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 experiencing unprecedented growth and transformation. Traditional retail outlets are rapidly digitalizing their operations, while pure-play e-commerce platforms are expanding their market presence. This evolution has created an exponential increase in the volume of sales data generated daily, presenting both opportunities and challenges for businesses.

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 Scope of Work

2.1 Market Growth and Transformation

The retail and e-commerce landscape in Vietnam is experiencing unprecedented growth and transformation. Traditional retail outlets are rapidly digitalizing their operations, while pure-play e-commerce platforms are expanding their market presence. This evolution has created an exponential increase in the volume of sales data generated daily, presenting both opportunities and challenges for businesses.

2.2 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.

2.3 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 respond-

ing 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.

3 Solution Architecture

3.1 Market Growth and Transformation

The retail and e-commerce landscape in Vietnam is experiencing unprecedented growth and transformation. Traditional retail outlets are rapidly digitalizing their operations, while pure-play e-commerce platforms are expanding their market presence. This evolution has created an exponential increase in the volume of sales data generated daily, presenting both opportunities and challenges for businesses.

3.2 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.

3.3 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.

4 Implementation Result

4.1 Ingest data directly with BigQuery

4.2 Ingest data with Apache Spark

4.3 Comparing between the solutions

4.4 Customer Data Platform tags