-bi.asp

Page, V. (2022, November 6). *What Is Amazon Web Services and Why Is It So Successful?* Investopedia. https://www.investopedia.com/articles/investing/011316/what-amazon-web-services-and-why-it-so-successful.asp

Ponniah, P. (2010). *Data Warehousing Fundamentals for IT Professionals* (2nd Edition). Wiley. https://www.wiley.com/en-us/Data+Warehousing+Fundamentals+for+IT+Professionals%2C+2nd+Edition-p-9780470462072

Rouse, M. (2022, July 11). Cloud Service Provider. *Techopedia*. https://www.techopedia.com/definition/133/cloud-provider

Top 6 Cloud Data Warehouse Solutions in 2023. (n.d.). Retrieved 5 December 2023, from https://www.scnsoft.com/analytics/data-warehouse/cloud

Twin, A. (2023, May 10). *Key Performance Indicator (KPI): Definition, Types, and Examples*. Investopedia. https://www.investopedia.com/terms/k/kpi.asp