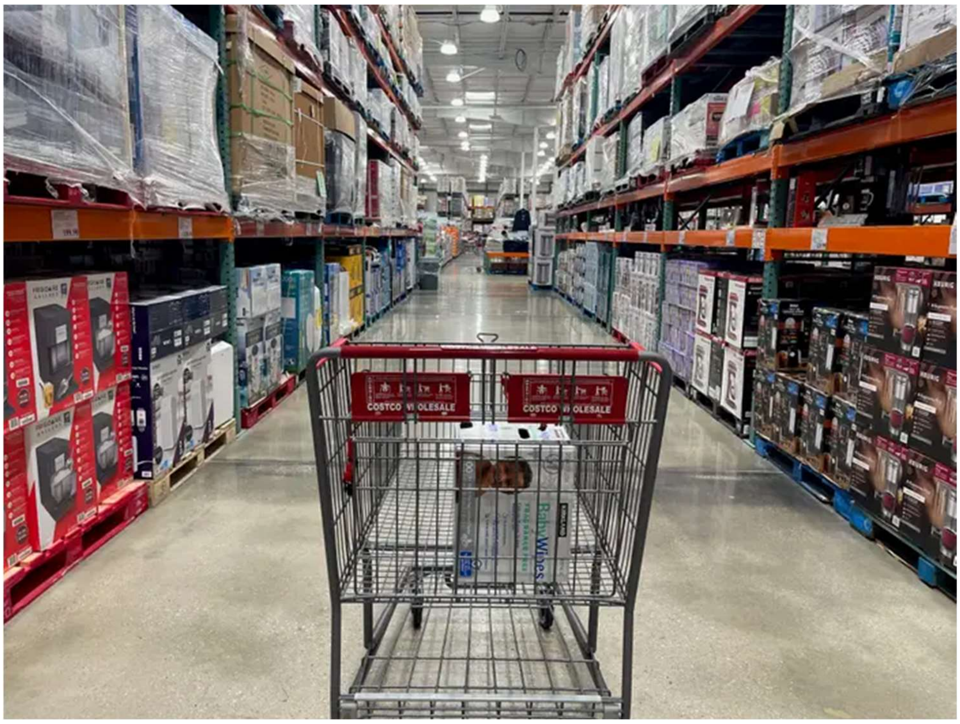


Wholesale World Cloud Architecture

OHM SAI BHARGAV ALURI

https://github.com/OHMALURI/DATASCIENCE

LinkedIn

Contents:

Abstract:

Wholesale World has adopted a cloud-based Lakehouse architecture on Microsoft Azure to effectively manage and analyse different kind of data sources, including membership data, inventory systems, and sales transactions. Using Azure Synapse Analytics, Azure Data Lake Storage Gen2, and Azure Databricks, the system employs a structured Bronze-Silver-Gold pipeline for seamless data ingestion, transformation, and reporting. With strong failure recovery strategies and built in scalability, this architecture supports Wholesale World's objectives of optimizing supply chains, reducing operational costs, and improving customer satisfaction. This creative solution empowers data-driven decision-making, driving sustainable growth and reinforcing the company's commitment to operational efficiency.

Keywords: *Cloud-based, Lakehouse architecture, Microsoft Azure, Azure Synapse Analytics, Azure Data Lake Storage Gen2, Azure Databricks, Bronze-Silver-Gold pipeline, Data ingestion, Supply chain optimization, Data-driven decision-making.*

Introduction:

Wholesale World, A global membership-based retailer and whole sale seller, provides high-quality products at competitive prices through its vast supply chain and bulk sales strategy . Focused on value, innovation, and sustainable growth, the company adopted a cloud-based Lakehouse architecture on Microsoft Azure. This system integrates Azure Synapse Analytics, Azure Data Lake Storage Gen2, and Azure Databricks, using a Bronze-Silver-Gold pipeline (lake house) to transform raw data into actionable insights.

This scalable, reliable architecture optimizes supply chain, helps in reducing costs, and enhances customer experiences, provides reports, supporting Wholesale World's mission to drive sustainable growth, operational excellence, and reinforce its position as an industry leader.

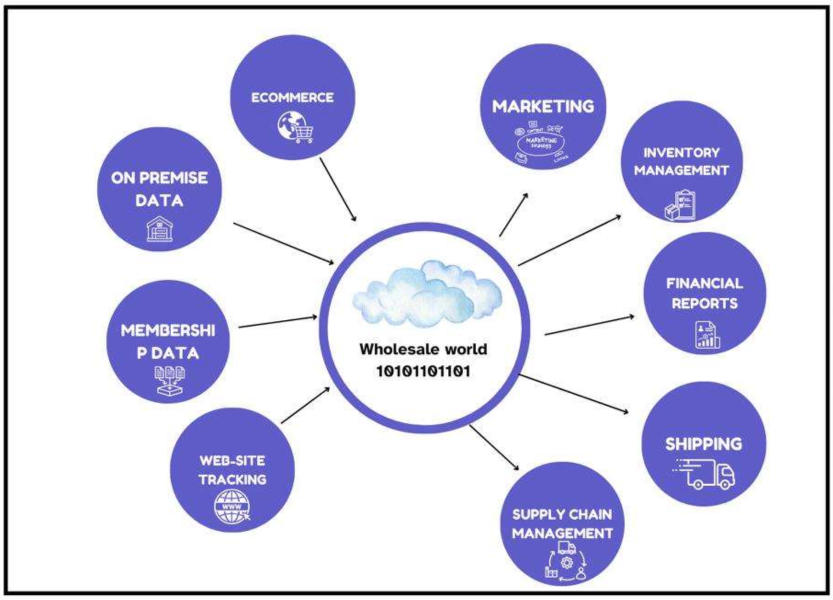
Mission:

To streamline and optimize wholesale warehouse data flow processes to ensure real-time, accurate, and actionable insights for improving operational efficiency, enhancing customer satisfaction, and driving sustainable growth.

Objectives:

1. **Optimize Inventory Management:** Leverage real-time data to avoid overstocking or stockouts and ensure seamless supply chain operations.
2. **Enhance Member Experience:** Provide personalized recommendations and improve customer service using predictive analytics.
3. **Support Data-Driven Decisions:** Empower leadership with detailed insights on sales trends, operational efficiency, and financial health.
4. **Cost Efficiency:** Minimize cloud resource costs through reusable datasets and scalable infrastructure.
5. **Scalable Infrastructure:** Adopt robust cloud-based architecture to handle increasing data volumes and complexity.

Vision diagram:



This vision diagram of the architecture depicts the flow of data from different sources to imp business functions. data sources include e-commerce data, on premises data, membership data, and website tracking data. After aggregating and processing this data, it supports critical business operations. The processed data is then helpful for key sinks, including marketing, inventory management, financial reporting, shipping, and supply chain management, helping in decision-making and improving operational efficiency.

Data Sources:

1.On-Premises data:

- **Type of Data:** Transactional, Numeric, and Categorical
- **Batch Processing:** Daily or periodic aggregation for reporting.
- **Uses:** Inventory management**,** Sales performance tracking, Personalized marketing and promotions.

2. E-Commerce Data:

- **Type of Data:** Transactional, Numeric, and Categorical
- **Example:** Customer Information, Product Details, Order Data
- **Dataflow:** streaming data
- **Uses:** Inventory and supply chain management. Sales performance tracking and revenue forecasting. Customer behaviour analysis and improving user experience.

3. Membership Details:

- **Type of Data:** Categorical, Numeric, and Date/Time
- **Dataflow: Batch Updates:** Member details are often updated periodically for billing or reporting.
- Data sources Stores customer profiles, membership levels, renewal history, and contact information.

4. Website Tracking:

- **Type of Data:** Numeric, Categorical, Event-based
- **Dataflow:** Batch processing
  - Tracks user behaviour on the Wholesale world website (clickstream data).
  - Includes page visits, navigation paths, and session durations.
  - Semi-structured data useful for understanding user engagement.
- Uses: User Behaviour Analysis.

Data Sinks:

1.Financial:

- **Integrates Feedback and Transaction Data:** Analyses revenue, costs, and profit margins, offering a comprehensive view of business performance. Provides insights for budgeting, financial reporting, and profitability analysis.

Delivery Method: power BI

1.Interactive Reports: Power BI delivers interactive, real-time financial reports and dashboards.

2.Visual Analytics: Offers intuitive visualizations of key performance indicators and financial insights for quick, data-driven decision-making.

2.Inventory Management:

• Uses on-premises data and membership details for stock forecasting and shelf replenishment.

• Helps reduce overstocking or stockouts through precise inventory control.

Delivery Method: Power BI

1. **Interactive Inventory Reports:** Power BI generates interactive, real-time reports on inventory levels, providing visibility into stock status across locations.

2. **Visual Insights:** Power BI visualizations help track trends, identify low-stock items, and optimize reorder processes.

3. Supply Chain Operation:

• Utilizes sales and inventory data to optimize replenishment and distribution.

• Enables tracking of stock levels and demand planning.

Delivery Method: reports

1.**Comprehensive Supply Chain Reports:** Generates detailed reports on stock levels, demand forecasts, and replenishment schedules, providing insights into operational efficiency.

2.**Data-Driven Insights:** Helps identify bottlenecks, track performance, and optimize distribution strategies across the supply chain.

4.Marketing:

- Uses sales, e-commerce, and membership data for targeted promotions.
- Enables personalized marketing strategies based on customer behavior, preferences, and purchase history to improve campaign effectiveness.

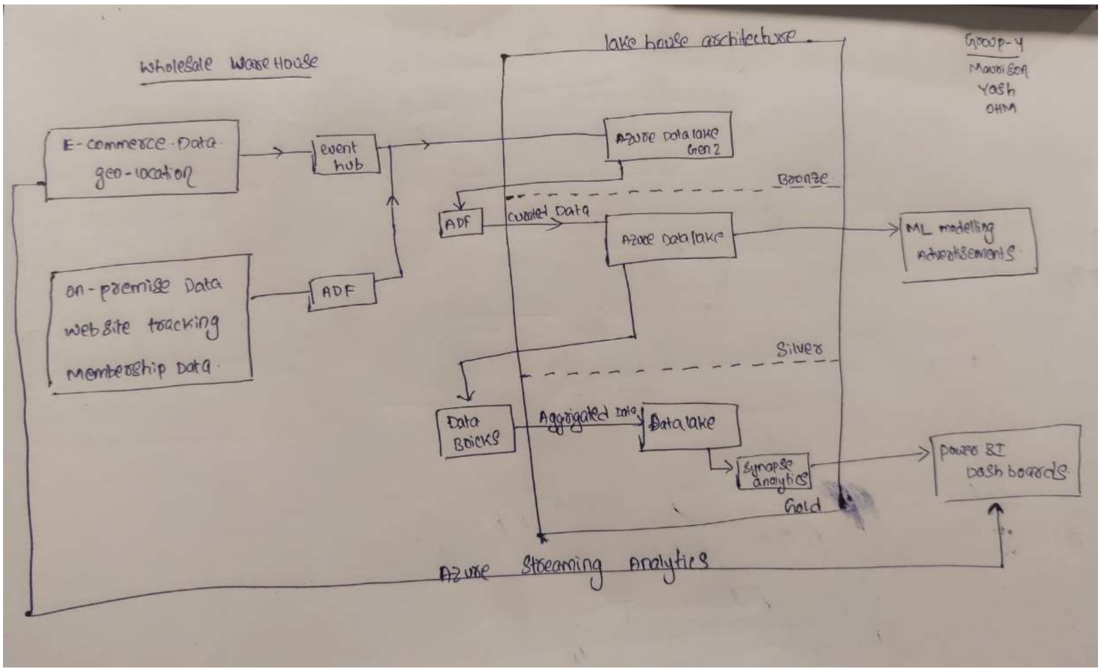
Delivery Method: Digital Channels (Email, Social Media, Push Notifications, SMS)

5.Shipping:

- Uses e-commerce details for shipping information outsourced to third party contractor

Delivery Method: live dashboards

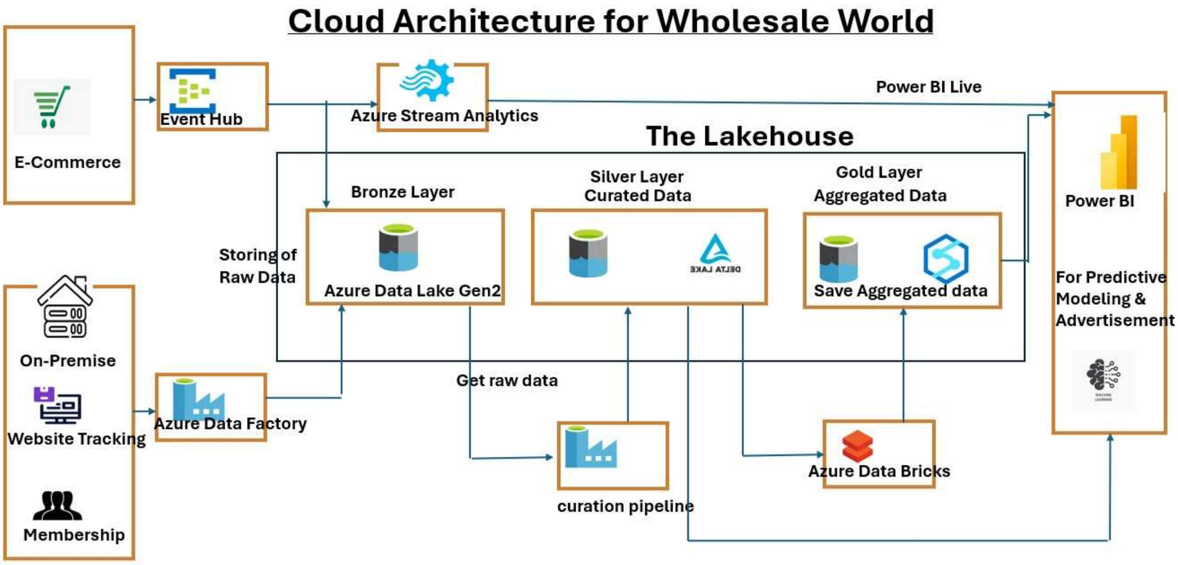
Initial phase of cloud architecture:



In the initial phase of design of the cloud architecture, I went with Lakehouse architecture as it provides the ideal framework for layer-based processing of retail data. The data sources were initially divided into two groups: streaming data and batch data. The streaming data was processed through Event Hub and then aggregated across different layers. On other hand, batch data was ingested through Azure Data Factory (ADF) and processed through the bronze, silver, and gold layers, with the data being aggregated before being sent to the sink.

At first, geolocation data was used as a batch data, but this was later removed from the architecture. Secondly, web site tracking data was initially considered as streaming data but was eventually reclassified as batch data due to cost-effectiveness and better optimized with the volume of data being processed. The final architecture, which reflects these changes, is proposed below.

Final cloud architecture:



USING OF LAKE HOUSE ARCHITECTURE:

Bronze Layer:

In this architecture, the data sources are segregated into two groups: batch data and streaming data. The batch data contains membership data, website tracking data, and on-premise data, while e-commerce data is considered as streaming data due to its real-time nature.

For the ingestion process, **Azure Data Factory (ADF)** is used to effectively handle the batch data and extracting it, and loading it into the system. **Event Hubs**, on the other hand, is used to ingest the streaming data, providing a scalable and reliable way for ingesting real-time data from e-commerce transactions.

Once the data is ingested, both batch and streaming data are stored in the bronze layer of this architecture. This layer serves as the raw data storage, where all data, in its raw form, is stored in **Azure Data Lake Storage Gen2**. Gen2 is preferred for storage in the bronze layer because of its high scalability, cost-efficiency, and integration with big data processing tools. It provides optimized option for handling large volumes of raw, unstructured data, and its hierarchical namespace allows for better data management, making it an ideal solution for further processing and analysis.

Azure Stream Analytics is used to aggregate streaming data in the real time, dealing with e-commerce transactions and customer interactions. The aggregated data is then directly pushed to **Power BI** for creating live dashboards and interactive reports. This facilitates real-time insights into sales, inventory, and customer activity for immediate decision-making.

Silver Layer:

The raw data which is stored in the bronze layer in **Azure Data Lake Storage Gen2** is moved to the **silver layer** with the help of **Azure Data Factory (ADF)**. ADF is used to transform the data by using curation techniques such as filling the missing values, correcting inconsistencies, and cleansing the data. The transformed and curated data is then ingested in the silver layer.

In the Silver layer, **Azure Data Lake Storage Gen2** is again used for storage, but this time in addition with **Delta Lake**. Delta Lake offers ACID transaction support, ensuring data consistency and reliability. It also facilitates the scalable data processing, allowing for effective updates and deletions of data in the silver layer.

The curated data in the silver layer is now available for advanced analytics and can be used for machine learning modelling, customer behaviour analysis, and targeted advertising. **Azure Communication Services (ACS)** can be leveraged to enable personalized customer communication, such as targeted email campaigns, SMS alerts, or notifications based on customer preferences and behaviours identified in the curated data. This makes the Silver layer crucial for driving business strategies and enhancing customer engagement through real-time communication.

Gold layer:

The **Gold layer** depicts the final stage in the data processing pipeline where curated and transformed data from the **Silver layer** is now aggregated and optimized for reporting and analytics. This layer is important for providing business-critical insights that support decision making across various departments, such as finance, inventory management, and supply chain.

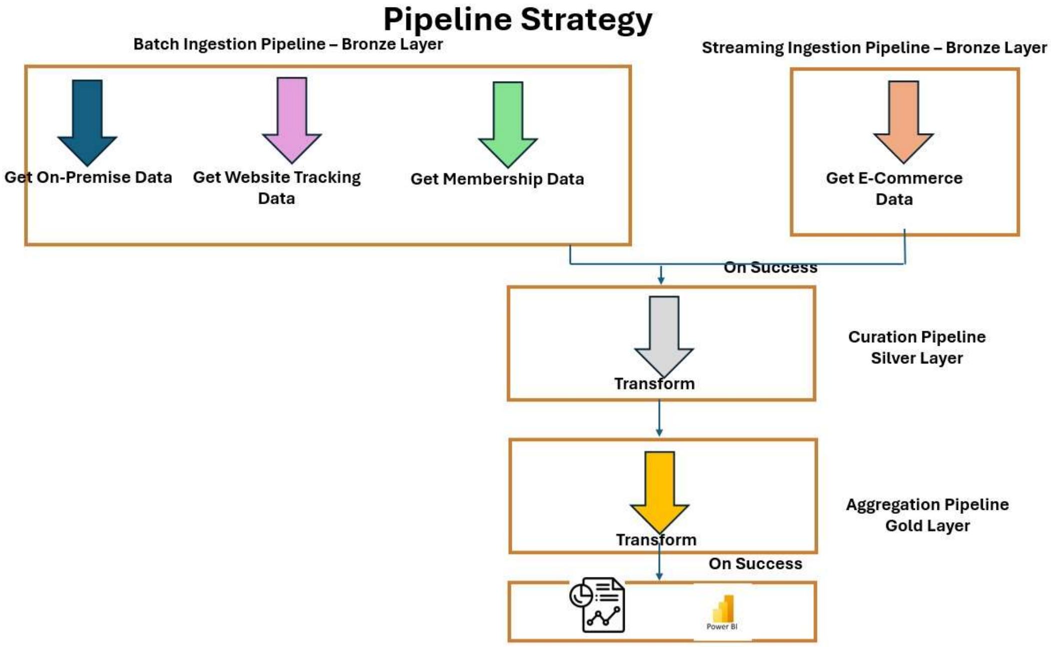
This process begins with **Azure Databricks**, which is used to perform complex aggregations on the curated data. Databricks offers efficient data processing at scale, applying various aggregation techniques to summarize key metrics like **financial data, inventory levels, stock counts, and other business KPIs**. These aggregated results are important for generating reports and performing high level analysis that directly depicts business strategies.

Once the data is aggregated, it is stored again in **Azure Data Lake Storage Gen2** in the Gold layer, maintaining high availability and optimized storage for big volumes of structured and semi-structured data. This will allow the organization to store vast amounts of business-critical data securely, with scalable access for future analysis.

To connect this aggregated data to actionable insights, **Azure Synapse Analytics** is used. Synapse used as a bridge between the gold layer and the **sink layer**, enabling seamless integration with external tool like **Power BI**. Synapse provides powerful analytics capabilities, allowing for data to be processed, queried, and prepared for visualization in Power BI.

With the **Power BI**, the data from the Gold layer is utilized to create interactive, real-time visualizations, live dashboards, and detailed reports. These reports provide decision-makers with a detailed view of the business, offering insights into financials, inventory health, customer trends, and other critical metrics. Live dashboards make sure that stakeholders can access up-to-date data instantly, improving response and enabling more agile business operations.

Pipeline Design:


Pipeline Strategy
Batch Ingestion Pipeline – Bronze Layer
Streaming Ingestion Pipeline – Bronze Layer
Get On-Premise Data
Get Website Tracking Data
Get Membership Data
Get E-Commerce Data
On Success
Transform
Curation Pipeline Silver Layer
Transform
Aggregation Pipeline Gold Layer
On Success
Power BI

The pipeline strategy makes sure of successful execution at each stage:

1. **Data Ingestion:** Data from diff sources like e-commerce and membership is ingested via batch and streaming pipelines and stored in the **Bronze Layer**. Successful ingestion ensures raw data is available for processing.

2. **Data Processing and Transformation:** The **Curation Pipeline** cleans and transforms data into the **Silver Layer,** making it ready for analysis after successfully running. The **Aggregation Pipeline** in the **Gold Layer** aggregates all the data for reporting and analytics, ensuring accurate insights.

3. **Data Output:** The Gold Layer data is used for **Power BI dashboards** and **analytics**, and **predictive modelling**. Successful execution ensures the data is actionable and ready for decision-making.

Each step is monitored, with any issues triggering alerts for a quick resolution, ensuring the pipeline consistently giving high-quality, actionable insights.

Pipeline Failure Strategy:

- **Data Ingestion**: Failed data ingestion (both batch and streaming) triggers automatic retries. Invalid data is rerouted to **Dead Letter Queues** for manual reviewing. Alerts are sent to notify the team members of any disruptions.
- **Bronze Layer (Raw Data):** Data integrity checks ensure completeness, and any issues result in reprocessing from the source. Regular backups allow restoration of raw data if needed.
- **Silver Layer (Curation):** Transformation failures are handled with **validation rules**, and if errors still exist, the system go back to the raw data in the Bronze Layer for reprocessing.
- **Gold Layer (Aggregation & Reporting):** Aggregation failures shows the partial results on dashboards while issues are fixed. **Detailed error logs and alerts** help them addressing problems quickly.
- **General Recovery:** Each stage is supported using retry mechanisms, dead letter queues, and alerting systems which ensures high availability and timely resolution of issues, persisting data quality and insights for decision-making.

Conclusion:

In conclusion, the cloud architecture for the Wholesale **World** gives a comprehensive, scalable, and cost-efficient solution that utilizes the power of Azure services to transform the data into actionable insights. By adopting a **Lakehouse architecture** with 3 layers (Bronze, Silver, and Gold), Wholesale World can effectively manage and process both batch and streaming data from multiple sources. The use of **Azure Data Lake Storage Gen2**, **Azure Data Factory**, **Event Hubs**, **Databricks**, and **Power BI** gives seamless data ingestion, transformation, and visualization, making key business processes like inventory management, marketing, and supply chain optimization.

The well-defined **pipeline failure strategy** ensures the data reliability at each stage, maintaining data integrity and minimizing downtime. By integrating the data-driven solutions, Wholesale World will improve the customer experience, supports the real-time decision-making, and improves operational efficiency. This architecture not only provide smarter, data-backed decisions but also helps in company's long-term goal of sustainable growth and agility in an ever-evolving market.