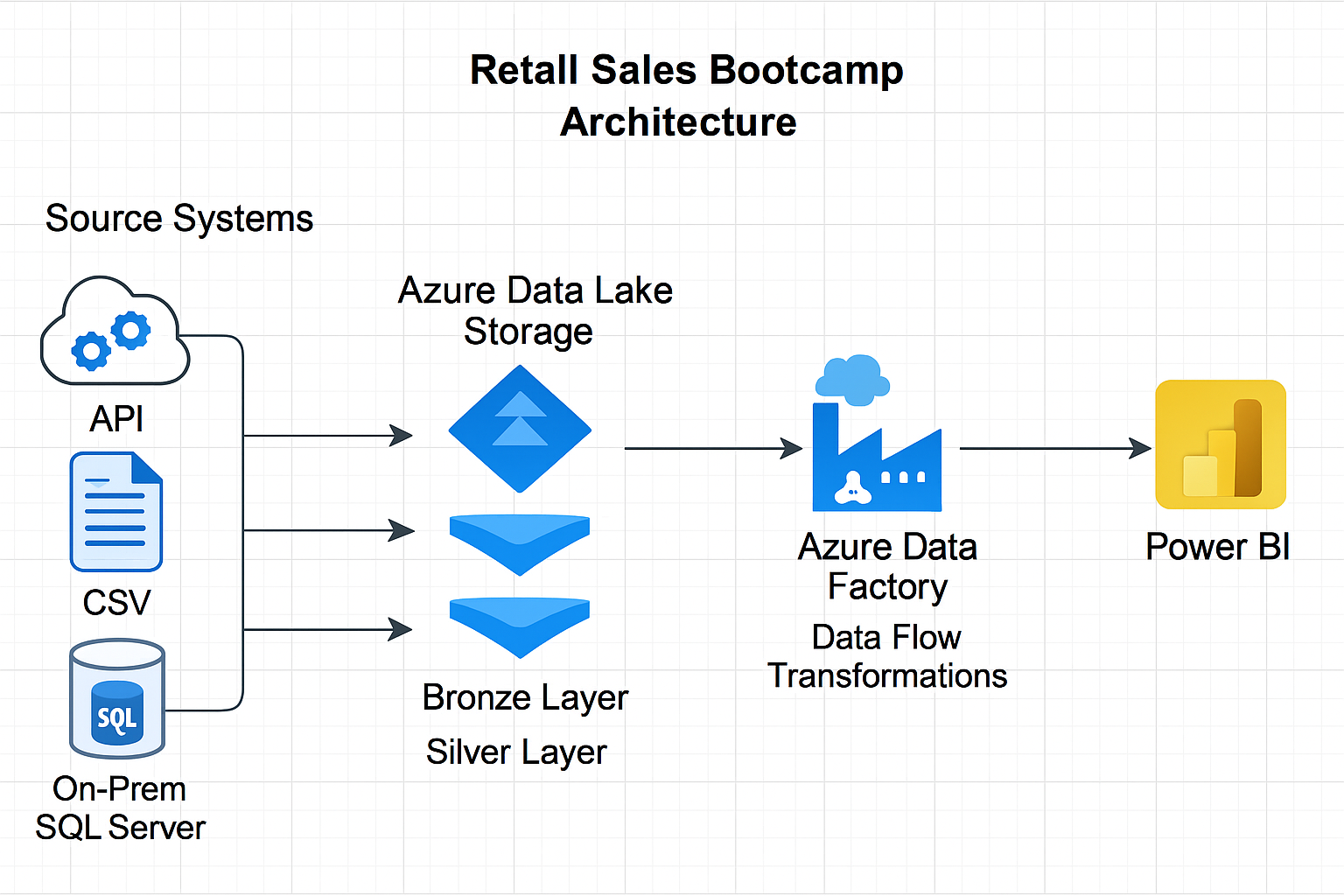
Retail Sales Data Pipeline - Project Documentation

# 1. Introduction

This document provides a comprehensive overview of the implementation of a Retail Sales Data Pipeline using Azure services. It includes design details, data flow processes, configuration, and deployment strategies. The solution handles data ingestion from multiple sources, transformation, storage, and finally visualization using Power BI.

# 2. Architecture Overview

## 2.1 Architecture Diagram

Diagram: 

## 2.2 Component Description

• Retail API (JSON): Provides real-time sales data via RESTful API using token-based authentication.

• CSV Files: Structured sales data in CSV format from public sources like Kaggle.

• On-Prem SQL Server: Contains customer, product, and historical sales data.

• Azure Data Factory: Orchestrates data ingestion, transformation, and movement.

• ADLS Gen2: Stores raw (Bronze) and transformed (Silver) data in hierarchical structure.

• Azure SQL Database: Stores cleaned and enriched data for reporting.

• Power BI: Connects to Azure SQL Database to build dashboards and visualizations.

# 3. Data Ingestion

## 3.1 API Ingestion

• Uses Web Activity and Copy Activity in ADF.

• Target Path: /bronze/api/

• Format: JSON

• Validation: Row count checks using If Condition activity.

Screenshot: A screenshot of a computer

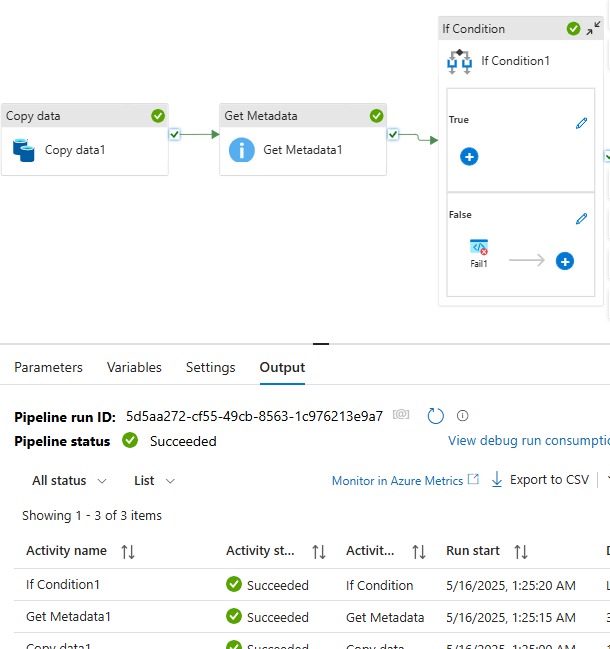
AI-generated content may be incorrect.

## 3.2 CSV Ingestion

• Loads CSV from public dataset (e.g., Kaggle) to /bronze/csv/

• Format: CSV

• Validation: Row count match with original file.

Screenshot: 3.3 On-

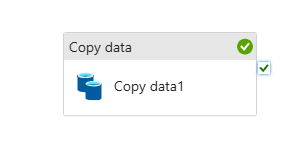
**Prem SQL Server Ingestion**

• Uses Self-hosted Integration Runtime (SHIR) to connect.

• Loads tables (Customers, Products, Sales) to /bronze/sql/

• Format: Parquet

• Validation: Compare row count with source tables.

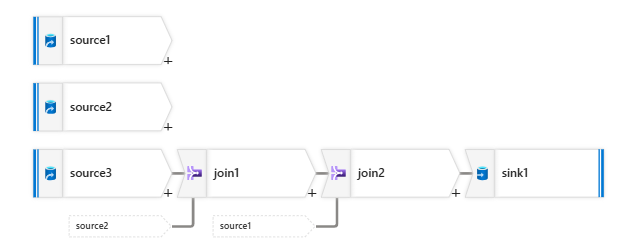


# 4. Data Transformation

• Transformation done using Mapping Data Flows in ADF.

• Cleans and joins datasets; aggregates monthly product sales.

• Outputs to /silver/ in Parquet format.

Screenshot: **5. Data Loading**

• Loads transformed data from /silver/ to Azure SQL Database.

• Write Behavior: Truncate and Load

• Optimized with batch size setting.

# 6. Data Visualization

• Power BI connected via Azure SQL Database connector.

• Data Model: Includes Customers, Products, Sales with defined relationships.

• Visuals: Sales over time, product performance, customer insights.

Screenshot: A close-up of a graph

AI-generated content may be incorrect.

# 7. Appendices

## 7.1 DDL Scripts

• Customers (CustomerID, Name, Email, Region)

• Products (ProductID, Name, Category, Price)

• Sales (SaleID, ProductID, CustomerID, Date, Quantity, TotalAmount)

## 7.2 Pipeline JSON Definitions

• ADF JSON pipeline files available upon request or export.

## 7.3 Notes

• Error Handling: Retry policies, failure alerts.

• Performance: Parallel copy, data partitioning, optimized queries.

• Security: RBAC, Private Endpoints, Key Vault integration.

• Future Improvements: Delta Lake integration, CI/CD for ADF.