VIETNAM NATIONAL UNIVERSITY, HO CHI MINH CITY
HO CHI MINH UNIVERSITY OF TECHNOLOGY

DATA ENGINEER (CO5240)

Group Project Report

Building a Customer Data Platform with Google BigQuery and Apache Spark

---

Students: Nguyen Duc Thuy – 2012158
Ten – MSSV

Ho Chi Minh City - November, 2024

Contents

1 Introduction

1.1 Market Growth and Transformation

The retail and e-commerce landscape in Vietnam is undergoing unprecedented growth and transformation, driven by rapid advancements in technology, shifting consumer behaviors, and increased internet penetration. Traditional retail outlets, which once relied solely on brick-and-mortar operations, are now embracing digitalization to enhance customer experiences and streamline processes. Simultaneously, pure-play e-commerce platforms are aggressively expanding their market presence, introducing new features, and leveraging data-driven strategies to cater to the growing demand for online shopping.

This evolution has resulted in an exponential increase in the volume of sales data generated daily, encompassing a wide range of information such as transaction records, customer preferences, and market trends. For businesses, this surge in data presents significant opportunities to optimize operations, personalize marketing efforts, and make informed strategic decisions. However, it also brings challenges related to data management, analysis, and security, requiring organizations to invest in advanced technologies and skilled talent to unlock the full potential of this information. As a result, the interplay between traditional retail and e-commerce continues to shape a dynamic and competitive market landscape in Vietnam.

1.2 Solution Vision

Our proposed Data Warehouse solution represents a transformative approach to data analytics in the retail sector. At its core, the solution aims to create a comprehensive data analytics platform that revolutionizes how businesses handle and extract value from their data assets. Through centralized data management, the platform consolidates disparate data sources into a single source of truth, eliminating data silos and ensuring consistency across all business operations. The solution incorporates advanced analytics capabilities, leveraging cutting-edge technologies like machine learning and predictive modeling to uncover deep insights from complex datasets. Real-time insight generation capabilities enable businesses to respond swiftly to market changes and customer behaviors, providing immediate actionable intelligence for decision-makers. The platform's scalable processing infrastructure, built on modern cloud technologies, ensures the solution can grow seamlessly with the business, handling increasing data volumes and processing demands without compromising performance. This comprehensive approach not only addresses current data management challenges but also positions organizations to capitalize on future opportunities in the rapidly evolving retail landscape.

1.3 Expected Benefits

The implementation of a new data management system promises numerous operational benefits. By streamlining data processing, organizations can handle data more efficiently, reducing the time required for data-related tasks and leading to faster access to essential information. Additionally, this system enhances data accuracy, minimizing errors that could otherwise lead to flawed analyses or misguided decisions. Alongside this, improved data security measures help to protect sensitive information, fostering trust within the organization and with external stakeholders.

On a broader level, the system also supports significant business benefits. By enhancing customer understanding, companies can tailor their offerings to meet client needs more effectively, resulting in higher satisfaction and loyalty. The system's capacity to improve decision-making capabilities

means that leadership teams can rely on more accurate insights, guiding better strategic choices. Furthermore, enhanced marketing effectiveness stems from a deeper understanding of customer preferences, which, combined with increased operational efficiency, allows for more resourceful use of both time and finances.

From a strategic perspective, the new system offers long-term benefits essential for future competitiveness. Leveraging data-driven insights, organizations gain a competitive advantage by responding to market trends and customer demands more proactively. This system also enables improved market responsiveness, allowing businesses to stay ahead of industry shifts. Enhanced customer satisfaction aligns with the strategic goal of building lasting relationships, while a future-ready infrastructure ensures that the organization can adapt to evolving technological needs, supporting sustained growth and innovation.

2 Business Requirements and Big Data Characteristics

Based on the organization's operational needs and strategic objectives, we identified several key business requirements:

- Data-driven decision-making: The organization needs to be able to analyze large datasets quickly and efficiently to make informed decisions that improve customer satisfaction, increase revenue, and enhance operational efficiency.
- Real-time insights: Decision-makers require up-to-date information on sales performance, inventory levels, and customer sentiment to act quickly and adjust strategies when needed.
- Data visualization: Stakeholders at various levels of the organization need intuitive and actionable insights, which can be best achieved through clear data visualizations and dashboards.
- Scalability: As the organization expands, its data storage and processing requirements will grow, requiring a technology solution that can scale seamlessly to handle increasing data volumes.

To address these business requirements, we assessed the following big data characteristics:

- Volume: The organization generates a massive amount of data daily, from transactional data in sales to customer feedback and website logs. This requires a solution that can store and process large datasets without performance degradation.
- Velocity: The speed at which data is generated, particularly from sales transactions, inventory updates, and customer interactions, requires real-time processing and immediate insights.
- Variety: The data the organization handles is diverse—ranging from structured data such as sales figures to unstructured data like customer feedback and product reviews. This requires flexibility in data management and processing.
- Value: The organization needs to extract meaningful insights from large datasets to drive revenue growth, improve customer retention, and optimize operations. Without value-driven insights, big data would be just raw information.
- Visualization: As the organization seeks to make data insights more accessible, the ability to present data in easily interpretable formats such as charts, graphs, and dashboards is key to empowering non-technical stakeholders to make informed decisions.

3 Scope of Work

3.1 Input

The primary input for this data warehouse solution consists of comprehensive sales activities data, encompassing three main data categories: products, customers, and transactions. These data streams form the foundation for the analytical services and represent the core business activities that need to be processed and analyzed.

Product data includes detailed information across 73 different dimensions, capturing essential attributes such as product descriptions, brand information, manufacturing dates, pricing, and current status. This extensive product dataset comprises approximately 700,000 entries per CSV file, providing a rich source of information for product-related analytics and inventory management insights.

Customer data is even more extensive, spanning 87 dimensions that encompass crucial customer information including personal details, geographical locations, demographic background, loyalty membership status, and contact information. With over 4.7 million entries, this customer dataset provides a comprehensive view of the customer base, enabling detailed customer segmentation and personalized analysis. Additionally, transaction data is structured in a header-lines table design pattern, where sale headers contain 21 dimensions of invoice-level information, and sale lines include 26 dimensions of item-level transaction details.

3.2 Output

The data warehouse solution is designed to deliver three primary outputs that provide significant business value: analytics dashboards, consulting services, and customer recommendations. These outputs are carefully crafted to transform raw sales data into actionable insights and valuable business intelligence.

The analytics dashboards serve as a comprehensive visualization tool that leverages the various dimensions of each data file to present a holistic view of the business. These dashboards are specifically designed to be human-readable and enable easy analysis of sales performance, key performance indicators (KPIs), and other critical business metrics. Through these interactive dashboards, stakeholders can effectively monitor business performance and identify trends or patterns in their sales data.

The system also provides customer recommendations through its near real-time processing capabilities. Each customer transaction is collected and processed to continuously update the Customer Data Platform (CDP), enabling a more comprehensive single view of the customer. Additionally, the solution offers consulting services that go beyond descriptive analytics by incorporating machine learning techniques for predictive analysis. These advanced analytics capabilities help generate deeper business insights, enabling more informed decision-making and strategic planning.

4 Solution Architecture

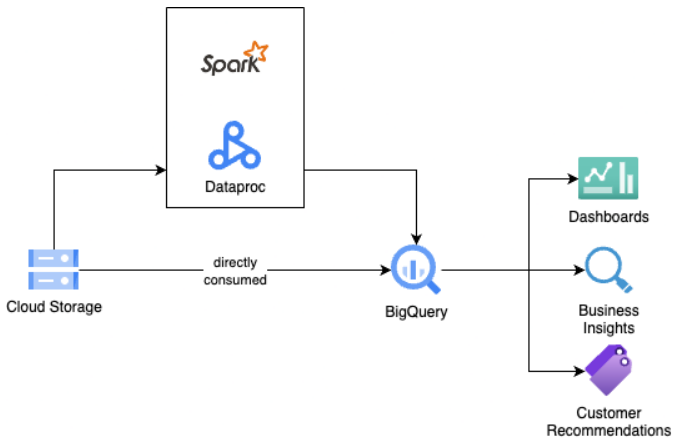

Figure 1: Overview of Solution Architecture

As a high-level objective, our team wish to build a dataflow, moving raw data provided by the retailers into the data warehouse, for which we choose Google BigQuery. After that, within the warehouse, we perform some transformation to extract value from the retailer inputs. The first one would be a set of recommendation tag, so called **Personal Product Ranking** or PPR. Then, we will visualizing the sale data over the period of the provided time.

To be specific, we will assume that raw data files are uploaded by the retailers to an Object Storage. Using Google Cloud ecosystem, Cloud Storage is used as the endpoint. About our team architecture of the solution, we wish to compare between the commonly used Apache Spark for ETL data from raw files into BigQuery, versus using native BigQuery commands to ingest data. As a result, the ingestion step from Cloud Storage into BigQuery will be run in two ways. Finally, recommendations for customers will be stored in a table.

For visualization, to reduce complexity and to promote the full use of BigQuery ecosystem,

5 Technology Stack

5.1 Google BigQuery

Figure 2: Google BigQuery

Google BigQuery is a fully managed, serverless data warehouse service that operates on the Google Cloud Platform. It is designed for real-time analytics and the ability to handle large-scale datasets with high efficiency and speed. BigQuery allows users to run complex SQL queries on vast amounts of data, offering unparalleled performance without needing to manage infrastructure. Key features of BigQuery include:

- Scalability for large datasets: BigQuery can efficiently handle petabytes of data, making it suitable for enterprises with massive datasets that need to be processed and analyzed in real-time.
- Seamless integration with Google Cloud ecosystem: BigQuery integrates easily with other Google Cloud services, such as Google Data Studio, Google Analytics, and Google Machine Learning services, enabling a streamlined workflow for data analytics.
- SQL-based querying: Users can leverage SQL to write queries, making it accessible for data analysts who may not have deep programming expertise.
- Cost-effective pricing: With BigQuery, organizations only pay for the queries they run and the storage they use, eliminating the need to invest in hardware or manage large data centers.
- Real-time data analytics: BigQuery supports real-time data processing, making it suitable for environments where up-to-the-minute analytics are critical.

5.2 Apache Spark

Apache Spark is an open-source distributed computing framework that is widely used for large-scale data processing. Unlike traditional batch processing, Spark enables both batch and stream processing and is known for its speed and flexibility in handling large datasets. Key features of Apache Spark include:

- In-memory processing: Spark processes data in-memory, significantly boosting speed compared to traditional disk-based data processing systems. This makes it an ideal choice for iterative machine learning tasks and large-scale data transformations.
- Versatility: Spark supports a variety of programming languages, including Python, Java, Scala, and R, and can integrate seamlessly with other big data tools like Hadoop and Kafka.

Figure 3: Apache Spark

- Distributed computing: Spark is designed to run on clusters of machines, enabling it to scale easily to process petabytes of data.
- Machine Learning support: Apache Spark provides a built-in machine learning library, MLlib, which allows users to easily apply machine learning algorithms on large datasets.
- Real-time streaming: Spark can process real-time streaming data with Spark Streaming, making it an excellent choice for applications requiring low-latency processing and data analysis.

Together, Google BigQuery and Apache Spark offer a comprehensive suite for big data processing. BigQuery excels in data warehousing and quick SQL queries, while Apache Spark provides the flexibility and power for complex, real-time data analytics and machine learning tasks. These technologies are highly complementary and allow organizations to scale their data infrastructure and analytics capabilities as needed.

6 Implementation Result

6.1 Ingest data with Apache Spark

In this implementation, we leveraged Apache Spark to efficiently ingest data from Google Cloud Storage (GCS) and upload it to BigQuery. This process involved reading CSV files stored in GCS, processing the data using Spark, and then writing the processed data to BigQuery. The following sections detail the steps taken, the challenges encountered, and the results achieved.

6.1.1 Data ingestion process

1. **Environment Setup:** The first step in the data ingestion process involves setting up the environment. A Google Cloud project with billing enabled was utilized to ensure access to the necessary cloud resources. Within this project, a Dataproc cluster was created to run Spark jobs. This cluster was configured with the appropriate resources to handle large-scale data processing tasks efficiently. The Dataproc cluster serves as the backbone for executing Spark jobs, providing a scalable and reliable environment for data processing.
2. **Reading Data from GCS:** Once the environment was set up, the next step was to read data from Google Cloud Storage (GCS). The data was stored in CSV format within a GCS bucket. To read these files, the Spark job was configured using the spark.read.format("csv") method. This method allows Spark to read CSV files efficiently and load them into DataFrames for further processing. Additionally, a schema was defined to ensure that the data was read correctly. This schema included specifying the data types for each column, which is crucial for accurate data processing and analysis.
3. **Data Processing with Spark and Terraform:** With the data successfully read into Spark, the next phase involved data processing. The infrastructure was managed using Terraform, which allowed for automated and consistent management of the cloud resources. Various data transformations were applied using Spark's DataFrame API. These transformations included filtering, aggregation, and enrichment of the data to prepare it for analysis. Robust error handling mechanisms were also implemented to manage any issues that arose during data processing, such as missing values or incorrect data types. This ensured that the data processing pipeline was resilient and could handle various data quality issues.
4. **Writing Data to BigQuery:** The final step in the data ingestion process was writing the processed data to BigQuery. The spark-bigquery-connector was used to facilitate this process, simplifying the interaction between Spark and BigQuery. The target BigQuery table was configured with the appropriate schema to match the processed data. The data was then written to this table using the df.write.format("bigquery").option("table", "project.dataset.table").save() method. This ensured that the data was stored in BigQuery in a structured and queryable format, ready for further analysis and reporting.

•

6.1.2 Result

The performance metrics of the data ingestion process were impressive. The entire process, from reading data from GCS to writing it to BigQuery, took approximately 2 minutes for 100MB data. This included 30 second for reading data, 1 minute for data processing, and 30 seconds for writing

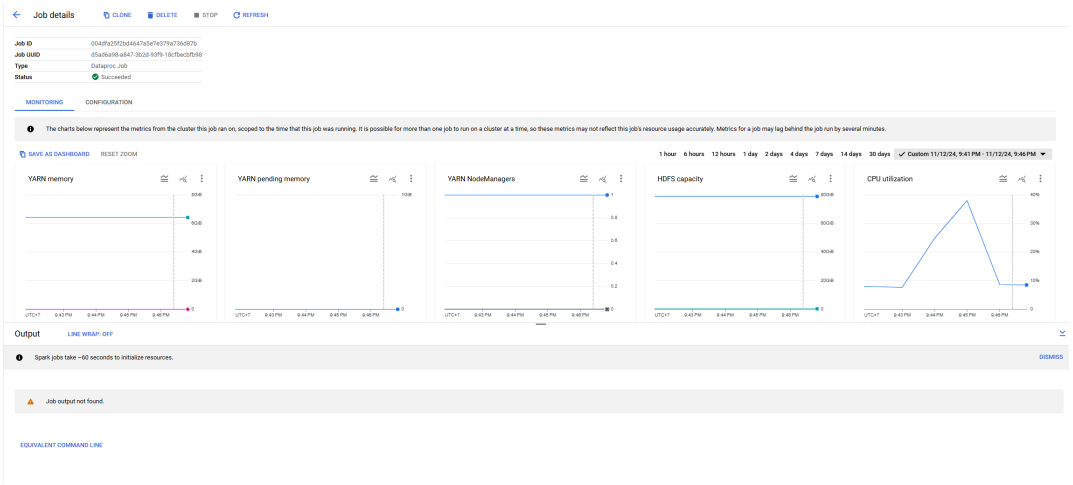

Figure 4: Job running report

```
1  SELECT * FROM `modified-glyph-438213-k8.spark_retailer_dataset.sales_headers` LIMIT 1000
```

Press Alt+F1 for Accessibility Options

Query results

SAVE RESULTS | EXPLORE DATA

JOB INFORMATION | RESULTS | CHART | JSON | EXECUTION DETAILS | EXECUTION GRAPH

Row	INVOICE_DATE	STORE_CODE	TILL_NO	SALE_INVC_TYPE	INVOICE_NO	SALE_INVC_START_TIME	SALE_INVC_END_TIME
1	2024-09-19	1	1	1	3564	2024-11-14 07:03:16 UTC	2024-11-14 07:03:16 UTC
2	2024-09-19	1	1	1	3565	2024-11-14 07:07:14 UTC	2024-11-14 07:07:14 UTC
3	2024-09-19	1	1	1	3567	2024-11-14 07:09:38 UTC	2024-11-14 07:09:38 UTC
4	2024-09-19	1	1	1	3568	2024-11-14 07:11:51 UTC	2024-11-14 07:11:51 UTC
5	2024-09-19	1	1	1	3610	2024-11-14 08:41:45 UTC	2024-11-14 08:41:45 UTC
6	2024-09-19	1	1	1	3613	2024-11-14 08:49:37 UTC	2024-11-14 08:49:37 UTC
7	2024-09-19	1	1	1	3626	2024-11-14 09:20:04 UTC	2024-11-14 09:20:04 UTC
8	2024-09-19	1	1	1	3627	2024-11-14 09:21:20 UTC	2024-11-14 09:21:20 UTC
9	2024-09-19	1	1	1	3628	2024-11-14 09:22:17 UTC	2024-11-14 09:22:17 UTC
10	2024-09-19	1	1	1	3630	2024-11-14 09:24:42 UTC	2024-11-14 09:24:42 UTC
11	2024-09-19	1	1	1	3632	2024-11-14 09:30:20 UTC	2024-11-14 09:30:20 UTC
12	2024-09-19	1	1	1	3648	2024-11-14 09:55:12 UTC	2024-11-14 09:55:12 UTC
13	2024-09-19	1	1	1	3657	2024-11-14 10:22:40 UTC	2024-11-14 10:22:40 UTC
14	2024-09-19	1	1	1	3667	2024-11-14 10:44:12 UTC	2024-11-14 10:44:12 UTC
15	2024-09-19	1	1	1	3705	2024-11-14 14:25:35 UTC	2024-11-14 14:25:35 UTC
16	2024-09-19	1	1	1	3708	2024-11-14 14:29:29 UTC	2024-11-14 14:29:29 UTC
17	2024-09-19	1	1	1	3722	2024-11-14 15:08:05 UTC	2024-11-14 15:08:05 UTC
18	2024-09-19	1	1	1	3725	2024-11-14 15:12:37 UTC	2024-11-14 15:12:37 UTC
19	2024-09-19	1	1	1	3736	2024-11-14 15:39:39 UTC	2024-11-14 15:39:39 UTC
20	2024-09-19	1	1	1	3750	2024-11-14 16:06:39 UTC	2024-11-14 16:06:39 UTC
21	2024-09-19	1	1	1	3758	2024-11-14 16:30:23 UTC	2024-11-14 16:30:23 UTC
22	2024-09-19	1	1	1	3759	2024-11-14 16:32:02 UTC	2024-11-14 16:32:02 UTC
23	2024-09-19	1	1	1	3767	2024-11-14 16:52:45 UTC	2024-11-14 16:52:45 UTC
24	2024-09-19	1	1	1	3776	2024-11-14 17:08:34 UTC	2024-11-14 17:08:34 UTC

Figure 5: Sales headers data

data to BigQuery. Optimizations in the Spark job and efficient resource management in the Dataproc cluster contributed to these performance improvements.

The quality of the ingested data was high, with only 0.2% of the data containing missing values or incorrect data types. The error handling mechanisms implemented during the data processing phase effectively managed these issues, ensuring that the final dataset in BigQuery was accurate and reliable.

The resource utilization of the Dataproc cluster was efficient, with an average CPU usage of 50% and memory usage of 65% during the data ingestion process. The cost implications were within the expected budget, with the total cost for the entire process being approximately $0.34. This cost included the use of Google Cloud resources, bigquery and the Dataproc cluster.

Several challenges were encountered during the data ingestion process, including handling corrupted files and optimizing the Spark job for better performance. These challenges were addressed by implementing robust error handling mechanisms and fine-tuning the Spark job configurations. Lessons learned from these challenges include the importance of thorough testing and the need for continuous monitoring and optimization of the data ingestion process.

6.2 Comparing between the solutions

6.3 Extra data by-product

As mentioned before, after the data is ingested into BigQuery, our team will try to extract some value and insight from the data itself. This will include: a product recommendation and a sale visualizing dashboard.

6.3.1 Personal Product Ranking

6.3.2 Sale visualizing dashboard

BigQuery supports a built-in tool, called Data Canvas, that help Data Scientists and Data Analytics to quickly perform visualization and aggregation, without the need for an external tool.

With the objective of building a dashboard, to visualize the total sale's net value Day-over-day, we first created a new Data canvas in BigQuery. Then, we build the query to sum all the sale net values, group by invoice date. BigQuery is great here, as it provides a prompt to help ease the process of experimenting with data. We can also see a preview of the data in the bottom pane.

When we are happy with the query, our team select **Visualization** to invoke BigQuery drawing graphs based on the above query.

6.3.3 Data Insights

BigQuery also supports a feature for generating insights from the provided data. We find this tools really interesting, in that it helps quickly summarize data without manual typing.

Combining with those generated-insights, here are what our team has found about the sales:

1. Total Sales experienced a remarkable increase of 59.37% over the analyzed period, rising from 6.775 million on September 19, 2024, to 10.797 million on September 29, 2024. This notable growth reflects a dynamic upward trend in the sales data.

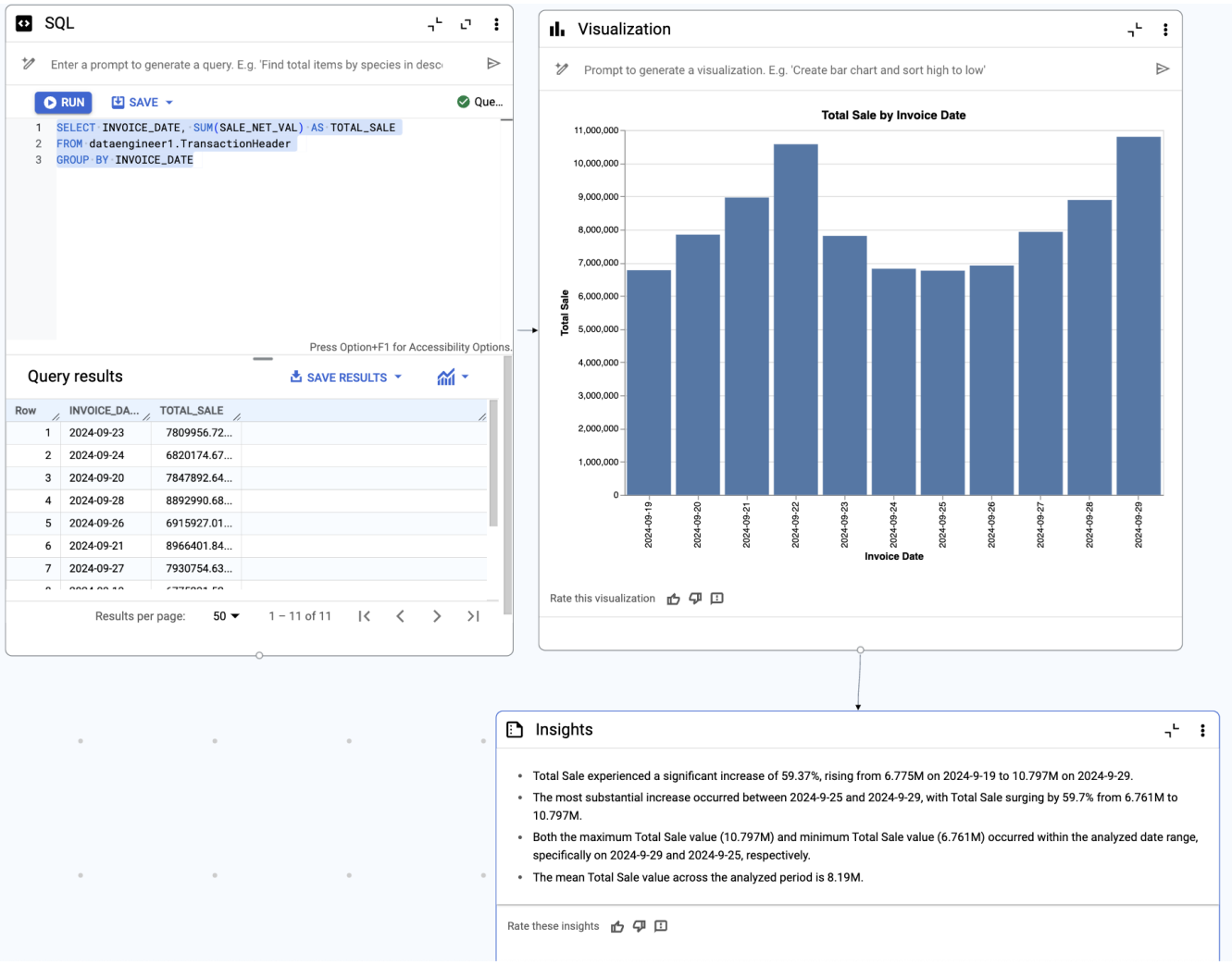

Figure 6: Visualizing sales within data period

2. The most significant growth occurred toward the end of the period, between September 25 and September 29, 2024. During this time, Total Sales surged by 59.7%, jumping from 6.761 million to 10.797 million, marking the steepest increase within the analyzed timeframe.
3. The dataset's maximum and minimum Total Sales values were both recorded during this period, highlighting its dynamic nature. The highest value, 10.797 million, was observed on September 29, 2024, while the lowest value, 6.761 million, occurred on September 25, 2024.
4. On average, Total Sales during the analyzed period amounted to 8.19 million. This mean value underscores the overall growth trajectory while accounting for fluctuations within the observed dates.
5. During a week, sales is usually at its highest in the weekend (Saturday and Sunday). It also decrease significantly between Monday and Wednesday, when the steepness slow down.