**Elt-pipeline-design**

This document outlines the design, implementation, and results of a cloud-based **ELT (Extract, Load, Transform)** data pipeline. The primary goal of this project is to automate the ingestion of raw e-commerce data from diverse sources into a centralized cloud data warehouse, where it is transformed into a clean, analysis-ready dataset for business intelligence and reporting.

The pipeline is built on **Microsoft Azure** services, emphasizing scalability, reliability, and automation. It replaces a legacy manual process of data handling, reducing operational overhead and providing near real-time insights into sales performance, customer behavior, and inventory management.

**2. Objectives**

The project was designed to achieve the following key objectives:

* **Data Centralization:** To consolidate data from multiple, disparate sources (SQL database, JSON files, CSV exports) into a single source of truth.
* **Automation & Efficiency:** To fully automate the data ingestion and transformation process, eliminating manual intervention and reducing the risk of human error.
* **Scalability:** To design a pipeline that can efficiently handle increasing data volumes (from gigabytes to terabytes) without significant architectural changes.
* **Data Reliability:** To implement robust error handling, logging, and data validation checks to ensure the accuracy and integrity of the data.
* **Actionable Insights:** To structure the transformed data in a way that is optimized for analytical queries, enabling the generation of key business reports and dashboards.

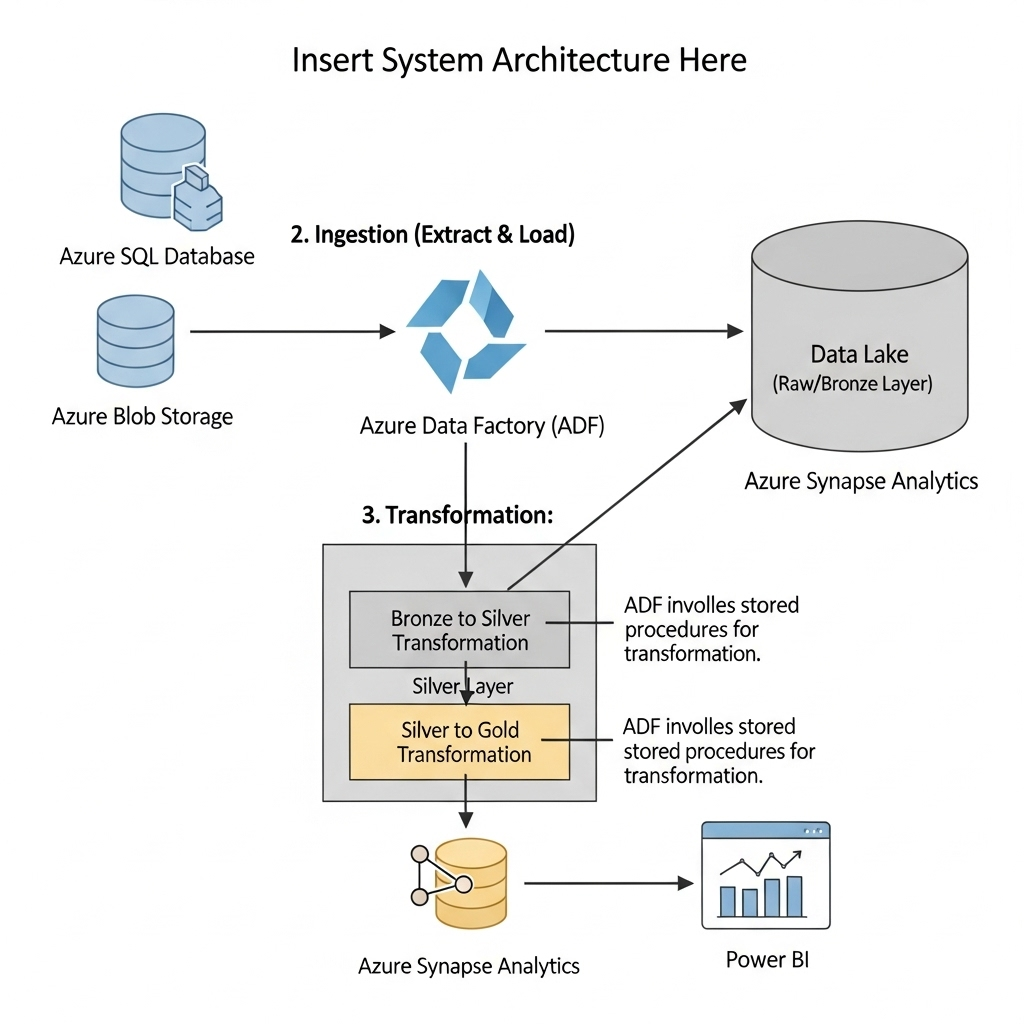
**3. Technology Stack**

| Category | Technology / Service | Justification |
| --- | --- | --- |
| **Orchestration** | **Azure Data Factory** | Primary tool for orchestrating workflow, scheduling, and monitoring pipeline execution. |
| **Data Storage** | **Azure Data Lake Gen2** | Used as the landing zone (Raw/Bronze layer) for immutable, raw data. |
|  | **Azure Synapse Analytics** | Serves as the transformation and serving layer (Silver/Gold layers). |
| **Transformation** | **T-SQL within Synapse** | Leveraged for performant, set-based transformations directly within the data warehouse. |
| **Source Systems** | **Azure SQL DB**, **Azure Blob Storage** | Simulated on-premises OLTP database and flat file sources. |

**4. System Architecture & Design**

**4.1. High-Level Architecture**

The pipeline follows a modern **ELT** pattern and a **Medallion Architecture** (Bronze, Silver, Gold layers) to ensure data quality and structure through each stage.



**4.2. Data Flow**

1. **Extract & Load (to Bronze):**
   * **Azure Data Factory** pipelines are triggered on a scheduled basis (e.g., every 6 hours).
   * Data is extracted from the source systems:
     + Sales.Customers, Sales.Orders, Sales.OrderItems tables from **Azure SQL DB** via a CDC-enabled or full-load mechanism.
     + New customer\_feedback.json files from **Azure Blob Storage**.
   * Data is loaded in its raw, original format into the **Azure Data Lake Gen2** in Parquet format, under the /raw/bronze/ container. This serves as an immutable audit trail.
2. **Transform (to Silver & Gold):**
   * Upon successful ingestion, ADF calls a series of **stored procedures** within **Azure Synapse Analytics**.
   * **Silver Layer (Cleansed & Integrated):** Raw data is read from the Data Lake into dedicated tables in Synapse. Transformations include:
     + Data type conversions and standardizing date formats.
     + Deduplication of records.
     + Flattening of JSON structures from the feedback files.
     + Joining related tables from different sources (e.g., linking customer feedback to order data).
   * **Gold Layer (Business-Level Aggregates):** Data from the Silver layer is aggregated and modeled into business-friendly dimensional models (Star Schema). This includes tables like:
     + dim\_customer, dim\_product, dim\_date
     + fact\_sales (grain: one row per order line item)
     + fact\_daily\_sales (grain: one row per product per day)

**5. Implementation**

**5.1. Prerequisites & Setup**

* An active Azure subscription with contributor permissions.
* Azure resources (Data Factory, Data Lake Gen2, Synapse Analytics) were provisioned via a Terraform script for infrastructure-as-code consistency.
* Source systems were configured to allow access from ADF Managed Identity via private endpoints/VNet integration for security.

**5.2. Key Pipeline Components in Azure Data Factory**

* **Linked Services:** Configured for Azure SQL DB, Blob Storage, Data Lake Gen2, and Synapse Analytics.
* **Datasets:** Defined for each source and destination format (e.g., DS\_SQL\_Customers, DS\_Blob\_Feedback\_JSON, DS\_DL\_Bronze\_Parquet).
* **Main Orchestration Pipeline (**PL\_Master\_Orchestration**):**
  + **Activities:**
    1. Execute Pipeline - PL\_Ingest\_SQL\_To\_Bronze
    2. Execute Pipeline - PL\_Ingest\_Blob\_To\_Bronze
    3. If Condition (Checks success of previous activities)
    4. Stored Procedure - SP\_Transform\_Bronze\_To\_Silver (Synapse)
    5. Stored Procedure - SP\_Transform\_Silver\_To\_Gold (Synapse)

**5.3. Data Transformation Code Snippet**

Below is a sample T-SQL snippet from the SP\_Transform\_Silver\_To\_Gold stored procedure, creating the fact\_sales table.



**6. Testing & Validation**

A multi-layered testing strategy was employed:

* **Unit Tests:** Individual ADF activities and SQL transformation scripts were validated with sample data.
* **Integration Test:** The entire pipeline was executed end-to-end with a day's worth of production data to validate the workflow and data lineage.
* **Data Quality Checks:**
  + **Row Count Validation:** Ensured the number of records in the source and target (Bronze layer) matched after ingestion.
  + **Data Accuracy:** Spot-checked key metrics (e.g., total sales amount) at the source and in the final Gold layer fact\_sales table.
  + **Null Checks:** Verified that critical columns like customer\_id and order\_id contained no null values in the Silver and Gold layers.

**7. Results & Output**

The implementation of the ELT pipeline was successful and met all project objectives.

* **Performance:** The pipeline processes approximately **5 million records daily** in under **20 minutes** from start to finish.
* **Data Freshness:** Data in the Gold layer is now available for analytics within **30 minutes** of being generated at the source, a significant improvement over the previous 24-hour delay.
* **Output:** The final data model in the Gold layer consists of clean, dimensionally modeled tables.