Report

On

Cloud-Based Architecture for Retail Store

(DATA 040 -Data Delivery)

Submitted by:

Parjot Kaur

Table of Contents

Introduction

Retail establishments like Dollarama must efficiently handle and analyse enormous volumes of data from different sources, such as physical stores, supply chain systems, and e-commerce platforms, in the current digital era. We offer a cloud-based data engineering solution that makes use of Azure resources to build a scalable, effective, and reliable architecture in order to meet this goal. This solution's powerful analytics and visualization tools provide insightful information by processing data in both real-time and batch modes.

Objectives

Following some primary objectives of this architecture:

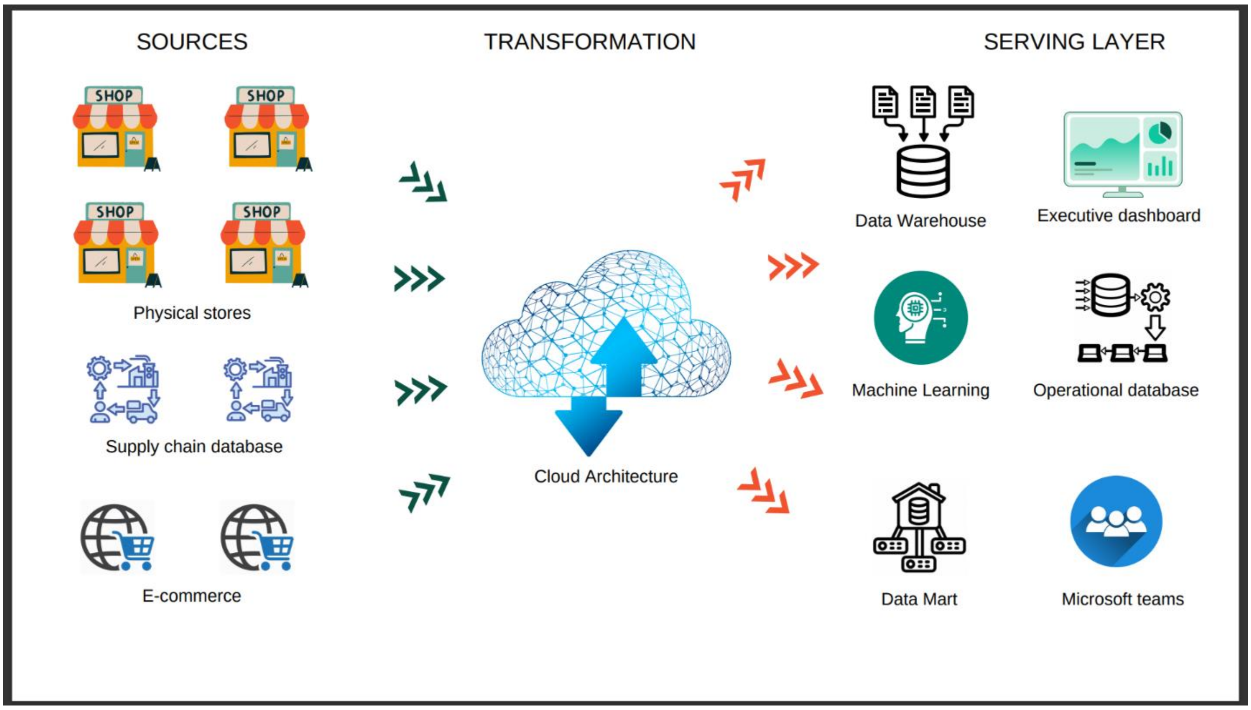
- **Improving Operational Efficiency:** By centralizing data from different sources, the architecture ensures that all business operations are informed by the most current data available. This streamlines decision-making and optimizes daily operations**.**
- **Leveraging Data Analytics for Store Expansion:** The system is designed to identify patterns and trends in the data that can inform decisions about where and when to open new store locations.
- **Implementing Real-Time Inventory Management:** Real-time data processing allows for accurate inventory tracking, helping stores to maintain optimal stock levels, reduce waste, and avoid stockouts.

Understanding the problem

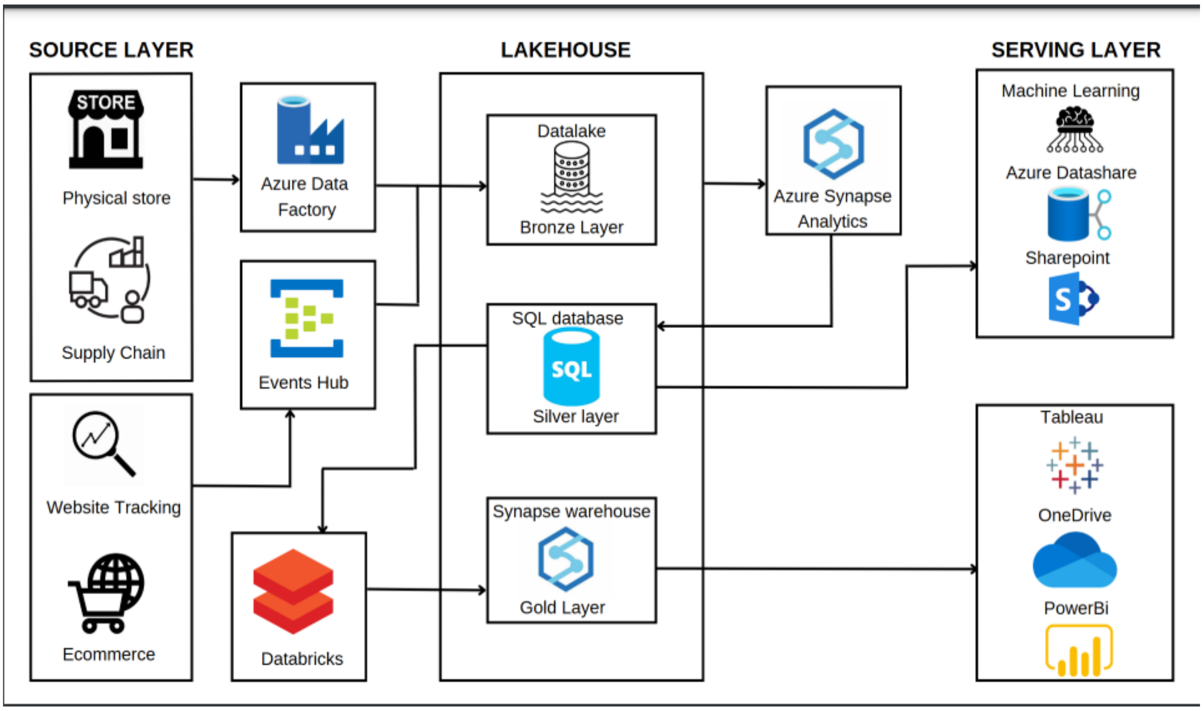
Managing data from several sources, such as in-store sales, online transactions, and supply chain operations, presents major issues for retail businesses. This data needs to be efficiently handled and analyzed because it frequently comes from several systems and formats. We can gather this data, carry out the necessary transformations, and derive useful insights that inform business choices by utilizing cloud-based technologies.

Cloud Architecture Vision

The core vision of this cloud architecture is scalability. As the business grows, so does the amount of data generated. The architecture is designed to evolve alongside the data, ensuring it remains effective and efficient over time. The process of Continuous Improvement and Development (CI/CD) is integral to this vision, allowing the architecture to adapt and improve with changing data needs.



CLOUD ARCHITECTURE



This architecture integrates various data sources, processing frameworks, and analytical tools to create a cohesive system that can handle the end-to-end data lifecycle.

1. Source Layer

The Source Layer is where the data originates. In a retail context, data can come from multiple sources:

- **Physical Store:** Data from in-store transactions, inventory systems, and customer interactions.
- **Supply Chain:** Data from logistics, shipping, inventory management, and supplier interactions.
- **Website Tracking:** Data on user behavior, page views, click-through rates, and more.
- **E-commerce:** Data from online transactions, customer profiles, product searches, etc.

2. Lakehouse (Processing and Storage Layer)

The Lakehouse layer is the central hub where data is ingested, processed, and stored. This layer integrates the benefits of both data lakes (for raw data) and data warehouses (for processed, structured data). In this architecture Azure data factory and Event hub used for the Data ingestion

Within the Lakehouse:

- **Bronze Layer (Data Lake):** Raw, unprocessed data is stored here. It acts as the initial landing zone for all ingested data, allowing for flexible storage without a predefined schema.
- **Silver Layer (SQL Database):** Data is cleaned, transformed, and stored in a structured format. This layer is optimized for querying and analytics.
- **Gold Layer (Synapse Warehouse):** Fully processed, high-quality data is stored here. This data is ready for advanced analytics, reporting, and machine learning models.

3. Serving Layer

The Serving Layer is where processed data is made available for business use. This layer includes tools for analytics, reporting, collaboration, and decision-making.

- **Azure Synapse Analytics:** A powerful analytics service that integrates big data and data warehousing to provide insights across all data in the system.
- **Machine Learning:** Processed data from the Gold Layer is used for training and deploying machine learning models.
- **Azure Datashare:** A secure data-sharing service that allows businesses to share data with external partners while maintaining control over the data.
- **SharePoint:** Used for document management and collaboration across teams.

For business intelligence (BI) and reporting:

- **Power BI:** A business analytics service that provides interactive visualizations and business intelligence capabilities.
- **Tableau:** Another powerful data visualization tool, often used for creating dashboards and reports.
- **OneDrive:** Cloud storage that facilitates collaboration and sharing of documents and reports across the organization.

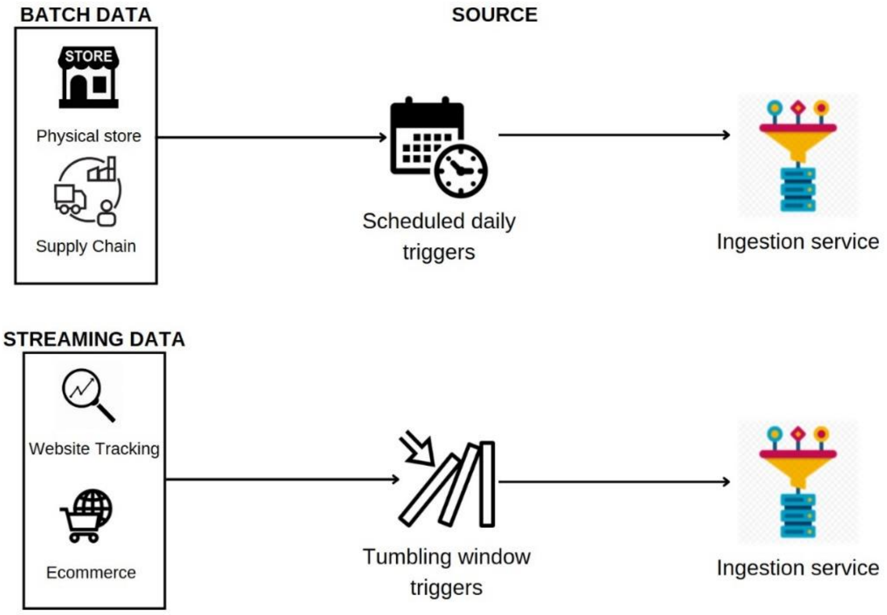
Pipeline Strategy

The data pipeline strategy is the heart of this architecture, ensuring that data flows smoothly from source to consumption. The pipeline strategy is describe below.

Source Layer

The source layer is where data is initially collected. There are two main types of data:

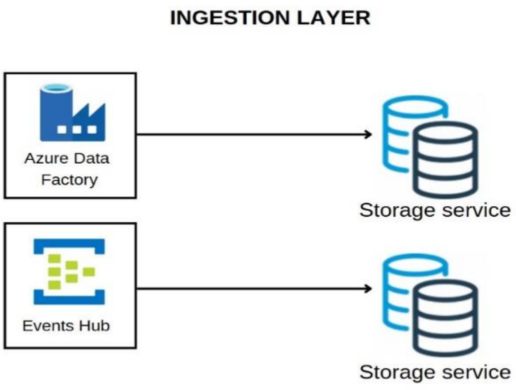
- **Batch Data:** This data is collected in daily batches, typically from physical stores or periodic reports.
- **Streaming Data:** This is real-time data, such as online transactions or real-time inventory updates.



Ingestion Layer

Once collected, the data is ingested into the system. The ingestion layer uses two primary Azure services:

- **Azure Data Factory:** Used for batch data ingestion, this service schedules and automates the movement and transformation of data.
- **Azure Event Hubs:** This service handles streaming data, enabling real-time data ingestion and processing.



Storage Layer

The data is then stored in different layers based on its level of processing:

1. Bronze Layer: Ingestion and Initial Storage

- **Data Sources:** The process begins with data ingestion from various sources, which could include transactional systems, logs etc.
- **Azure Data Lake Gen 2 and Blob Storage:** The ingested data is initially stored in the **Bronze Layer** within **Azure Data Lake Gen 2** or **Blob Storage**. This layer is typically used to store raw, unprocessed data as it arrives. The primary goal at this stage is to capture data in its original form without any transformations or cleaning.

2. Silver Layer: Data Transformation and Intermediate Storage

- **Transformation Process:** After the data is stored in the Bronze Layer, it undergoes a transformation process. This involves cleaning, filtering, and organizing the data to make it more useful for analysis.
- **SQL and NoSQL Databases:** The transformed data is then moved to the **Silver Layer**, which typically involves storing the processed data in structured formats within **SQL databases** or semi-structured formats within **NoSQL databases**.

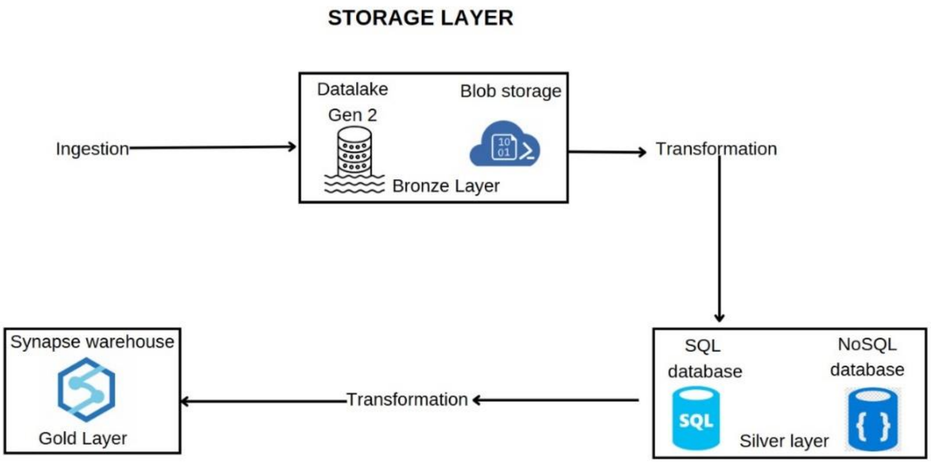
  The Silver Layer acts as an intermediate storage stage where data is ready for more complex transformations and aggregations.

3.Gold Layer: Final Transformation and Storage for Analysis

Further Transformation: Aggregations, dataset joins, and analytical view creation are a few examples of extra transformations that can be applied to data in the Silver Layer.

Synapse Warehouse (Gold Layer): The finished, completely changed data is kept in the Azure Synapse **Warehouse's Gold Layer**. The high-quality, polished data that is prepared for in-depth analysis and reporting is contained in this layer.

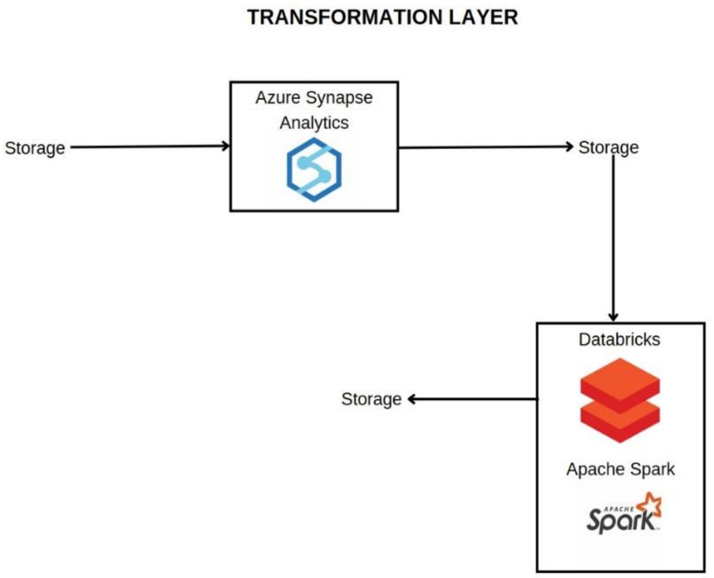
LINKEDIN



Transformation Layer

In the transformation layer, the data is cleaned, enriched, and aggregated to make it useful for analysis:

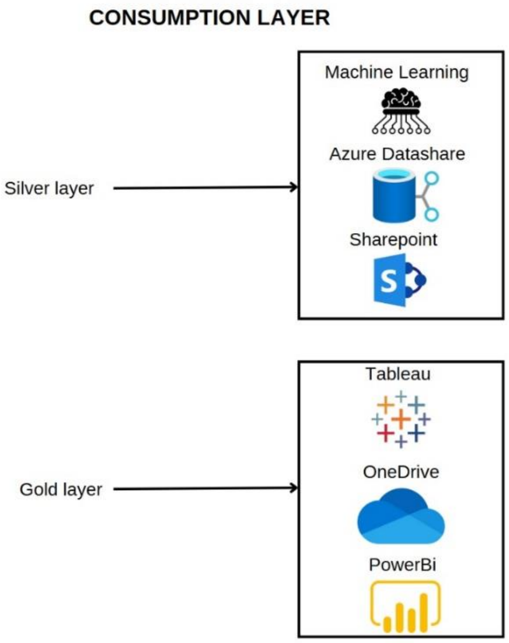
- **Azure Synapse Analytics:** Used in the Silver Layer for removing duplicates, handling missing values, and basic formatting.
- **Databricks:** Used in the Gold Layer for adding context, applying business rules, and ensuring consistency.



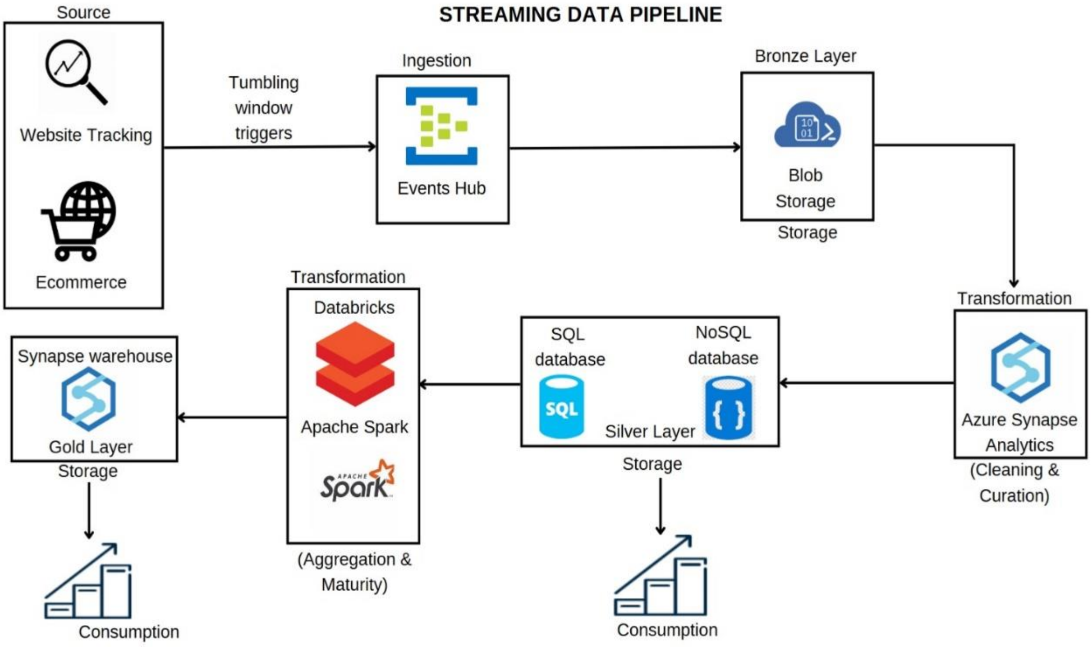
Consumption Layer

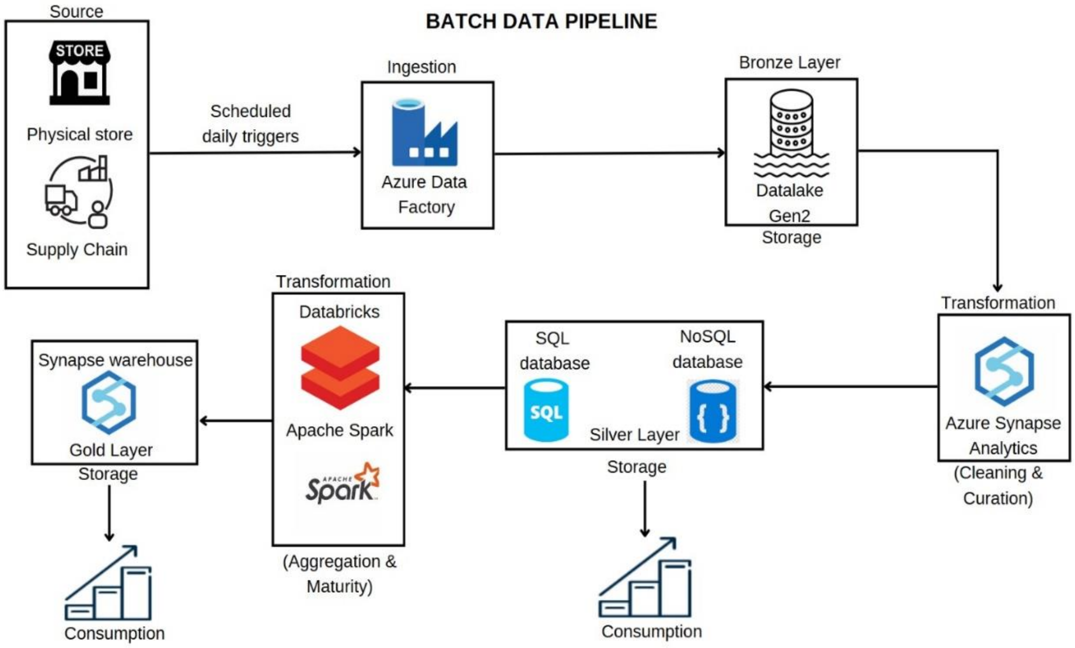
Finally, the data reaches the consumption layer, where it is used by business users for analysis and decision-making. Depending on the needs, data can be accessed from either the Silver or Gold layers.

This layer connects directly to visualization tools like Power BI, allowing users to create reports, dashboards, and insights that drive business decisions.



FOLLOWING ARE DIFFERENT PIPELINE STRAGIES FOR BOTH STREAMING DATA PIPELINE AND BATCH DATA PIPELINE





Monitoring and Failure Strategies

The following complete tracking and failure method is in place to guarantee the pipeline's reliability:

Centralized Logging: By collecting all logs in one place, all information about the system's condition is visible.
Incident Management: Clear communication protocols and ways to escalate are only two of the many systems in place for dealing with problems.
Failure Recovery: To manage failures and reduce downtime, the system is built with automatic retries, checkpointing, and failover techniques.

Conclusion and Prospects for the Future

The future is a priority in the design of this cloud-based data engineering solution. Businesses' requirements for data will change in parallel with the retail industry's continuous development. Through the utilization of Azure's cloud services, this architecture offers an effective, scalable, and adaptable solution that can expand along with the company.

This architecture is positioned to take advantage of emerging technologies and tools as they become available, keeping it at its best and ready to take on new problems. The practice of constant improvement guarantees that the system is always tuned for maximum efficiency.