CASE STUDY:
SKY MART EXPRESS

BY
NITIN

Southern Alberta Institute of Technology Calgary, Canada

https://www.linkedin.com/in/nitin-b57867168/
https://github.com/nitin10-sys/Project1

INDEX

1. **INTRODUCTION**

   Sky Mart Express relies on a robust logistics framework, including top carriers and advanced planning strategies, to ensure reliable, flexible, and timely deliveries. The company employs a global network and adaptive technology to effectively navigate disruptions like weather or traffic. This adaptability ensures goods are routed efficiently while minimizing costs and logistical complexity. This introduction underscores the importance of using technology to optimize the supply chain.

2. **MISSION**

   Designing cloud architecture for Sky Mart that's reliable, fast, secure, and able to grow easily supports its online store and data analysis needs.

3. **OBJECTIVES**

   • Speed up order processing for smooth and accurate transactions.

   • Track inventory, orders, and customer interactions in real time.

   • Give useful insights into market trends to help make better decisions.

   • Make the best use of resources like staff, inventory, and infrastructure.

   • Provide personalized customer experience with recommendations and targeted campaigns.

   • Share clear and detailed reports with stakeholders for better decisions.

4. DESIGN AND DISCOVERY PHASE
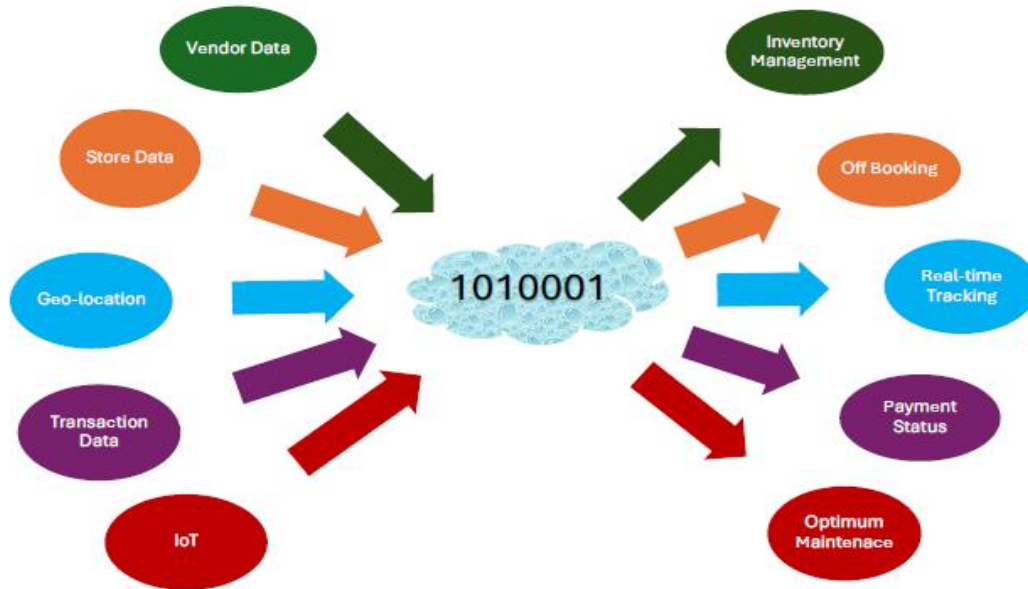
5. DATA SOURCES OVERVIEW

1. Vendor Data

• **Type**: Structured

• **Flow**: Batch processing, as updates occur periodically.

• **Source**: Supplier portals, ERP systems, APIs.

2. Transaction Data

• **Type**: Structured

• **Flow**: Typically, batch but can also support real-time for payment success/failure tracking.

• **Sources**: E-commerce platforms, payment gateways, POS systems, CRM systems.

3. IoT (Internet of Things)

• **Type**: Unstructured or semi-structured

• **Flow**: Primarily streaming; may use batch for historical analysis.

• **Sources**: Smart sensors, GPS systems, IoT-enabled devices

4. Store Data

• **Type**: Structured and unstructured

• **Flow**: Batch, as updates are typically daily.

• **Source**: POS systems, store management software.

5. Geo-Location Data

• **Type**: Unstructured

• **Flow**: Streaming, as locations are updated live.

• **Source**: Mobile app GPS, third-party location APIs.

6. KEY DELIVERABLES

1. (a) Inventory Management

- Low-stock alerts and predictive restocking.
- Inventory turnover rates and real-time stock availability.

(b) Delivery Method

- Power BI Dashboards
- Automated Reports
- API Integration

2. (a) Payment Status

- Detailed sales reports.
- Payment success/failure analysis.
- Refund and return insights.
- Fraud detection reports for unusual transactions.

(b) Delivery Method

- Power BI Dashboards
- Alerts
- Automated Reports

3. (a) Optimum Maintenance

- Monitor live delivery and inventory movement.
- Track machine performance to predict failures.

(b) Delivery Method

- Power BI Dashboards
- Alerts

4. (a) Off-Booking

- Analysis of unfulfilled orders and stockouts.
- Root cause reports for booking issues.

(b) Delivery Method

- Power BI Dashboards
- Alerts
- Vendor Reports

5. (a) Real-Time Tracking

- Live tracking of deliveries and inventory updates.
- Geo-location data for targeted marketing or delivery route optimization.

(b) Delivery Method

- Power BI Dashboards
- Notifications

Data Processing

- **Data Flow:** The data is ingested in either batch or real-time modes, depending on its nature.
  - **Batch Processing**: Processes periodic updates (e.g., daily sales, and inventory reports).
- **Streaming**: Supports live updates like geo-location tracking or IoT sensor data.

    The central processing system (cloud) harmonizes these diverse data formats (structured, semi-structured, and unstructured).

Purpose of the Design

- **Reliability:** Handles disruptions like traffic or weather delays with adaptive routing.
- **Scalability:** Easily expands to manage increased data flows.
- **Insights-Driven Decisions:** Provides detailed analytics for operational and strategic improvements.
- **Cost-Efficiency:** Optimizes resource utilization (staff, inventory, delivery routes).

7. PHASE 1
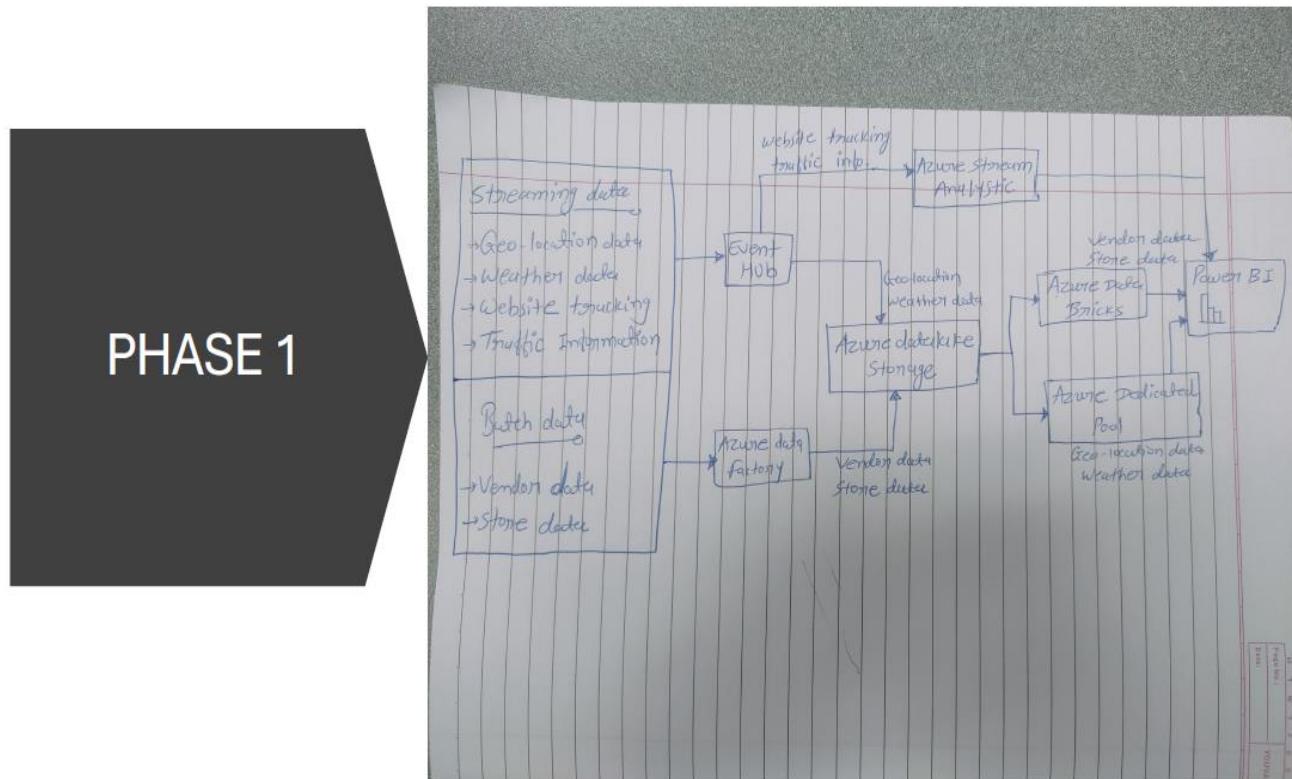
Explanation of Phase 1 Operations (as shown in the diagram)

Phase 1 focuses on setting up the foundation for Sky Mart's data ingestion and processing system. Here's a detailed breakdown of the components, data flows, and their respective functions:

---

1. Data Sources

- **Streaming Data:**
    - **Geo-location Data**: Real-time data from GPS devices, apps, or sensors to track asset or delivery locations.
    - **Weather Data**: Information about weather conditions that might affect logistics and operations.
    - **Website Tracking**: Data about customer interactions, clicks, and navigation patterns from Sky Mart's website.
    - **Traffic Information**: Live traffic updates used for route optimization.
- **Batch Data**:
    - **Vendor Data**: Information on product availability, supplier updates, and stock orders.

    o **Store Data**: Periodic sales reports, inventory levels, and operational metrics from physical stores.

---

2. Event Hub

- **Role**: A central entry point for all streaming data.
- **Function**: Collects data from real-time sources (geo-location, weather, website tracking, traffic) and prepares it for further processing.
- **Why it's Important**: Event Hubs ensure the seamless handling of high-volume real-time data streams, acting as a buffer before analytics processing.

---

3. Azure Streaming Analytics

- **Role**: Processes real-time streaming data from the Event Hub.
- **Function**:
  - Performs filtering, aggregation, and transformations on the live data.
  - For example, it can calculate delivery delays due to traffic or weather conditions.
- **Output**: Insights are passed on to downstream systems or dashboards for immediate action.

---

4. Azure Data Lake Storage

- **Role**: Centralized storage for both streaming and batch data.
- **Function**:
  - Stores raw and processed data for long-term use.
  - Serves as a repository for further data processing or analytics.
- **Why it's Important**: Azure Data Lake provides scalability for handling large volumes of structured and unstructured data.

---

5. Azure Data Factory

- **Role**: Manages batch data ingestion (Vendor and Store data).
- **Function**:
  - Automates data extraction, transformation, and loading (ETL) into the Azure Data Lake.
  - Schedules and orchestrates workflows to ensure data is ingested at regular intervals.

6. Azure Databricks

- **Role**: Advanced analytics platform for big data processing.
- **Function**:
  - Processes data from the Azure Data Lake.
  - Performs data cleaning, transformations, and modeling.
  - For example, it can integrate streaming weather and traffic data with batch vendor data for predictive analytics.
- **Output**: Outputs clean and enriched data to the Azure Dedicated Pool.

7. Azure Dedicated Pool (Synapse Analytics)

- **Role**: Handles advanced analytics and high-performance queries.
- **Function**:
  - Acts as a data warehouse for processed and refined data.
  - Enables fast querying and supports Power BI dashboards.
- **Why it's Important**: Provides structured, query-optimized data for reporting and business intelligence.

8. Power BI

- **Role**: Visualization and reporting tool.
- **Function**:
  - Consumes data from Azure Synapse Analytics.
  - Provides interactive dashboards for stakeholders to monitor KPIs such as sales trends, inventory levels, and delivery performance.
- **Why it's Important**: Enables actionable insights and decision-making.

Data Flow in Phase 1

1. **Streaming Data**: Collected through the Event Hub → Processed by Azure Streaming Analytics → Stored in Azure Data Lake.
2. **Batch Data**: Ingested through Azure Data Factory → Stored in Azure Data Lake.

3. Data from the Azure Data Lake is processed by **Azure Databricks** for cleaning and transformation.
4. The transformed data is loaded into the **Azure Dedicated Pool** for structured querying.
5. **Power BI** consumes data from the Azure Dedicated Pool to generate dashboards and reports.

Purpose of Phase 1

- To establish a robust pipeline for ingesting, processing, and storing both real-time and batch data.
- To provide a foundation for analytics, reporting, and decision-making.
- To ensure scalability and reliability for future operational phases.

8. PHASE 2
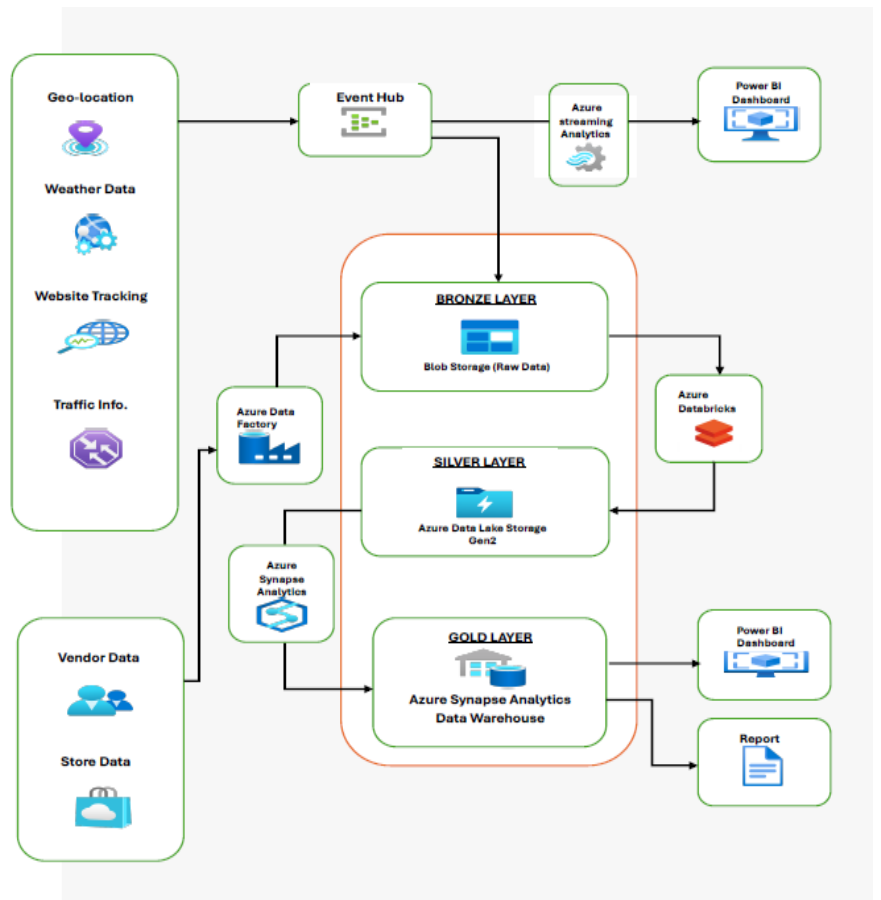
Phase 2 Explanation Based on the Provided Diagram

Overview

Phase 2 builds upon the foundational data ingestion and processing pipeline established in Phase 1. It further integrates structured and unstructured data from diverse sources, enhances analytics capabilities, and introduces layered storage (Bronze, Silver, and Gold) for better data management and analysis.

Components and Workflow

1. **Data Sources**:
   o **Streaming Data**:
      ▪ **Geo-Location Data**: Real-time tracking data sourced from GPS devices or apps to monitor asset and delivery locations.
      ▪ **Weather Data**: Continuous weather updates impacting delivery or logistics.
      ▪ **Website Tracking**: Behavioral data collected from user interactions on the Sky Mart website.
      ▪ **Traffic Information**: Real-time data used for dynamic route optimization.
   o **Batch Data**:
      ▪ **Vendor Data**: Structured updates from suppliers regarding inventory, orders, and availability.
      ▪ **Store Data**: Periodic updates on in-store sales, inventory, and operations.
2. **Event Hub**:
   o Acts as a central ingestion hub for streaming data sources (e.g., Geo-location, Website Tracking, and Traffic).
   o Buffers and prepares real-time data for immediate or near-immediate analysis.
3. **Azure Streaming Analytics**:
   o Processes data ingested by Event Hub.
   o Execute transformations, filtering, and aggregations (e.g., calculating delivery delays from weather and traffic data).
   o Outputs actionable insights directly to dashboards or other analytics systems.
4. **Azure Data Factory**:
   o Facilitates batch processing by automating the extraction, transformation, and loading (ETL) of Vendor and Store Data.
   o Transfers cleaned data into Azure Data Lake for integration with streaming data.
5. **Layered Data Architecture**:
   o **Bronze Layer (Blob Storage)**:
      ▪ Stores raw data directly from the source without any processing.
      ▪ Acts as a repository for long-term historical analysis.

- o **Silver Layer (Azure Data Lake Gen2)**:
    - Houses processed and cleaned data, suitable for advanced analytics.
    - Integrates batch and streaming data to enable comprehensive analysis.
  - o **Gold Layer (Azure Synapse Analytics)**:
    - Contains refined, structured data optimized for querying and reporting.
    - Supports interactive dashboards and advanced business intelligence.
6. **Azure Databricks**:
   - o Performs advanced analytics on data stored in the bronze and silver layers.
   - o Execute transformations, data cleaning, and machine learning models (e.g., demand forecasting, route optimization).
   - o Outputs enriched data to the Gold Layer for reporting.
7. **Azure Synapse Analytics**:
   - o Acts as a data warehouse for the Gold Layer.
   - o Handles complex queries and supports large-scale analytics workloads.
   - o Serves as the primary source for business intelligence and reporting.
8. **Power BI**:
   - o Visualizes processed data from the Gold Layer.
   - o Create dashboards for tracking KPIs like sales, inventory, and logistics performance.
   - o Enables stakeholders to make data-driven decisions through interactive insights.
9. **Reports**:
   - o Automated reports generated for deeper insights into operational performance.
   - o Shared with stakeholders for strategic planning and analysis.


Data Flow in Phase 2

1. **Real-Time Streaming**:
   - o Data flows from sources (Geo-location, Weather, Website Tracking, Traffic) → Event Hub → Azure Streaming Analytics → Azure Data Lake (Bronze Layer).
2. **Batch Processing**:
   - o Data flows from Vendor and Store Data sources → Azure Data Factory → Azure Data Lake (Bronze Layer).
3. **Integration and Transformation**:
   - o Data in the Bronze Layer is processed via Azure Databricks and stored in the Silver Layer.
   - o Enriched data in the Silver Layer undergoes further optimization and is stored in the Gold Layer (Azure Synapse Analytics).
4. **Insights and Reporting**:

    o   Data in the Gold Layer is consumed by Power BI for dashboard visualization and reporting.

Purpose of Phase 2

- **Enhance Data Usability**: Introduces a layered architecture (Bronze, Silver, Gold) for improved data management and analysis.
- **Scalable Insights**: Enables real-time and batch processing for a holistic view of operations.
- **Improved Decision-Making**: Provides actionable insights via Power BI dashboards and reports.
- **Operational Efficiency**: Combines data from diverse sources to optimize routes, predict demand, and enhance customer experience.

9. PHASE 3
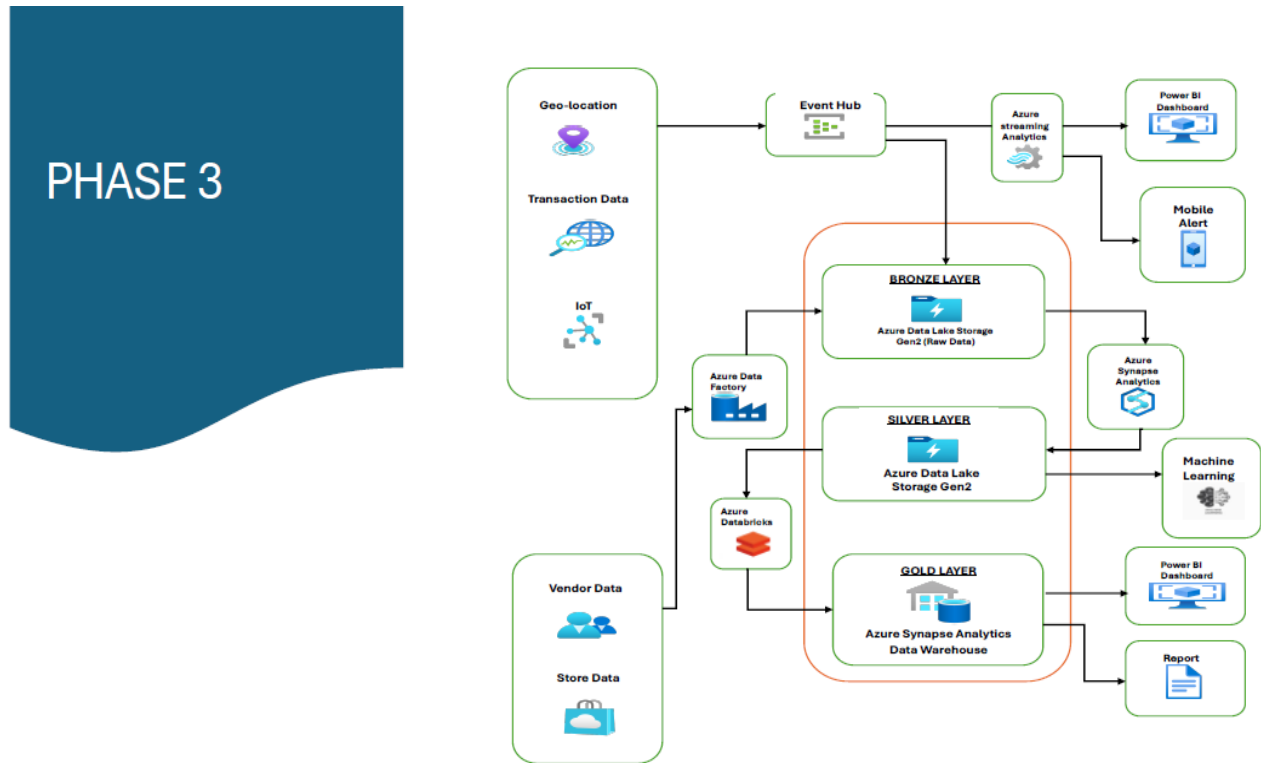
Phase 3 integrates multiple data sources (real-time and batch) into a cohesive data processing pipeline using Azure cloud services. It introduces advanced analytics and machine learning capabilities to transform raw data into actionable insights, enhancing decision-making and business intelligence.

Key Components in Phase 3

1. **Data Sources**
   - **Geo-location Data**: Captures real-time location updates for deliveries or assets.
   - **Transaction Data**: Includes customer transactions like orders and payment details.
   - **IoT Data**: Gathers data from smart devices, sensors, and connected systems.
   - **Vendor Data**: Provides supplier-related updates and stock availability.
   - **Store Data**: Consists of in-store inventory and sales updates.
2. **Event Hub**
   - **Role**: A centralized service for ingesting real-time streaming data (e.g., geo-location, IoT data).
   - **Functionality**: Acts as a buffer to handle high-volume streaming data before further processing.
3. **Azure Streaming Analytics**
   - **Role**: Real-time data processing and transformation.
   - **Functionality**: Performs calculations (e.g., traffic analysis, delivery delays) and forwards insights to dashboards or downstream systems.
4. **Bronze, Silver, and Gold Layers**

- o **Bronze Layer** (Raw Data Storage):
  - **Azure Data Lake Gen2** stores unprocessed, raw data from all sources.
  - Acts as a centralized repository for historical and unprocessed data.
- o **Silver Layer** (Processed Data):
  - Data is cleansed, structured, and integrated using **Azure Databricks**.
  - Makes the data analysis ready while combining real-time and batch data sources.
- o **Gold Layer** (Refined Data):
  - **Azure Synapse Analytics** serves as the data warehouse.
  - Stores refined, query-optimized data for advanced analytics and visualization.

5. **Azure Data Factory**
   - o **Role**: Batch data ingestion and orchestration.
   - o **Functionality**: Automates the movement and transformation of batch data (e.g., vendor and store data) into the Bronze Layer.

6. **Azure Databricks**
   - o **Role**: Big data processing and advanced analytics platform.
   - o **Functionality**:
     - Cleans and integrates data from multiple sources.
     - Enables machine learning workflows for predictive insights (e.g., demand forecasting).

7. **Azure Synapse Analytics**
   - o **Role**: Data warehouse for refined data.
   - o **Functionality**:
     - Supports complex analytical queries.
     - Integrates with Power BI for visualization and reporting.

8. **Power BI Dashboards**
   - o **Role**: Visualization and reporting.
   - o **Functionality**: Presents insights via interactive dashboards for stakeholders to monitor key metrics like sales trends, inventory levels, and delivery performance.

9. **Mobile Alerts**
   - o Provides real-time notifications to users or system administrators regarding critical updates, such as delays or inventory issues.

10. **Machine Learning Integration**
    - o **Role**: Advanced predictive analytics.
    - o **Functionality**:
      - Apply machine learning models for tasks like demand forecasting, anomaly detection, and route optimization.
      - Informs business strategies and improves operational efficiency.

Workflow in Phase 3

1. **Data Ingestion**
   - o Streaming data is routed through the **Event Hub** into the **Bronze Layer** via **Azure Streaming Analytics**.
   - o Batch data is ingested using **Azure Data Factory** and stored in the **Bronze Layer**.

2. **Data Transformation and Processing**
   - The **Bronze Layer** stores raw data, which is processed and cleaned using **Azure Databricks**.
   - Processed data is moved to the **Silver Layer** for integration and further transformations.
   - Final refined data is stored in the **Gold Layer** within **Azure Synapse Analytics**.
3. **Analytics and Reporting**
   - Data in the **Gold Layer** is queried using **Azure Synapse Analytics**.
   - Insights are visualized in **Power BI Dashboards** and shared via mobile alerts or automated reports.
4. **Machine Learning**
   - Integrated models analyze data for trends and predictions, feeding results back into the system for reporting and decision-making.

Purpose and Benefits of Phase 3

- **Real-Time Insights**: Combines batch and streaming data for a complete view of operations.
- **Enhanced Decision-Making**: Machine learning and advanced analytics provide actionable insights.
- **Operational Efficiency**: Automates and optimizes key business processes like inventory management and route planning.
- **Scalability**: Cloud-native architecture supports growing data needs and advanced analytics workloads.
- **Stakeholder Visibility**: Dashboards and reports empower stakeholders with clear, data-driven insights.
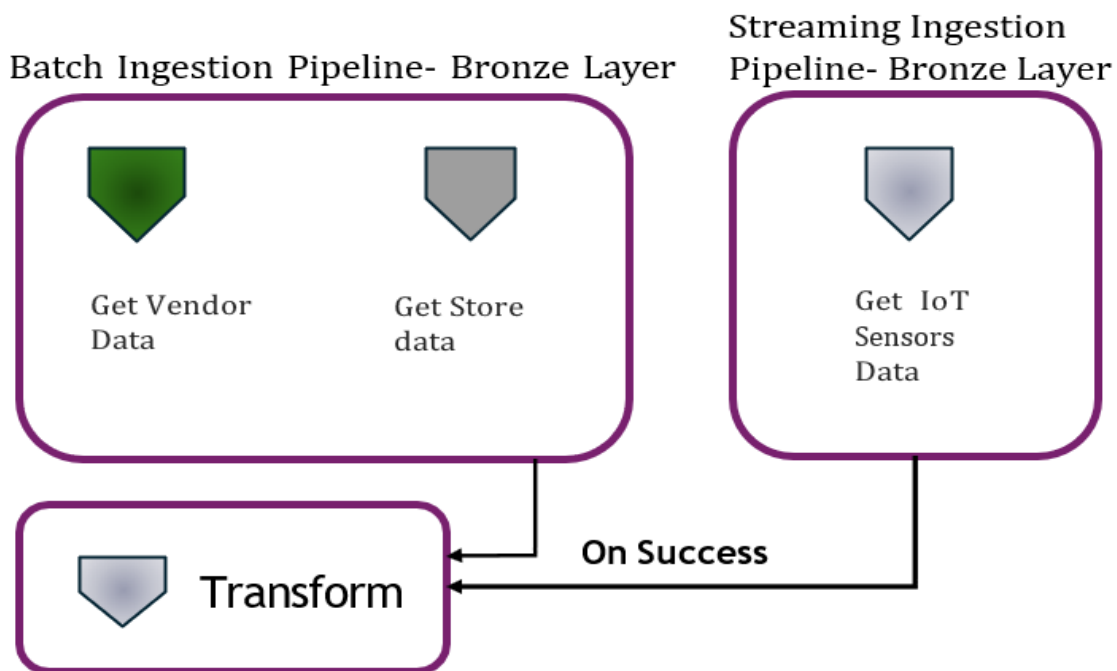
10. PROCESS OF CREATING PIPELINE
Bronze Layer (Ingest Data)

We decided to create two ingestion pipelines:

- Batch ingestion and
- Streaming ingestion.

Each one of these pipelines will be built using a different set of Azure services.

- For batch ingestion, we will use Azure Data Factory, and
- For streaming ingestion, we will use Azure IoT hubs.

Pipeline Failure

A pipeline failure happens when a data pipeline cannot process, transfer, or load data as expected.

Failure attempt: - Max3 time

Attempt time: - After 1 hour

11. CONCLUSION

The Sky Mart Cloud Architecture provides a robust, scalable, and efficient framework for managing and optimizing business operations. By integrating advanced data ingestion, processing, and analytics pipelines across its three phases, Sky Mart ensures a seamless blend of real-time and batch data handling. This cloud-based solution empowers the company to make data-driven decisions, enhance customer experience, and improve operational efficiency. The layered architecture—Bronze, Silver, and Gold— ensures optimal data usability while enabling advanced insights through machine learning and predictive analytics. Moreover, the use of Azure services for storage, analytics, and visualization creates a highly reliable and secure environment, fostering continuous growth and adaptability to dynamic business needs. Through these innovations, Sky Mart Express is positioned to maintain its competitive edge in the logistics and retail industry, offering timely deliveries, personalized customer experiences, and resource optimization. This architecture not only fulfills current objectives but also lays a strong foundation for future technological advancements and scalability.