**Data Pipeline with Azure Data Factory, Databricks and Dashboard on Restaurants Dataset | Swiggy**

**1. Overview**

**Purpose**

This document outlines the architecture and design of a data pipeline that supports data ingestion, transformation, and visualization using Azure Data Factory, Delta Lake, Databricks.

The purpose of this pipeline is to:

1. Ingest raw data from various data sources.
2. Transform the raw data into refined and structured formats (Silver and Gold layers).
3. Enable insights and decision-making through interactive dashboards.

**2. Architecture**

**Pipeline Stages**



The architecture consists of three key stages:

1. **Staging**
2. **Transformation**
3. **Visualization**

**2.1 Staging**

**Components**

* **Data Source (CSV):** Represents raw data generated by external systems or business processes in CSV format.
* **Azure Data Factory (ADF):**
  + **Role:** Orchestrates data ingestion.
  + **Functionality:** Extracts data from the source and loads it into the raw data storage (Data Lake).
  + **Triggers:** Scheduled or event driven.
* **Azure Data Lake (Raw Data Store):**
  + **Role:** Stores raw ingested data.
  + **Storage Format:** Parquet, CSV, or JSON.

**Process Flow**

1. ADF connects to the source (CSV) and ingests the data.
2. Data is stored in its raw format in the Data Lake for further processing.

**2.2 Transformation**

**Components**

* **Databricks:**
  + **Role:** Data transformation and refinement.
  + **Technology:** Apache Spark-based compute engine for distributed processing.
  + **Functionality:**
    - Cleans and prepares raw data.
    - Converts data into structured formats for Silver and Gold layers.
* **Delta Lake (Silver and Gold Tables):**
  + **Silver Layer:** Intermediate, cleansed data layer. Used for applying data quality checks and basic transformations.
  + **Gold Layer:** Final, business-ready data layer. Aggregated and enriched data for reporting and analytics.

**Process Flow**

1. Databricks reads raw data from the Data Lake.
2. **Silver Table:**
   * Cleansing and basic transformations (e.g., null handling).
   * Stores intermediate data for reusable processing.
3. **Gold Table:**
   * Aggregation, enrichment, and final transformation.
   * Stores data ready for analysis.

**2.3 Visualization**

**Components**

* **Databricks:**
  + Provides the processed Gold Layer data for dashboards.
  + Enables interactive dashboards and visualizations for business users.

**Process Flow**

1. Connects to the Gold Table using Databricks.
2. Dashboards are built.

**3. Key Design Considerations**

**3.1 Scalability**

* Use Azure Data Factory's scalable pipelines to handle large data volumes.
* Leverage Databricks' distributed compute for efficient transformations.

**3.2 Performance**

* Optimize Data Lake storage with partitioning and efficient file formats (e.g., Parquet).
* Utilize Delta Lake's ACID properties for consistency and high-speed queries.

**3.3 Security**

* Implement role-based access controls (RBAC) on the Data Lake.
* Encrypt data in transit and at rest using Azure security features.

**4. Tools and Technologies**

| **Tool/Technology** | **Purpose** |
| --- | --- |
| CSV Files | Data source format |
| Azure Data Factory | Data ingestion orchestration |
| Azure Data Lake | Raw and intermediate data storage |
| Databricks | Data transformation and processing |
| Delta Lake | Data versioning and ACID storage |
| Power BI/Databricks | Dashboarding and visualization |

**5. Implementation Plan**

**5.1 Development Steps**

1. **Staging:**
   * Set up ADF pipeline to ingest data into the Data Lake.
2. **Transformation:**
   * Build Databricks notebooks to create silver and gold tables.
   * Apply Delta Lake for versioning and transactional consistency.
3. **Visualization:**
   * Connect Power BI to Gold Table for visualization.

**5.2 Testing**

* Verify data accuracy at each stage (Raw, Silver, Gold).
* Validate dashboards for correct insights.