



SOFA – SO GOOD

CLOUD ARCHITECTURE



Submitted By:
Praveen Chaudhary

<https://www.linkedin.com/in/praveen-chaudhary96/>

INDEX



1. INTRODUCTION
 - 1.1. Overview of Sofa-so good Cloud Architecture
2. Project Purpose and Objectives
 - 2.1. Objective
 - 2.2. Purpose
3. Mission and Goals
 - 3.1. Mission Statement
 - 3.2. Project Goals
4. Design and Discovery phase
 - 4.1. Requirements gathering
 - 4.2. Key Challenges and Opportunities
5. Data Sources Overview
 - 5.1. Data Inputs
 - 5.1.1. Sales Transactional Data
 - 5.1.2. E – commerce Data
 - 5.1.3. Website logs
 - 5.1.4. Warehouse Operational Data
 - 5.2. Data Outputs / Key Deliverables
 - 5.2.1. Sales Performance Dashboard
 - 5.2.2. Inventory Management Reports
 - 5.2.3. Customer Segmentation Reports
 - 5.2.4. Sales Forecast Reports
 - 5.2.5. Web Analytics Dashboard
 - 5.3. List of Data Users
 - 5.3.1. Marketing team
 - 5.3.2. Sales team



- 5.3.3. Operations team
- 5.3.4. Management team
- 5.3.5. Customer service team
- 5.3.6. Web development team
- 6. Cloud Architecture Phases
 - 6.1. (PHASE - 1): Initial Design
 - 6.2. (PHASE – 2): Refined Design
 - 6.3. (PHASE – 3): Final Architecture Design
- 7. Process of creating pipeline
 - 7.1. Bronze layer (Ingested Data)
 - 7.2. Silver layer (Curated Data)
 - 7.3. Gold layer (Aggregated Data)
 - 7.4. Master Pipeline Approach (Event-based)
 - 7.5. Pipeline Failures and Solutions
 - 7.5.1. Failure 1: Data Ingestion Timeout (Azure Data Factory)
 - 7.5.2. Failure 2: Corrupted data in Delta Lake (silver layer)
 - 7.5.3. Failure 3: Pipeline dependency or trigger failure
 - 7.5.4. Failure 4: PowerBI not refreshing from gold layer
- 8. Conclusion
 - 8.1. Summary of project outcomes
 - 8.2. Business Impact



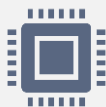
I. Introduction

I.1. Overview of Sofa-so good Cloud Architecture

Sofa-so-good's proposed cloud-based data architecture is designed to establish a comprehensive and agile data ecosystem, empowering the business to leverage its data assets for strategic advantage. This architecture focuses on building a scalable, secure, and efficient platform that centralizes diverse data sources, streamlines data processing, and facilitates advanced analytics. By adopting cloud-native technologies, Sofa-so-good will gain the flexibility to adapt to evolving business needs, optimize operational workflows, and drive data-informed decision-making across all departments. This solution emphasizes the creation of a unified, reliable data foundation that supports both real-time insights and long-term analytical capabilities, ultimately fostering a culture of data-driven innovation.



In today's dynamic business environment, data is a critical asset for Sofa-so-good.



The increasing volume and variety of data necessitate a modern, scalable solution.



This project addresses the need to overcome data silos and gain a unified view of business operations.



Our goal is to build a robust and scalable cloud-based data infrastructure for easy data access and informed decision-making.

<https://www.linkedin.com/in/praveen-chaudhary96/>



2. Project Purpose and Objectives

2.1. Objectives

To build a robust and scalable data infrastructure that enables Sofa-so-good to gain easy access to business data.

2.2. Project Purpose



Centralize and integrate disparate (unique) data sources.



Automate data processing and transformation for efficiency.



Enable easy data retrieval for advanced analytics and reporting.



Provide a single source of truth for all business data.



3. Mission and Goals

3.1. Mission Statement

To design and implement a high-performance, reliable data engineering pipeline that ensures accurate, readily accessible, and consistently available data, empowering Sofa-so-good with a foundational infrastructure for future analytics and operational improvements.

3.2. Project Goals

- **Centralize Data:** Create a unified repository for easy access and informed decision-making.
- **Optimize Data Flow and Pipeline Efficiency:** Improve data workflows by integrating real-time insights and automation for seamless processing and enhanced performance.
- **Future-Proof Data Pipelines:** Build adaptable data engineering systems that prioritize cost-efficiency, optimization, and reliability.



4. Design and Discovery Phase

4.1. Requirements Gathering



“KEY QUESTIONS”

- Which are our data sources? (Structured, Unstructured)?
- How data will be used?
- Who will be using the data?



4.2. Key Challenges and Opportunities

4.2.1. Key Challenges 🔑

- **Data Integration Complexity**
Integrating structured and unstructured data from multiple sources (e.g., sales systems, web logs, and inventory records) required precise ETL design and robust orchestration.
- **Cost Management**
Ensuring that the use of cloud services like Azure Synapse, Databricks, and Data Factory remained within budget while still delivering performance was a balancing act.
- **Data Quality and Governance**
Maintaining high data quality through the bronze, silver, and gold layers, while also implementing security and access control, required thoughtful planning and validation.
- **Tool Integration**
Ensuring seamless integration among different tools (Azure services, Power BI, ML frameworks) while maintaining scalability and reliability.

4.2.2. Key Opportunities 🚀

- **Scalability and Performance**
The architecture allows Sofa-so-good to scale up quickly as data volume grows, without a complete redesign.





<https://www.linkedin.com/in/praveen-chaudhary96/>



- **Advanced Analytics and AI**
With clean, organized data in the gold layer, there's huge potential for implementing machine learning models for forecasting, customer segmentation, and trend prediction.
- **Improved Decision-Making**
Interactive dashboards and real-time insights empower departments—from sales to operations—to make informed, data-driven decisions.
- **Operational Efficiency**
Automating data ingestion, transformation, and reporting saves time and reduces manual errors, freeing up staff for more strategic work.

5. Data Sources Overview

5.1. Data Inputs

	Sales Transaction Data (Physical stores)	Type: Structured Ingestion: Batch
	E-Commerce Data	Type: Structured Ingestion: Batch
	Warehouse Operational Data	Type: Structured Ingestion: Batch
	Website Logs	Type: Unstructured Ingestion: Batch



5.1.1. Sales Transactional Data:

- Nature: Detailed records of in-store and point-of-sale transactions, including product details, prices, quantities, dates, customer information, and payment methods.
- Format: Structured Data. (CSV, relational database tables (e.g., PostgreSQL, MySQL)).
- Ingestion type: Batch ingestion.
- Usage: Sales analysis, forecasting, customer behavior analysis, inventory management, and financial reporting.

5.1.2. E-Commerce Data:

- Nature: Online sales data, including product views, cart additions, checkout processes, order details, customer accounts, and online payment information.
- Format: Structured Data (Relational database tables).
- Ingestion type: Stream ingestion for real-time order processing, and batch ingestion for daily/weekly reports.
- Usage: Online sales analysis, customer journey tracking, website conversion optimization, and personalized online marketing.

5.1.3. Website Logs:

- Nature: Detailed records of user interactions on the Sofa-so-good website, including page views, clicks, search queries, session durations, and device information.



- Format: Unstructured data (log files).
- Ingestion type: Stream ingestion for real-time web traffic analysis.
- Usage: Website analytics, user behavior analysis, A/B testing, and marketing campaign performance evaluation.

5.1.4. Warehouse Operational Data:

- Nature: Data related to warehouse inventory, stock movements, shipping, receiving, and order fulfillment processes.
- Format: Structured Data (CSV, database tables).
- Ingestion type: Batch ingestion for daily/weekly inventory updates.
- Usage: Inventory management, stock optimization, order fulfillment tracking, and supply chain analysis.

5.2. Data Outputs / Key Deliverables

5.2.1. Sales Performance Dashboards:

- Real-time sales tracking, product performance, and regional sales analysis.
- User needs: Sales and management teams for monitoring sales trends and performance.

5.2.2. Inventory Management Reports:

- Stock levels, reorder points, and inventory turnover rates.
- User needs: Operations and inventory management teams for optimizing stock levels.

<https://www.linkedin.com/in/praveen-chaudhary96/>



5.2.3. **Customer Segmentation Reports:**

- Customer demographics, purchase behavior, and personalized recommendations.
- User needs: Marketing and sales teams for targeted campaigns and customer relationship management.

5.2.4. **Sales Forecasting Reports:**

- Predictive analysis of future sales based on historical data and market trends.
- User needs: Management and sales teams for resource allocation and planning.

5.2.5. **Web Analytics Dashboards:**

- Website traffic, user engagement, and conversion rates.
- User needs: Marketing and web development teams for optimizing website performance.

5.3. **List of Data Users**

5.3.1. **Marketing Team:** Uses customer segmentation, web analytics, and sales data for targeted campaigns and personalized marketing.

5.3.2. **Sales Team:** Uses sales performance dashboards, customer purchase history, and forecast reports to improve sales strategies and customer interactions.

<https://www.linkedin.com/in/praveen-chaudhary96/>



5.3.3. **Operations Team:** Uses inventory management reports and supply chain analysis to optimize stock levels and warehouse operations.

5.3.4. **Management Team:** Uses sales performance dashboards, forecast reports, and overall business analytics for strategic decision-making.

5.3.5. **Customer Service Team:** Uses customer data and purchase history to provide personalized and efficient customer support.

5.3.6. **Web Development Team:** Uses web analytics and API data to improve website functionality and customer experience.

6. Cloud Architecture Phases

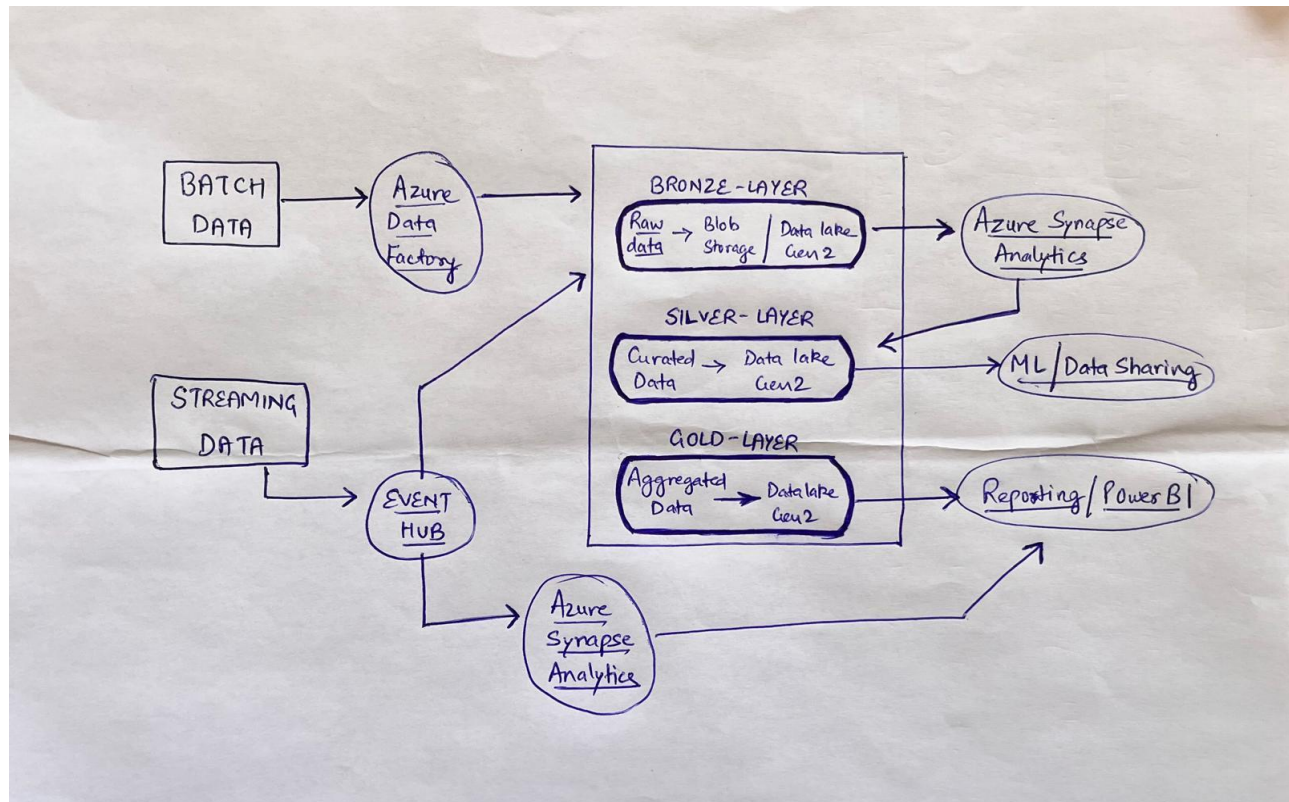
6.1. 🏠 (PHASE – 1): Initial Design

6.2. 🛠️ (PHASE – 2): Refined Design

6.3. 🚩 (PHASE – 3): Final Architecture Design



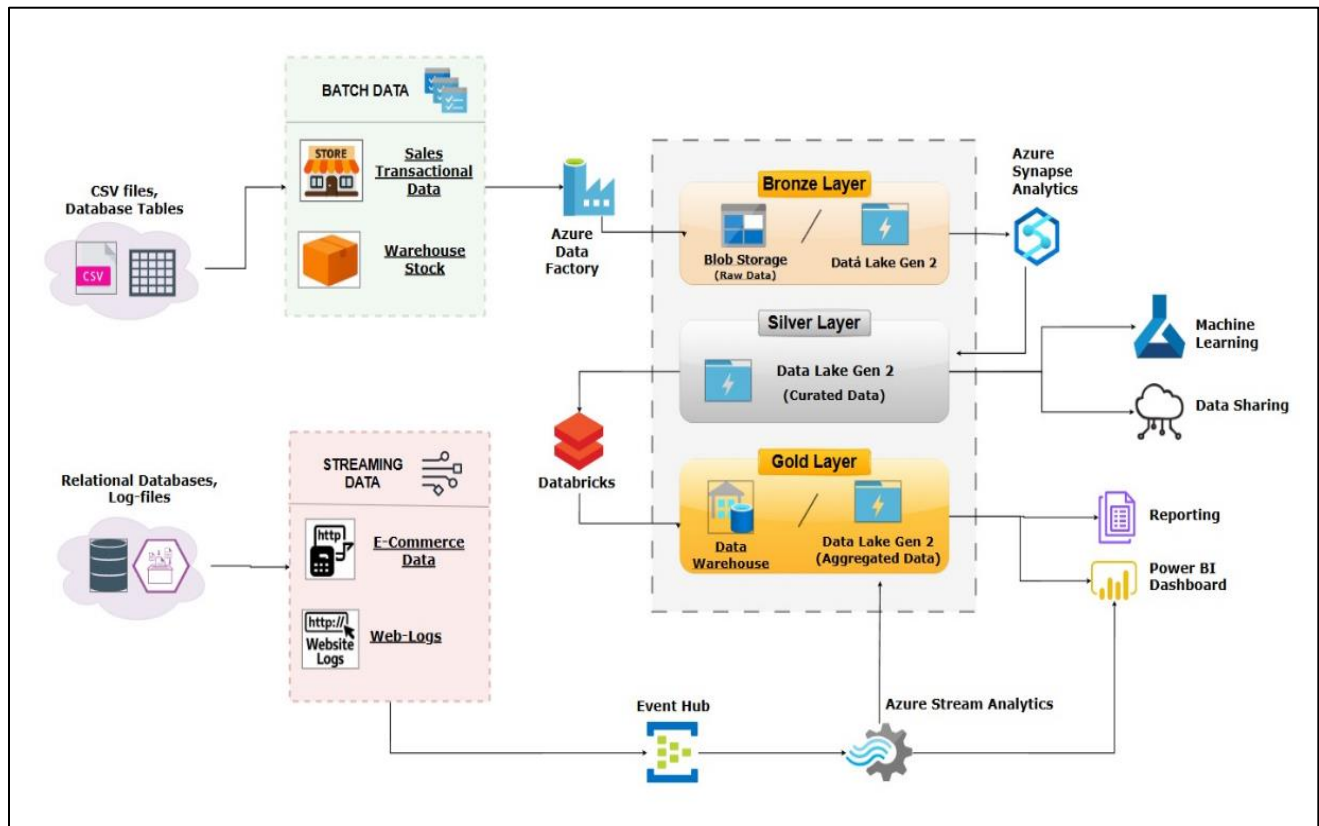
□ (PHASE – I): Initial Design



The architecture uses **Azure Data Factory** for batch data and **Event Hub** for streaming. Data flows through three layers (Bronze → Silver → Gold) in **Azure Data Lake Gen 2**, refining raw data into analytics-ready formats. **Azure Synapse Analytics** powers transformations, feeding insights into **Power BI** and ML models. This scalable pipeline ensures reliable, actionable data for Sofa So Good's business needs.



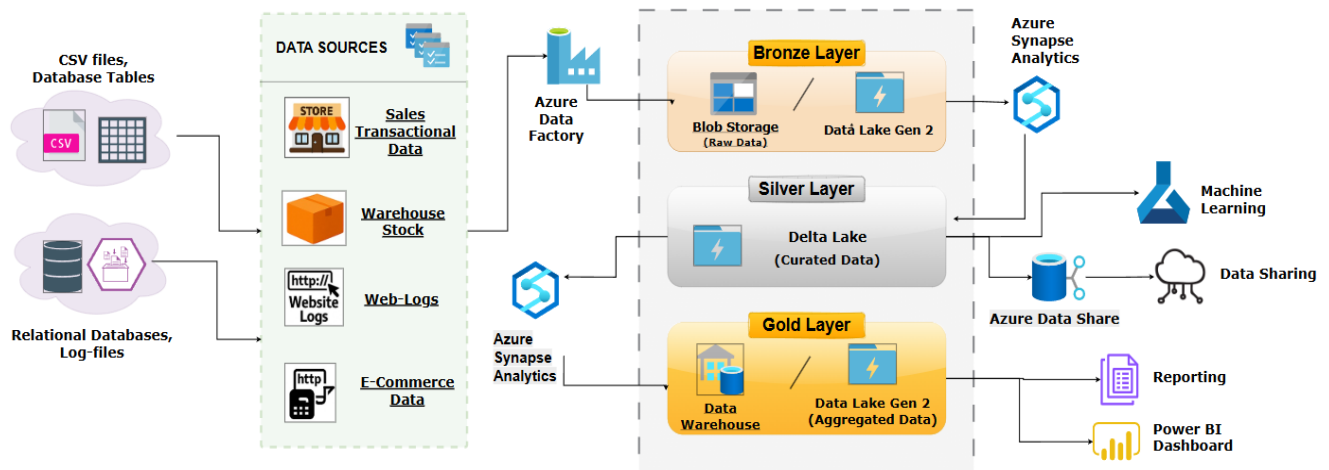
(PHASE – 2): Refined Design



This **Phase 2 Cloud Architecture** outlines the foundational setup for a data Lakehouse solution using Azure services. It integrates **batch data** (from sales and warehouse systems) and **streaming data** (e-commerce and web logs) into a unified data pipeline. Data is ingested through **Azure Data Factory** and **Event Hub**, then processed in **Databricks**, and stored in structured layers: **Bronze (raw)**, **Silver (curated)**, and **Gold (aggregated)** on **Data Lake Gen 2**. This structured approach supports analytics, machine learning, and business intelligence through tools like **Synapse**, **Power BI**, and **Azure ML**.



(PHASE – 3): Final Architecture Design



This final **Phase Cloud Architecture** represents a refined version of the initial design, streamlining data flow and enhancing processing capabilities. Data from various sources—CSV files, transactional systems, and log files—is ingested using **Azure Data Factory** and processed through **Azure Synapse Analytics**. The architecture utilizes a **Delta Lake** in the Silver layer for improved data curation and consistency, while the gold layer supports analytical workloads through a **Data Warehouse** and **Data Lake Gen 2**. Outputs serve key functions like **Power BI** dashboards, reporting, machine learning, and secure data sharing via **Azure Data Share**.



7. Process of creating pipeline

7.1. Bronze Layer (Ingested Data)

- ❖ In the Sofa-So Good architecture, we established data pipelines for collecting both structured and semi-structured data from multiple sources. This included batch data and real-time logs, ensuring comprehensive ingestion coverage.
- ❖ For batch ingestion, we utilized Azure Data Factory to extract and load data from CSV files, database tables, and inventory systems. This method is ideal for handling high-volume historical data for sales and warehouse transactions.
- ❖ For semi-streaming ingestion, data such as website logs and e-commerce activity are collected and moved periodically (Tumbling – Window Operation) through Azure Synapse Analytics, simulating near-real-time capture while maintaining processing efficiency.
- ❖ By routing all ingested data into Blob Storage and Data Lake Gen 2, the Bronze Layer ensures raw data is securely stored and readily accessible for transformation in later layers.

7.2. Silver Layer (Transform & Curate Data)

- ❖ In the Sofa-So Good architecture, the Silver Layer focuses on transforming raw data into structured, high-quality datasets suitable for analysis.
- ❖ Using Delta Lake, we cleaned, validated, and standardized data from the Bronze Layer. This includes refining sales records, filtering out anomalies in warehouse logs, and structuring web activity data. The transformation process ensures consistency and accuracy, enabling reliable downstream processing.
- ❖ Azure Synapse Analytics plays a key role here, allowing scalable querying and processing of curated datasets. This curated data is now ready for advanced analytics, business intelligence, and machine learning applications.

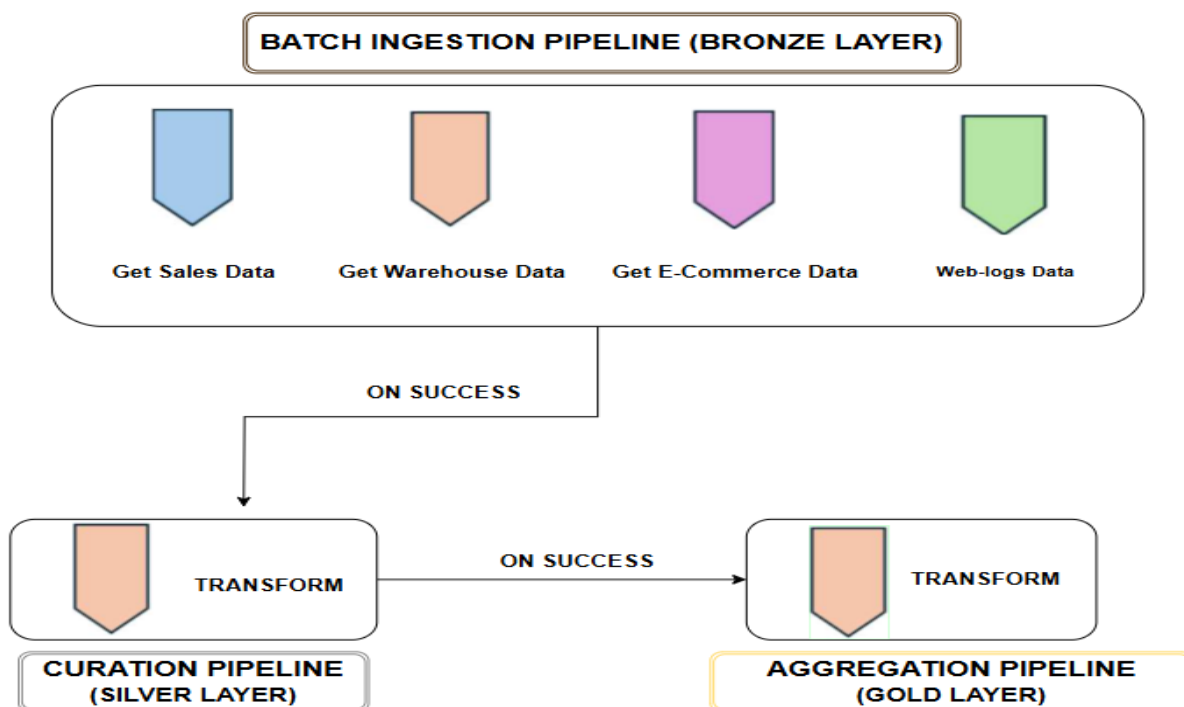


7.3. **Gold Layer (Aggregated Data)**

- ❖ The Gold Layer is where meaningful insights are derived from curated data. In Sofa-So Good, we aggregate the transformed data using Azure Synapse Analytics and store it in a Data Warehouse and Data Lake Gen 2.
- ❖ This layer supports complex analytics such as sales trend forecasting, inventory optimization, and customer behavior analysis. Aggregated datasets from this layer are used for:
 - Power BI Dashboards for interactive visualizations
 - Reporting on business operations and decision-making
 - Machine Learning models to predict demand and personalize experiences
 - Azure Data Share to securely share data with internal teams or partners
- ❖ The Gold Layer transforms data into a strategic asset, enabling data-driven decisions across the organization.

7.4. **Pipeline Approach (event-based)**

- ❖ We set up our data pipelines using an event-based approach to ensure smooth and efficient data processing.
- ❖ A master pipeline was created to oversee the execution of all other pipelines. It triggers each pipeline based on specific events, such as whether the previous pipeline ran successfully or encountered a failure



To keep the system running regularly, the master pipeline is scheduled to run every 24 hours, starting 12:01 AM. This setup ensures that all data processes are automated using 'tumbling – window operation' and completed on time, reducing manual effort and minimizing delays in data processing. All the processes in the master pipeline follows 'event – based approach' to oversee the execution of all other pipelines systematically.



7.5. Pipeline Failures and Solutions

7.5.1. Failure 1: Data Ingestion Timeout (Azure Data Factory)

Issue:

Data pipelines may fail to ingest large files due to network latency or resource throttling.

Solution:

- ❖ Increase timeout settings and enable retry policies in Azure Data Factory.
- ❖ Break large files into smaller chunks using chunking strategies.
- ❖ Use integration runtimes optimized for your data volume and region.

7.5.2. Failure 2: Corrupted Data in Delta Lake (Silver Layer)

Issue:

Ingested data may contain nulls, unexpected characters, or wrong formats, causing transformation errors.

Solution:

- ❖ Add schema validation and data profiling at the ingestion stage.
- ❖ Implement Try-Catch error handling logic in Data Flows.
- ❖ Use Delta Lake's MERGE and UPDATE capabilities to clean and overwrite only the problematic data.



7.5.3. Failure 3: Pipeline Dependency or Trigger Failure

Issue:

A downstream pipeline doesn't trigger due to a dependency failure in a prior step.

Solution:

- ❖ Enable pipeline dependency chaining with proper success/failure conditions.
- ❖ Use Azure Data Factory's custom alerts and logging to identify where and why the failure occurred.

7.5.4. Failure 4: Power BI Not Refreshing from Gold Layer

Issue:

Reports fail to refresh due to connection timeouts or expired credentials.

Solution:

- ❖ Schedule refreshes during off-peak hours to avoid contention.
- ❖ Monitor dataset refresh logs in Power BI Service.
- ❖ Ensure that credentials used to connect to Azure Synapse are valid and updated.



8. Conclusion □

The Sofa-So Good data architecture successfully delivers a modern, scalable, and reliable data pipeline using the medallion architecture (Bronze, Silver, Gold) with Azure services at its core. By integrating diverse data sources—ranging from CSV files, sales transactions, warehouse stock, to website logs—into a unified data Lakehouse framework, the project streamlines data ingestion, processing, and consumption with precision.

8.1. Summary of project outcomes ✓

- **Efficient Ingestion Pipelines**

Both batch (via Azure Data Factory) and streaming (via Azure IoT Hub) pipelines were deployed, allowing seamless collection of real-time and historical data.

- **Clean & Curated Datasets (Silver Layer)**

With Delta Lake at the center, raw data was transformed into curated datasets, ready for analytics and modeling. This ensured high data quality and consistency across the platform.

- **Aggregated Insights & Advanced Analytics (Gold Layer)**

Aggregated data in the Gold Layer was used for high-level reporting and business intelligence using Power BI and Azure Synapse Analytics.

<https://www.linkedin.com/in/praveen-chaudhary96/>



It also supported machine learning workloads, enabling predictive insights for business decisions.

- **Robust Error Handling**

Potential pipeline failures were mitigated through validation steps, automated retries, and proactive monitoring—ensuring resilience and continuous data availability.

8.2. **Business Impact**

This architecture empowers stakeholders to:

- Make real-time data-driven decisions.
- Monitor and optimize inventory, sales, and web traffic effectively.
- Gain actionable insights via interactive dashboards.
- Build a foundation for future integration of AI/ML models for forecasting and personalization.

<https://www.linkedin.com/in/praveen-chaudhary96/>