You should build a complete ETL pipeline in Snowflake Snowsight using a Sales dataset (CSV + Parquet). They will create stages, load raw data, clean and transform, build star-schema tables (fact + dimensions), and finally run reports for analytics.

## Assignment Tasks

### Task 1: Setup Database & Schema

1. Create a new database SALES\_DB and schema RAW\_SCHEMA.  
2. Create file formats for CSV and Parquet data (csv\_format, parquet\_format).  
3. Create a named stage sales\_stage in RAW\_SCHEMA.

### Task 2: Load Raw Data

1. Create a staging table sales\_raw with JSON columns for product\_details and customer\_info.  
2. Create another staging table Raw\_Parquet\_table for Parquet data.  
3. Load sales\_data\_1000.csv and sales\_data\_1000.parquet into the staging tables.

### Task 3: Data Quality Checks

1. Write queries to count total vs distinct transactions, null values, and negative values.

### Task 4: Flatten JSON Data

1. Extract fields from product\_details and customer\_info using Snowflake's JSON operators.

### Task 5: Clean & Transform

1. Create a cleaned table sales\_clean in CLEAN\_SCHEMA.  
2. Apply transformations like TRY\_TO\_DATE, replacing negatives with NULL, removing missing values.

### Task 6: Feature Engineering

1. Add new columns profit\_margin and sales\_quarter to sales\_clean.  
2. Update table with calculated values.

### Task 7: Create Fact & Dimension Tables

1. Create dimension tables DIM\_REGION, DIM\_PRODUCT, DIM\_CUSTOMER.  
2. Create fact table FACT\_SALES with transaction and sales metrics.

### Task 8: Reporting

Run queries:  
1. Sales by Region  
2. Customer Segment Analysis  
3. Product Brand Performance  
4. Order Status Distribution

Question 14: Which region had the highest sales?  
Question 15: Which product brand has the best rating-to-sales ratio?

## Deliverables

1. SQL scripts for each task (from stage → fact/dim → reports).