Data Warehousing

Building and Analysing Data Warehouse Prototype for METRO Shopping Store in Pakistan

Name: Abdul Rehman

Roll No: i17-0357

Date: 26th Nov, 2021

Section: B

FAST School of Computing

National University of Computer and Emerging Sciences

Islamabad, Pakistan

2021

Project Overview

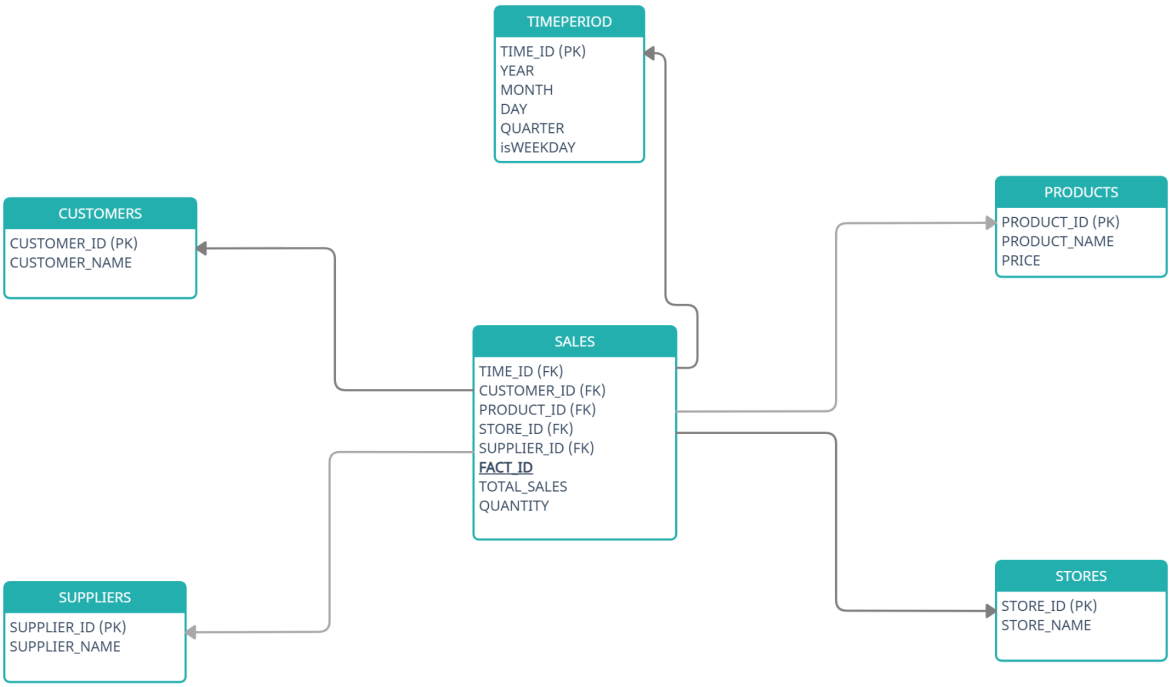
The project revolved around implementing a warehouse solution for a METRO Shopping Store in Pakistan. The main purpose of the project was to introduce the idea of real time implementation and how a certain solution works while dealing with fast-incoming data. Furthermore, the details involved in building such solutions included analysing the incoming data, choosing and building an appropriate schema for it and by implementation of Mesh join, we had to populate the schema and at the end, had to perform query analysis on it. The whole process comfortably sums up to the extent we have studied in the class and how warehouse analysis and engineering is done in the industry.

Schema for Data Warehouse

Since we were specified to create a star schema for our proposed warehouse solution, my star schema design included these dimensions,

1. Products
2. Customers
3. Suppliers
4. Stores
5. TimePeriod

The complete design of the star schema is as follow,



Mesh Join Algorithm

After reading the description of the Mesh-join approach given in the project-briefing document, I started with specific sizes for buffer, stream queues(Arraylists since they provided easier handling of inserting and deleting data). After that, I simply created a connection with the database and read transaction and master data according to their fixed size and populated the hashmap. Now, I matched the records of hashmap with the transactions inside the map and if any matched the key, I simply populated that inside the warehouse star schema according to tables. The entire process is sort of inside a loop which ends when all the records have been inserted in the data warehouse. Inside the loop,

transactions are read and added to the hashmap using a function called fillQueueAndHashMap(). After that, probIntoHash() function takes the product ids, finds them in the hashMap and if the both match, the transaction is then sent to saveTupleToDWHouse() function to insert records into the data warehouse.

Shortcomings

The shortcomings identified are,

1. Dependency between disk buffer and stream buffer which reduces the optimal disk allocation when dealing with joins.
2. Lesser record removal from hashmap in terms of large incoming data can be possible

Warehouse Queries

Query 06

The anomaly in the dataset is that if the customer buys the same product from the same store at the same time(DD/MM/YYYY) then it will create a clash based on the composite keys in the fact table which lead to a clash. So, the new record will not be inserted as the record with the same values is already in the data warehouse. To counter this, I used a surrogate key by the name SALE_ID.

Overall Learning

The main idea was to give a strong industry related experience in handling above mentioned and explained concepts. The whole process from identifying the schema implementation to the join implementation and later query processing on the warehouse data was pretty handy and practical.