Methodology Followed:

Visualize Data

- To understand the structure of the data.
- The 3 given JSON files were loaded into an online json formatter to easily comprehend the data and fixed their structure.

Normalization

- Normalization divides the receipts table into smaller tables to achieve achieve SRP (Single Responsibility Principle).
- Python script was used to read .json and split into 2 csv files namely receipts_main and receipts_iems.
- Receipts_main.csv: Contains high-level receipt details such as receipt ID, date, and user ID.
- Receipts_items.csv: Contains item-level details for each receipt, such as item name, price, and quantity, linked to the Receipts_main by the receipt ID.

Data Flattening

- Flattening converts nested structures present in the JSON files to flat tables in order to create csv files.
- Converted brands.json and users.json files into csv files using excel for further cleaning and processing.
- CSV files are more convinient for manipulations and to import into SQL Server.

Data Cleaning

- To ensure data is clean, consistent and ready for database import.
- Python scripts were used to clean and format all the csv files.
- Tasks like: removing duplicates, handling null values, datetime and other datatype conversions, standardizing text fields.
- **Handling delimiters in CSV fields:** Some field values contained commas(,), slashes(/), and other special characters that conflicted with csv delimiters. To resolve this, all commas were replaced with underscores(_).

ER Diagram

- Visualize the relationships between the tables.
- Using draw.io, an Entity-Relationship diagram was created to illustrate the relationships between the Users, Brands, Receipts_main, and Receipts_items tables.
- This diagram visually represents the data model and relationships between entities.

Schema definition

- Define the structure of the table to align with the normalized schema.
- Based on the ER diagram, SQL CREATE TABLE statements were written and executed in SQL Server to define the database schema
- Specifying the columns, data types, NULL/NOT NULL constraints, and primary keys for each table.
- Defining relationships between tables e.g., foreign key constraints linking tables.

Data Ingestion

- To load the data into the predefined tables in SQL Server using the BULK INSERT command.

Data Querying

- To retrieve and analyze data using SQL queries.