AMAZON SALES DATA PIPELINE USING SNOWFLAKE

PROJECT OVERVIEW:
A variety of data engineering components are included in the Amazon Sales Data Pipeline Project to convert unprocessed sales data into informative analytical reports. The pipeline uses Snowflake for compute and storage while processing data in a variety of forms, including CSV, JSON, and Parquet. Finally, explore the concept of dimensional modelling, a popular approach for organizing data into meaningful structures that facilitate analytical querying and reporting.

DATA PIPELINE ARCHITECTURE
The pipeline is divided into multiple zones within the architecture, each handling a distinct aspect:
- **Ingestion**: Snowflake's internal staging area receives data files through ingestion.
- **Source Zone**: First storage where information is organized by product, client, and sales.
- **Curated Zone**: Snowpark's DataFrame API is used to clean and modify data.
- **Consumption Zone:** Dimensional modeling of the data makes efficient and straightforward querying possible.
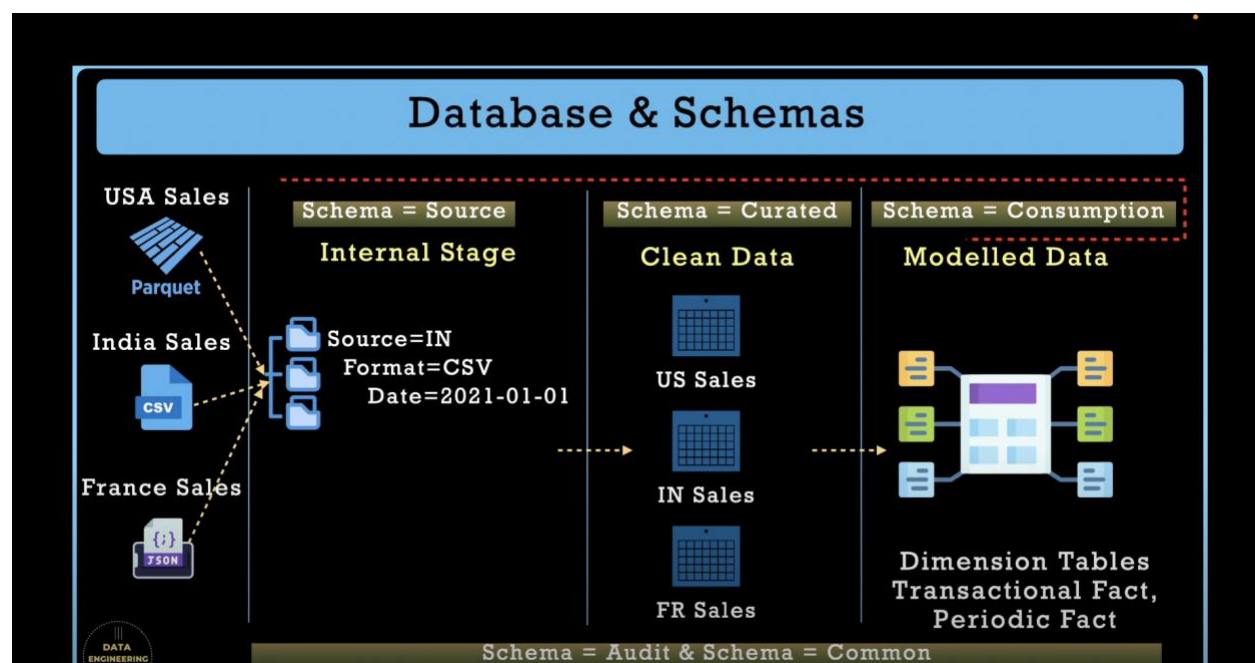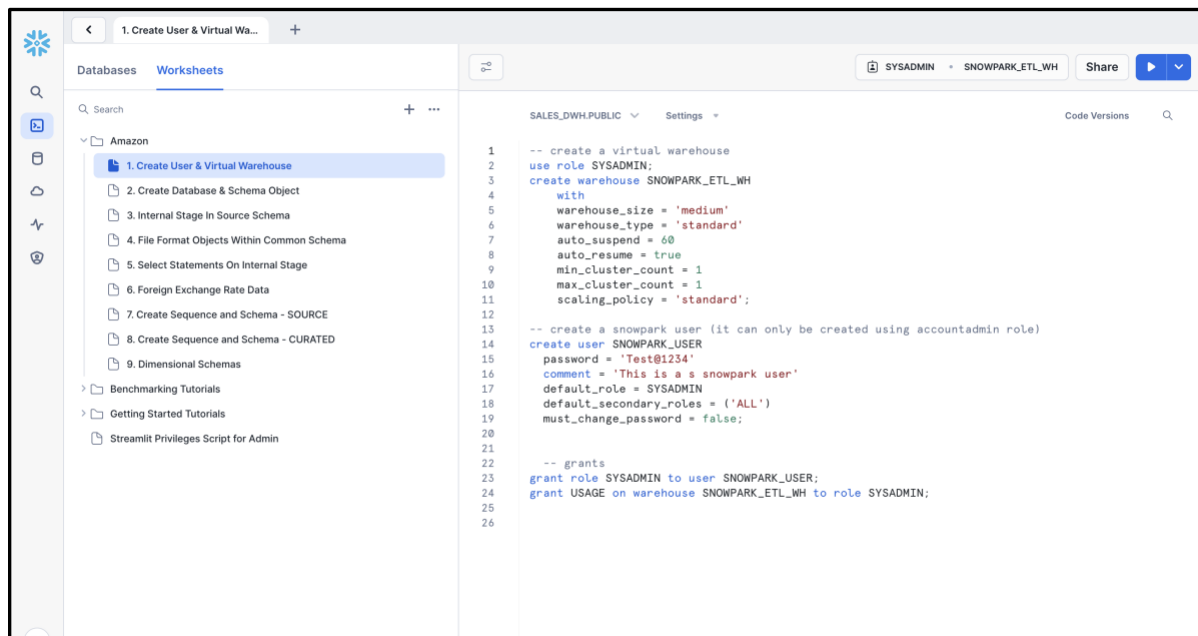
DATASET

The data encompasses Amazon's mobile phone sales data sales from three regions: India, USA, and France. The pipeline employs modern data engineering practices and tools to handle data ingestion, and transformation effectively.
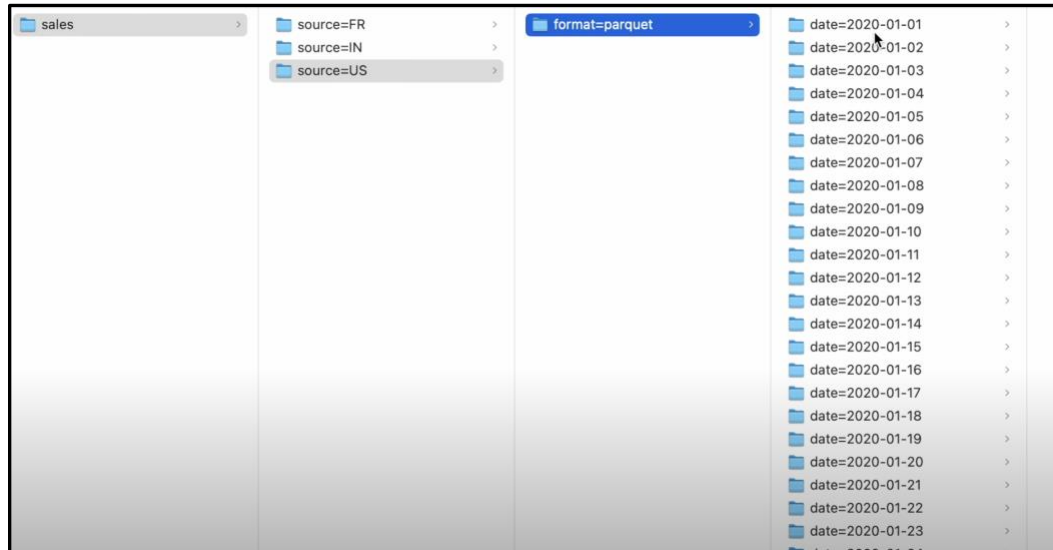
SNOWFLAKE WORKSHEETS

CONSUMPTION TABLES

CUSTOMER_DIM: Contains details about customers such as customer ID, gender, age, and other demographic information, enabling analysis based on customer profiles.

DATE_DIM: Stores information related to dates such as day, month, year, and any other relevant date attributes, facilitating time-based analysis and trend identification.

PAYMENT_DIM: Includes payment-related details such as payment method, transaction ID, and other payment-specific attributes, allowing for analysis of payment trends and preferences.

PRODUCT_DIM: Holds product-related information such as product ID, name, category, and other product attributes, facilitating analysis based on product performance and sales.

PROMO_CODE_DIM: Contains details about promotional codes such as code ID, discount percentage, validity period, and other promotional offer attributes, aiding in analyzing the effectiveness of promotions.

REGION_DIM: Stores information about regions such as region ID, name, location, and other region-specific attributes, enabling analysis based on geographical insights.

SALES_FACT: The central table that contains the core sales data, including details such as sales ID, customer ID, product ID, payment ID, date ID, promo code ID, region ID, and other relevant metrics like quantity sold and total amount, allowing for comprehensive analysis of sales performance and trends.