**Phase 3 Document: Advanced Databricks and Azure Integration**

**Objective**

Focus on building **data processing notebooks** in Azure Databricks, developing **workflows**, integrating with **Azure Data Factory**, and **connecting to Power BI** for reporting.

**1. Advanced Data Processing with PySpark**

Key areas to focus on for data engineering development inside Databricks notebooks:

* **PySpark Essentials**
  + Master DataFrame operations, transformations, and actions.
  + Implement efficient and scalable data processing pipelines.
* **Handling Slowly Changing Dimensions (SCD Type 1)**
  + Implement **SCD Type 1** logic using PySpark.
  + Focus on **overwriting** existing records with the latest data (no history tracking).
  + Use simple merge/upsert strategies into Delta Tables.
* **Star Schema Modeling**
  + Design and implement **Fact Tables** and **Dimension Tables** following star schema principles.
  + Ensure the data model is optimized for analytics and reporting.
* **Surrogate Keys Management**
  + Generate **surrogate keys** in PySpark to uniquely identify dimension table records.

**2. Create Databricks Notebooks and Workflows**

* **Create Databricks Notebooks**:
  + Develop modular ETL notebooks that handle:
    - Raw Data Ingestion
    - Data Cleansing and Validation
    - Data Transformation (Business Rules and SCD Type 1 updates)
    - Loading Processed Data into Delta Tables (Lakehouse)
* **Build Databricks Workflows**:
  + Set up **Workflows** to orchestrate multiple notebook executions.
  + Configure task dependencies and scheduling for automation.
  + Enable monitoring and alerting for successful and failed runs.

**3. Integration with Azure Data Factory (ADF)**

* **Create Azure Data Factory Pipelines**:
  + Link ADF to Databricks using **Databricks Linked Services**.
  + Trigger Databricks notebook activities from ADF pipelines.
  + Set up monitoring, retries, and error handling in the ADF pipelines.

**4. Connect Databricks to Power BI**

* **Create Direct Connection**:
  + Configure Databricks SQL Warehouses (Endpoints) or ODBC/JDBC connections.
  + Connect Power BI Desktop or Power BI Service directly to Databricks.
  + Schedule regular dataset refreshes to provide near real-time analytics.