**Project Flow: Sales Data Consolidation and Analysis Platform**

**1. Data Ingestion**

* **Sources**: Extract raw sales data from multiple databases (SQL Server and Oracle).
* **Process**: Use **Azure Data Factory** to set up pipelines that pull data from these sources at scheduled intervals.
* **Output**: Load the extracted data into a raw data layer in **ADLS Gen 2**, maintaining its original format for traceability.

**2. Data Transformation**

* **Platform**: Utilize **Azure Databricks** to process raw data.
* **Process**:
  + Write **PySpark jobs** to clean, transform, and standardize the data (e.g., handling missing values, formatting dates).
  + Perform aggregations to consolidate sales metrics, such as total sales by region or product.
* **Output**: Store the transformed data in a curated layer in **ADLS Gen 2**, with optimized partitioning (e.g., by region, date) for faster access.

**3. Data Validation and Quality Checks**

* **Process**: Implement automated validation scripts in **PySpark** to ensure data accuracy (e.g., row counts, format checks, duplicate handling).
* **Alerts**: Set up automated notifications for pipeline failures or validation errors using **Azure Monitor** or **Azure Data Factory alerts**.

**4. Reporting and Visualization**

* **Platform**: Use **Power BI** to create interactive dashboards.
* **Process**:
  + Connect Power BI to the curated data in **ADLS Gen 2** for direct querying.
  + Build dashboards to display key sales metrics (e.g., sales performance by region, product trends, year-over-year growth).
  + Share insights with stakeholders for informed decision-making.

**5. Monitoring and Optimization**

* **Cluster Management**: Monitor and optimize **Databricks clusters** to ensure efficient resource utilization and minimize costs.
* **Pipeline Monitoring**: Use **Azure Data Factory monitoring tools** to track the status of data pipelines in real-time.

**High-Level Tools and Techniques Used**

* **Azure Data Factory**: Data ingestion and pipeline orchestration.
* **Azure Databricks**: Data transformation and aggregation using **PySpark**.
* **ADLS Gen 2**: Scalable and optimized data storage.
* **Power BI**: Data visualization and reporting.
* **SQL Server and Oracle**: Data sources for sales data.