

Interview Case Study

1. Data Loading:

- Load the provided three files using either SQL or PySpark.
- Name each table with a raw_ prefix to differentiate between original and transformed data.

2. Data Review and Storage:

- Review the loaded data and assign appropriate data types based on your best judgment.
- Identify primary and foreign keys for each table.
- Store the transformed data using a store_ prefix.

3. Product Master Transformations:

- Perform the following transformations on the product master data and write the results into a table named publish_product:
 - i. Replace NULL values in the Color field with N/A.
 - ii. Enhance the ProductCategoryName field when it is NULL using the following logic:
 - If ProductSubCategoryName is in ('Gloves', 'Shorts', 'Socks', 'Tights', 'Vests'), set ProductCategoryName to 'Clothing'.
 - If ProductSubCategoryName is in ('Locks', 'Lights', 'Headsets', 'Helmets', 'Pedals', 'Pumps'), set ProductCategoryName to 'Accessories'.
 - If ProductSubCategoryName contains the word 'Frames' or is in ('Wheels', 'Saddles'), set ProductCategoryName to 'Components'.

4. Sales Order Transformations:

- Join SalesOrderDetail with SalesOrderHeader on SalesOrderId and apply the following transformations:
 - i. Calculate LeadTimeInBusinessDays as the difference between OrderDate and ShipDate, excluding Saturdays and Sundays.
 - ii. Calculate TotalLineExtendedPrice using the formula: OrderQty * (UnitPrice - UnitPriceDiscount).
 - iii. Write the results into a table named publish_orders, including:
 - All fields from SalesOrderDetail.

- All fields from SalesOrderHeader except SalesOrderId, and rename Freight to TotalOrderFreight.

5. Analysis Questions:

- Provide answers to the following questions based on the transformed data:
 - i. Which color generated the highest revenue each year?
 - ii. What is the average LeadTimeInBusinessDays by ProductCategoryName?