**Use Case: Retail Sales Analysis**

**Scenario:**

You are a data engineer at a large retail company. The company has stores in multiple regions, and you are tasked with analyzing the sales data to gain insights into customer behavior, product performance, and regional trends. You will need to process large volumes of sales data, apply transformations, and generate actionable insights.

**Data Sources:**

1. **Sales Data:**
   * sales.csv: Contains information about each transaction.
     + transaction\_id: Unique identifier for the transaction.
     + store\_id: Identifier for the store where the transaction took place.
     + product\_id: Identifier for the product sold.
     + quantity: Number of units sold.
     + price: Price per unit.
     + transaction\_date: Date of the transaction.
2. **Products Data:**
   * products.csv: Contains details about each product.
     + product\_id: Unique identifier for the product.
     + product\_name: Name of the product.
     + category: Category to which the product belongs.
3. **Stores Data:**
   * stores.csv: Contains details about each store.
     + store\_id: Unique identifier for the store.
     + store\_name: Name of the store.
     + region: Region where the store is located.

**Tasks:**

1. **Load and Prepare Data:**
   * Load the sales.csv, products.csv, and stores.csv files into PySpark DataFrames.
   * Perform necessary data cleaning (e.g., handling missing values, correcting data types).
2. **Join DataFrames:**
   * Join the sales DataFrame with the products DataFrame on product\_id.
   * Join the resulting DataFrame with the stores DataFrame on store\_id.
3. **Calculate Total Revenue:**
   * Calculate the total revenue for each transaction (i.e., quantity \* price).
   * Create a new column total\_revenue in the DataFrame to store this value.
4. **Analyze Product Performance:**
   * Group the data by product\_name and category and calculate the following:
     + Total units sold.
     + Total revenue generated.
     + Average revenue per unit.
5. **Analyze Regional Performance:**
   * Group the data by region and calculate the following:
     + Total revenue generated in each region.
     + Top-performing products in each region (based on revenue).
     + Store with the highest revenue in each region.
6. **Identify Seasonal Trends:**
   * Extract the month from the transaction\_date and analyze sales trends across different months.
   * Identify the peak sales months for each product category.
7. **Save the Results:**
   * Save the resulting DataFrames with the analyzed data (e.g., product performance, regional performance) as CSV files or Parquet files.

**Expected Deliverables:**

* **Product Performance Report**: A DataFrame showing the total units sold, total revenue, and average revenue per unit for each product, grouped by product category.
* **Regional Performance Report**: A DataFrame showing total revenue by region, top-performing products in each region, and the store with the highest revenue.
* **Seasonal Trends Report**: A DataFrame showing sales trends by month, highlighting peak sales periods.

**Extension:**

* **Advanced Analysis**: Use PySpark SQL to perform more complex queries, such as identifying the top 5 products by revenue in each region or finding the month-over-month growth rate for different product categories.