Vending Machine Dataset Issues

1. Data Modelling - Schema

The dataset is not in a normal form, which means that some entities are not separated into their own dimension tables, but rather mixed with the fact or transaction tables. For example, the Bán hàng table contains vending machine attributes that belong to the Danh sách máy table. These attributes are: Dòng máy, Quận, Tên máy, Tên địa điểm, Tên Đường, Phường, Lat, Long, WarehouseID, WarehouseParentId, and WarehouseParentName. Moreover, the same table has product attributes such as Tên sản phẩm and Tên danh mục SP that should be part of a Product dimension table. The table seems to have a denormalized or flat structure. This kind of schema is suitable for simple calculations such as aggregate count and sum, as it does not require joining tables to aggregate. However, this schema also takes longer to load filtered data or calculate time series changes (MoM, YoY change), which reduces its efficiency in more complex calculations. Additionally, this kind of tables is not compatible with Similarly, other tables have the same problem, such as the supplier attribute (Tên NCC) in the Đơn đặt hàng and Nhập kho từ NCC tables; the warehouse attributes such as Kho xuất, Kho nhập in the Luân chuyển tới máy and Nhập kho từ NCC tables; and the [Vending Machine] Type attribute in the Đơn đặt hàng table. The dataset also lacks some columns in the dimension tables and some dimensions altogether. For instance, there is no date dimension or a product dimension table in the dataset. For the two warehouse dimension tables, there is only one ID column and no other information, which requires data transformation to obtain.

2. Redundant Columns

The dataset has data redundancy, which means that the same information is stored in multiple columns within the same table or across different tables. A clear example of this is the use of Tháng, Ngày, Giờ columns when this information is already available in the DateCreate column in the Bán hàng table. This information is also duplicated in the CreatedOn column. It is best practice to split the date and time information from a column and create date and time dimensions accordingly. Therefore, it is essential to choose one option and remove the others depending on how the data is modelled later. Another example of redundancy is the different Code and ID columns in all of the fact tables. These columns refer to the same entity and their purpose is to distinguish one transaction from another. However, in this case, it is possible that the code, in a longer format, is needed in the real-world for the paperwork involved in the delivery and receipt of goods, while the ID is used in the company's system to make the data querying more efficient.

3. Low-cardinality Columns

The dataset has some columns that have only one unique value or no value at all (NULL). These include the Note columns in all tables, the TaxVAT & Company ID columns in the Đơn đặt hàng table, and the Kho nhập column in the Nhập kho từ NCC table. A special case is the UpdatedBy, UpdatedByName, & UpdatedOn columns in the three fact tables. In a conventional database, these columns are used for ETL processes and are stored in a separate logging table to create a layer of abstraction and make the analysts' job easier. In this case, it is better to remove them during the data modelling process because they are not relevant for understanding the data itself and they increase the query load time.

4. Naming Convention

The table and column names in the dataset use both English and Vietnamese, which can cause confusion and inconsistency. This also worsens the problem of data redundancy, as some columns have different names but store the same information. Using one language for naming would make the process of removing redundancy easier. Another issue is the absence of indicators for dimension or fact tables in the table names. For example, FCT_Bánhàng and DIM_Warehouse are more informative and easier to process by analysts than Bán hàng and WarehouseID, especially if the data model follows a star or snowflake schema. Finally, the use of a single term for different types of entities can be clarified. For example, WarehouseID or RefWarehouseID can refer to either a receiving warehouse (a warehouse that receives goods from suppliers) or a machine warehouse (a warehouse that delivers goods to vending machines) in different tables. In this case, it is better to rename the entities as Original Warehouse (kho gốc) and Machine Warehouse (kho máy) and the ID columns as OriginalWarehouseID and MachineWarehouseID. This approach would prevent unnecessary confusion.

5. Dimensional Modelling Integrity

The dataset has two main issues that affect its ability to be modelled dimensionally: primary key integrity and lack of keys and details to link tables. Primary key integrity means that a key should uniquely identify one entity with consistent descriptions. However, some keys in the data have inconsistent descriptions for the same entity. For instance, in the Bán hàng table, some keys for vending machines have different locations. This makes the vending machine dimensional columns in the Bán hàng table unreliable and unusable. Therefore, these columns are removed in the modelling process and the vending machine ID column is mapped to the id máy column in the Danh sách máy table, which serves as the single source of truth for vending machines. Likewise, there are several product IDs for the same product in Bán hàng table, hence new surrogate product IDs were created for product dimension table. Moreover, the Bán hàng table has no key columns to join with other fact tables (Đơn đặt hàng, Luân chuyển tới máy, and Nhập kho từ NCC). To analyze other aspects of the data in depth, such as which suppliers have the most profitable products, the data needs more details such as product ID column and product quantities in these other fact tables.