TECHNICAL TEST - Data Engineer Position

- Phan Thanh Tin  
- 0818355716  
- thanhtin8642@gmail.coma

1. Schema Design

To design the data storage schema for the data set "Iowa Liquor Sales 2017 thru 2023" I use the Star Schema model:

* Dimension Tables:

**date\_dim:** date\_id, date, day, month, year, quarter, day\_of\_week

Purpose: Allows for time-based data analysis, such as sales by day, month, quarter, year.

**store\_dim:** store\_id, store\_number, store\_name, city, snapshot\_date, last\_update

Purpose: Analyze sales by store and geographic location.

**category\_dim:** category\_id, category\_name, snapshot\_date, last\_update

Purpose: Analyze sales by product type.

**vendor\_dim:** vendor\_id, vendor\_number, vendor\_name, snapshot\_date, last\_update

Purpose: Analyze sales by vendor.

* Fact Table:

**sales\_fact:** sales\_id, date\_id, store\_id, category\_id, vendor\_id, bottles\_sold

Purpose: Stores metrics such as the number of bottles sold. This table is related to the dimension tables through foreign keys.

* Ensuring Completeness:

Dimension Tables: Dimension tables (date\_dim, store\_dim, category\_dim, vendor\_dim) are designed to store descriptive attributes, allowing for analysis of data from multiple perspectives such as time, store, product type, and vendor.

Fact Table: The sales\_fact table stores key metrics such as number of bottles sold. This ensures that all information required for analysis is available.

* Performance Optimization:

Query Optimization: Star schemas help optimize data analysis queries, as dimension tables are typically small and can be quickly joined to fact tables. This reduces query time and improves performance.

* Extensibility:

Design the schema so that new dimension tables or fact tables can be easily added without affecting the existing structure.

New data can be easily added to existing tables without changing the table structure, as long as the new data follows the same format.

Using snapshot\_date, last\_date in dimension tables allows tracking changes over time, allowing for better historical analysis and data management.



2. Data issues encountered during processing and storage.

* I noticed columns with missing values ​​like: City(53231), Category Name(8751), Verdor Name(7), Vender Number(7)
* I noticed that the Date column was not in date format so I converted it to the correct format
* I noticed that the 'Bottles Sold' column value has 4761 negative values, this may not be appropriate because the number of products sold cannot be negative

Because I don't yet know the business information and requirements on the dataset, I will create a DataFrame to store error records so that I can process them later. Only load valid data into the database after handling errors.

* I have a problem with duplicate primary keys when updating data into sql. To solve the problem, I added ‘snapshot\_date’ and ‘last\_update’ columns to the dimension tables. This can help me store daily snapshot records and track changes in data.