**WNS Case Study for Data Engineering**

WNS Services, a leading retail company, needs to analyze sales data from both physical and online stores. The goal is to process and analyze this data to gain insights into customer behavior and sales trends. The data sources include:

* Batch data in CSV and Parquet formats from point-of-sale systems.
* Real-time sales data in JSON format from online transactions.

It should include the following components:

1. Event Hubs for real-time data ingestion.
2. Azure Storage and Azure Data Lake Gen 2 for batch data storage.
3. Azure Synapse Analytics for data ingestion, processing, and storage.
4. Spark Notebooks for data transformation.
5. Dedicated SQL Pool for storing processed data.
6. Azure Stream Analytics for real-time data processing.
7. Power BI for data visualization.

**Step-by-Step Implementation**

**1. Data Ingestion**

**Azure Storage and Data Lake Gen 2**

1. **Setup Azure Storage Account**:
   * Create an Azure Storage account with Data Lake Gen 2 enabled.
   * Create containers for storing CSV and Parquet files.
2. **Upload Batch Data**:
   * Upload CSV and Parquet files containing sales data to the respective containers in Data Lake Gen 2.

**Event Hubs**

1. **Setup Event Hubs**:
   * Create an Event Hub namespace and an Event Hub instance.
   * Configure the online sales platform to send real-time data in JSON format to the Event Hub (we can use Data Lake as an input for the online transaction data).

**2. Data Ingestion with Azure Synapse Analytics**

**Azure Synapse Analytics Workspace Setup**

1. **Create Synapse Workspace**:
   * Create an Azure Synapse Analytics workspace.
   * Set up linked services to connect to Event Hubs and Azure Data Lake Gen 2.

**Data Pipelines for Batch Ingestion**

1. **Create Batch Ingestion Pipeline**:
   * In Azure Synapse Studio, create a new data pipeline.
   * Add a Copy Data activity to move data from Azure Data Lake Gen 2 to a staging area in Synapse.

**Example Pipeline Steps:**

* + Source: Azure Data Lake Gen 2 (CSV/Parquet files)
  + Sink: Synapse Analytics staging table

**Data Pipelines for Streaming Ingestion**

1. **Create Streaming Ingestion Pipeline**:
   * In Azure Synapse Studio, create a new data pipeline for streaming data.
   * Use Azure Stream Analytics to process the streaming data and store it in a dedicated SQL pool.

**Example Pipeline Steps:**

* + Source: Event Hubs (JSON data)
  + Sink: Dedicated SQL Pool

**3. Data Processing and Transformation**

**Spark Notebooks**

1. **Create Spark Notebook**:
   * In Azure Synapse Studio, create a Spark Notebook.
   * Use PySpark to clean, transform, and aggregate data from CSV, JSON, and Parquet files.

**4. Data Storage**

**Dedicated SQL Pool**

1. **Create Dedicated SQL Pool**:
   * In Azure Synapse Studio, create a dedicated SQL pool.
   * Define tables to store the transformed data.

**5. Real-Time Data Processing**

**Azure Stream Analytics Job**

1. **Create Stream Analytics Job**:
   * In the Azure portal, create a Stream Analytics job to process real-time data from Event Hubs.
   * Define input from Event Hubs and output to the dedicated SQL pool.
2. **Stream Analytics Query**:
   * Write a query to analyze the streaming data and store it in the SQL pool.

**6. Data Visualization**

**Power BI**

1. **Connect Power BI to Dedicated SQL Pool**:
   * Open Power BI Desktop and connect to the dedicated SQL pool.
   * Import tables containing aggregated sales data.
2. **Create Dashboards and Reports**:
   * Use Power BI to create interactive dashboards and reports.

**Example Visualizations:**

* + **Real-time Sales Dashboard**: A dashboard showing current sales performance using real-time data.
  + **Historical Sales Trends**: A line chart displaying sales trends over time.
  + **Sales by Region and Product**: A bar chart comparing sales performance across different regions and products.