Electronic Store

Azure Cloud Architecture & Data Engineering Pipelines

~Author~

Kunjal Simzia

in : KunjalSimzia

Table of Contents

1. Introduction

The cloud architecture for an electronic store, considering the business of Best Buy, represents a modern approach to building and managing online retail platforms by using Azure cloud services. Unlike traditional on-premises systems, cloud architecture allows e-commerce businesses to utilize scalable, flexible, and cost-effective resources provided by Microsoft Azure.

2. Mission Statement and Objectives

The mission is to leverage Azure cloud services and build a scalable and robust system which ensures system security and availability.

There are some objectives along with the mission which need to be addressed.

- **Efficient storage:** The efficient storage is to balance the need for quick data retrieval with the costs of storage infrastructure, ensuring that the system operates smoothly and cost-effectively as data volumes grow.

- **High availability:** The system needs to be available and accessible for a high percentage of time, often 99.99% or more, minimizing downtime.

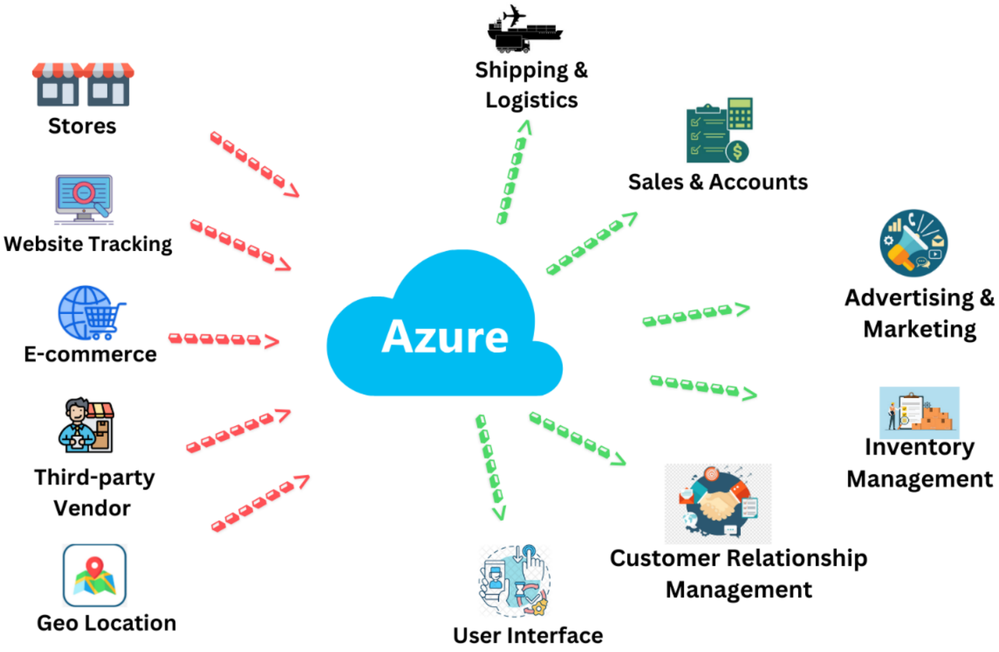
- **Ensure data security:** Data security is implementing measures and protocols to protect data from unauthorized access, breaches, theft, or corruption.

- **Improve sales:** With the help of analytics, the sales team can make different strategies to improve the sales of the company and increase the revenue.

- **Improve operational efficiency:** Enhancing the effectiveness and productivity of an organization's processes and resources to achieve better results with less effort, time, and cost.

3. Architecture Vision

Cloud adoption allows for dynamic resource allocation, scaling resources to match demand. E-commerce transaction data, website analytics, geolocation, and currency information will be gathered, processed, refined, and prepared for end-users to access and utilize. The figure below illustrates the vision for this architecture, depicting the data sources and end users involved.



a. Data Source Layer

The **Data Source Layer** represents the foundational level where raw data is collected from various origins. This layer is crucial as it feeds data into the subsequent layers of the architecture for processing, storage, and analysis. This layer consists of multiple data source:

- **Onsite Store** – An onsite store gathers data from in-store transactions, customer behavior, and inventory management. This information helps optimize stock levels, improve customer experiences, and enhance operational efficiency in the physical retail environment.
- **Website tracking** – It tracks user behavior on the website, including the pages they visit, and the time spent on each page. This data is crucial for developing marketing strategies tailored to user activity and references.
- **Online Store** – An online store collects diverse data from transactions, user interactions, and customer feedback across various platforms.
- **Third-party vendor** – Third-party vendor data involves collecting information from external suppliers or partners, including product availability, pricing, and shipping details. This data is crucial for managing supply chains, ensuring accurate inventory levels, and maintaining competitive pricing strategies across the platform.
- **Geo location** – It offers insights into user locations and regional preferences, enabling businesses to effectively target ads and promotions.

b. Data Consumption Layer

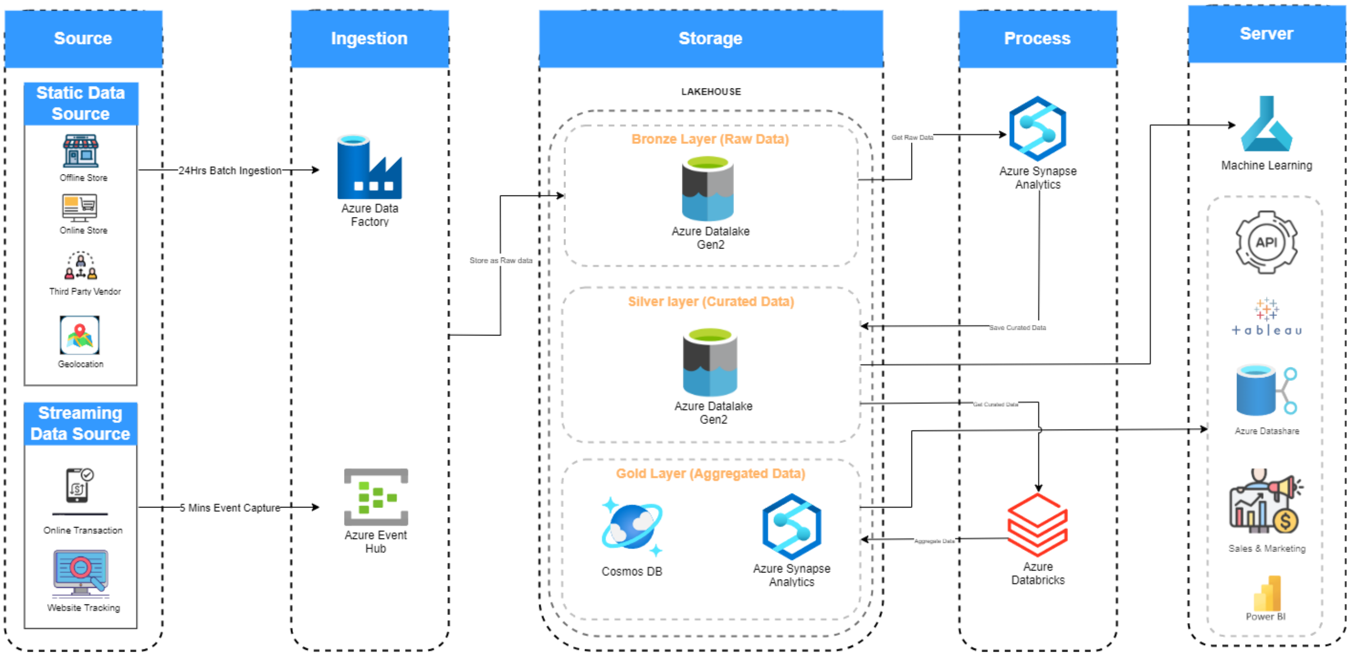
The **Data Consumption Layer** is the layer responsible for delivering processed and analyzed data to end users or applications for decision-making and operational purposes.

- **Shipping and logistics** – To optimize delivery routes, track shipments in real-time, manage inventory across multiple locations, and ensure timely delivery to customers.
- **Sales and accounting** – To monitor financial performance, track revenue and expenses, manage invoicing, and analyze sales trends.
- **Advertising and marketing** – To assess the effectiveness of campaigns, understand customer engagement, and optimize promotional strategies.
- **Inventory management** – To monitor stock levels, track product movement, and analyze supply chain efficiency.
- **Customer relationship management** – To enhance interactions and relationships with customers. This involves analyzing customer behavior, preferences, and feedback to tailor communications, improve service quality, and drive customer satisfaction.
- **On-device experience** – To enhance user interactions directly on their devices. This involves analyzing user behavior, app performance, and interaction patterns to optimize the user interface

4. Cloud Architecture

The architecture below illustrates the data flow within the cloud, from the source to the consume stage.

Electronic Store: Azure Cloud Architecture



a. Source Layer

The source layer in cloud architecture refers to the initial stage where data is collected from various sources before it is processed and stored in the cloud. The source layer has static data source and streaming data source.

Static data sources refer to data that is relatively stable and does not change frequently. In this system static data source consists of store data, online store data, third-party vendor and geolocation data.

Store Data Online Store Data Third Party Vendor Geo Location

Streaming data sources are responsible for providing continuous and real-time data to the system. In this system, streaming data sources are website tracking and E-commerce transactions.

Website Tracking E-commerce Transaction

b. Ingestion Layer

The ingestion layer in cloud architecture is responsible for the initial processing and transfer of data from various sources into the cloud environment. This layer focuses on efficiently bringing data into the cloud system and preparing it for further processing. Data ingestion in this setup utilizes two separate methods and services:

- **Azure Data Factory** is used for batch processing, handling the ingestion of static data source daily. This means it collects and processes large volumes of geolocation data every 24 hours.
- **Azure Event Hubs** is responsible for real-time data streaming, managing the continuous flow of streaming data sources. It handles the ingestion of data as it is generated, allowing for immediate processing and analysis.

This approach ensures efficient handling of both batch and real-time data.

Azure Data Factory Events Hub

c. Storage Layer

Further is the storage layer which is responsible for managing and storing data efficiently and securely. There are some basic layers in storage.

- **Bronze layer:** This layer consists of Raw data which is stored in Azure Data Lake Storage Gen 2. All the data are stored here in their original format without any transformation. This layer acts as a landing zone for all incoming data. Only reading is allowed at this stage

- **Silver layer:** In this layer, data that has been curated and processed is stored, and business rules are applied. Data from the Bronze layer undergoes validation and cleaning to ensure its quality and accuracy. This step includes verifying missing values, correcting data formats, and eliminating duplicates or anomalies.

- **Gold layer:** The curated data in the silver layer provides the basis for creating detailed insights in the gold layer. This layer includes data marts, which are targeted data repositories designed for business units or teams. For example, separate data marts may be established for marketing, finance, and operations, each holding aggregated datasets tailored to their specific needs and goals.

The combination of all these layers is called Lakehouse.

d. Process Layer

The process layer in cloud architecture handles the tasks of transforming, cleansing, and enhancing data. It ensures that the data is properly prepared for storage or analysis by executing various processing operations.

In this system, there are two main processing engines, Azure Databricks and Azure Synapse Analytics. Azure Databricks uses Apache Spark for big data processing and machine learning, while Azure Synapse Analytics integrates data warehousing and big data analytics, providing a unified platform for data management and analysis.

Azure Synapse Analytics   Data Bricks

e. Server Layer

The server layer in cloud architecture is the component responsible for delivering the processing power needed to run applications and manage workloads. All types of data are utilized for multiple purposes in the server layer. The server layer basically consists of PowerBI, Azure Data share, Tableau, Machine learning, sharepoint etc.

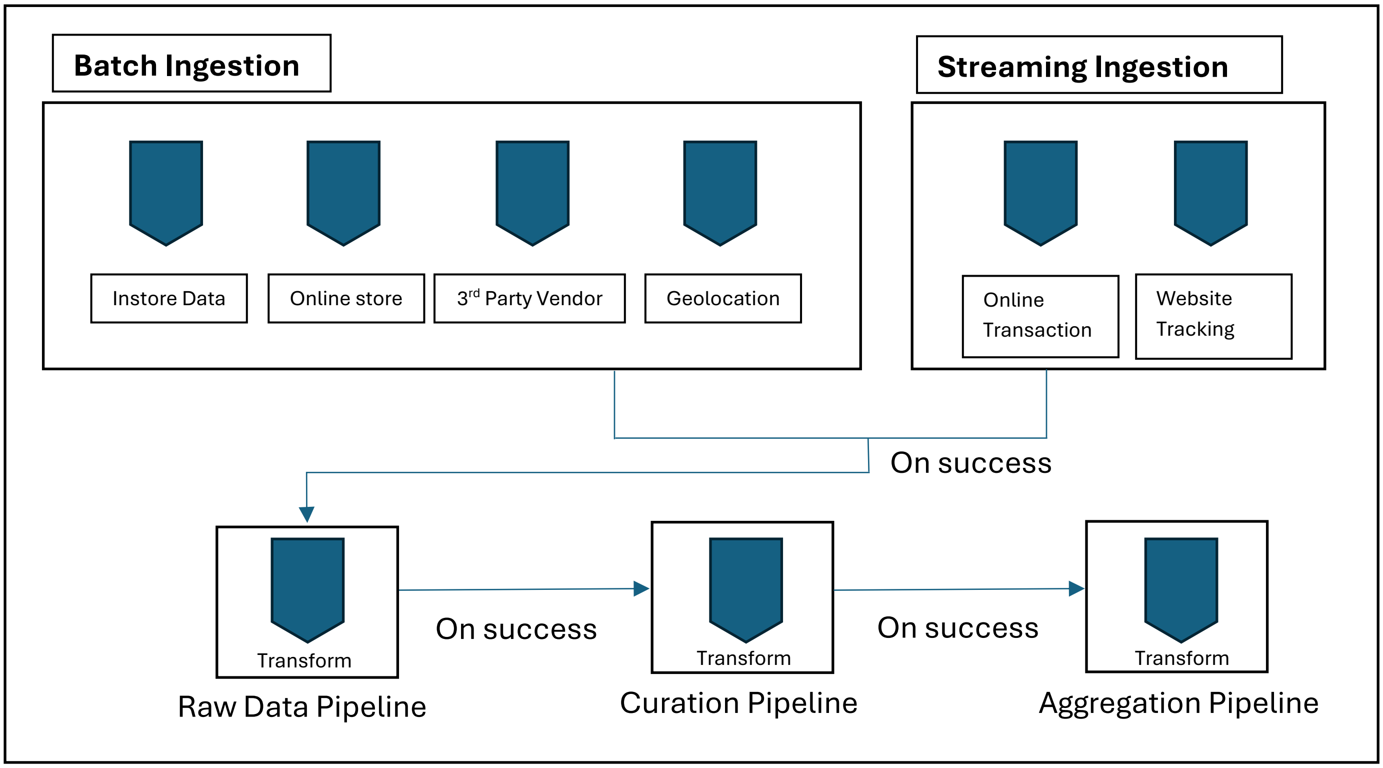
5. Pipeline Strategy

- Initially, the static data from the source layer will be ingested in batch of 24hrs using Azure data factory and the streaming data will be ingested using Azure event hub every 5 mins.
- As there is no urgent need for analytics of streaming data it is stored in database rather than using Azure Streaming Analytics as it is a costly service. Else it will be directly sent to Azure Streaming Analytics for processing and sent to PowerBI and Tableau for analytics.
- Both the data will be inserted into the first layer, bronze layer in the Azure data lake Gen 2 which can store both structured and unstructured data. Using

Datalake Gen 2 both types of data will be managed easily so there is no need for an SQL database.

- After storing it will be sent to Azure Synapse Analytics to remove the null values and data cleaning. The cleaned data will be stored in a silver layer in Azure Datalake Gen 2.
- Further, the curated data will be sent for some more formation and aggregation in the Azure Databricks. Databricks is used for handling big data, it is a bit costly service used for a big amount of data.
- The fine and aggregated data is stored in Cosmos DB for unstructured data and Azure Synapse Analytics for structured data which is named as gold layer.
- Now the silver layer data will be sent to the Data science team to do some machine learning and do some predictions which can help the marketing and sales team to make some improved decisions.
- The data in gold layer will be kept for Azure Data share for data sharing and share using API. Also, the gold layer data will be used for analytics and get some meaningful insights using PowerBI and Tableau.

Master Pipeline



6. Pipeline Failure Strategy

A pipeline failure strategy outlines the steps and actions taken when a data pipeline encounters issues or fails. Here's a general approach:

a) **Detection and Alerts:**

- **Monitoring**: Implement continuous monitoring of the pipeline using tools that can identify any issues or failures in real-time.
- **Alerting**: Set up notifications to immediately inform relevant teams or stakeholders when a problem is detected.

b) **Failure Identification:**

- **Error Logging**: Maintain detailed logs of errors to help diagnose the cause of any failure.
- **Root Cause Analysis**: Review the logs and pipeline components to accurately identify the source of the problem.

c) **Automatic Retry:**

- **Retry Mechanism**: Establish an automatic retry system for temporary or intermittent errors, with a limit on retries to prevent endless loops.

d) **Fallback Procedures:**

- **Failover Mechanism**: Utilize alternative routes or backup systems to keep the process running if the main pipeline fails.
- **Graceful Degradation**: If full functionality cannot be sustained, ensure that essential services continue to operate in a reduced capacity.

This strategy ensures that pipeline failures are managed effectively, minimizing downtime and data loss while continuously improving the system's resilience.

7. Conclusion

Hence, all the objectives are achieved after developing the Azure Cloud Architecture. The Azure cloud setup for an electronic store is designed to be strong and adaptable, meeting the changing needs of online shopping. It ensures data is safe, processed quickly, and easily accessible for analysis. This setup helps the store run smoothly, grow as needed, and improve customer satisfaction, keeping the business competitive in the online market.