Data Warehousing

CS-408

NASEER UD DIN
18I-0407
SEC “A”

PROJECT REPORT

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

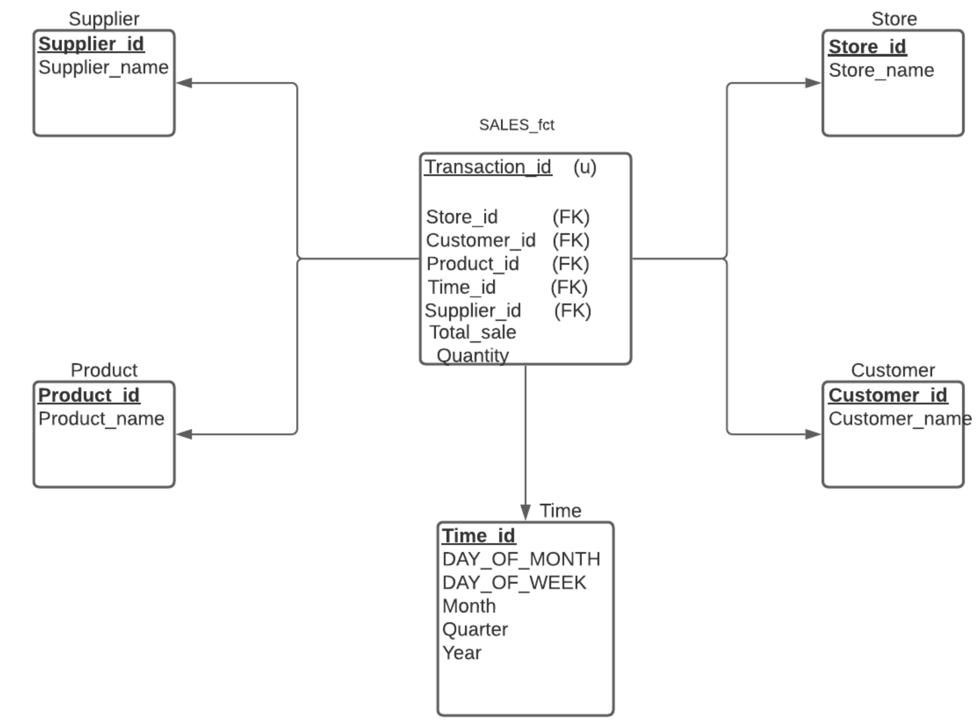
Project Overview

METRO is one of the biggest superstores chains in Pakistan. The stores has thousands of customers and therefore it is important for the store to online analyse the shopping behaviour of their customers. Based on that the store can optimise their selling techniques e.g. giving of promotions on different products.

To make this analysis of shopping behaviour practical there is a need of building a near-realtime DW and customers' transactions from Data Sources (DSs) are required to reflect into DW as soon as they appear in DSs. For this , I performed near-real time ETL using meshJoin algorithm for joining of transactional data with master data. As data from transaction data is incomplete for the desired data warehouse data structure, a complementary master data with detailed product and supplier information is used using join i.e., meshJoin.

Start schema is used to create Data warehouse of Metro store. Once all the transactional data is loaded into Metro Data warehouse which is METRO_DWH in my case, OLAP queries are performed for the analysis purpose.

Schema Diagram of METRO_DWH



MeshJoin Algorithm

The algorithm is implemented in three phases i.e., the Extraction, Transformation, and Loading.

1. **Extraction:**
   + 1st of all I read tuples from TRANSACTIONS table as an input data into the hash table with their join attribute values in the queue which is hashTableQueue in my case. Transactions data is accessed by using `getTransactionsData` function of DBHandler class.
2. **Transformation:**
   + Loading next MD partition into the disk buffer and MD partition in the disk-buffer is replaced by the new MD partition from disk using `getMasterData` function of DBHandler class.

   + Next, I look up each tuple from the disk buffer to the hash table. If matches, add the attributes which is `masterDataTuple` , into transaction tuple.
3. **Loading:**
   + The transaction tuple, that is updated in the previous step, is now loaded into METRO_DWH.

At last, after completing the look up of all disk buffer tuples into hash table, I remove the join attribute values from the last partition of the queue along with their transaction tuples from the hash table by checking against `PRODUCT_ID.`

Shortcomings in Mesh Join

1. If the master data becomes very large, MESHJOIN's performance will decrease.
2. Meshjoin reduces disk accesses to master data over a queue of stream tuples in memory.
3. The limitation of MESHJOIN is that when the number of tuples in R changes, the size of disk buffer also changes; that makes memory distribution suboptimal.

Anamoly

The date in transaction table only includes day, month and year. There is no time associated with it. So, when data is only without surrogate key there is possibility that some data is not loaded into the sales table. It would happen in a case where a customer buys same product from same store on same day more than one time.

What did I learn from the Project?

First of all, I learned from this project is that the ETL phase of Data warehousing is the most crucial part. The implementation of how the data is extracted from database, then performing transformation, and at last loading which is a time taking job at the end of ETL, in practical scenario is really helpful to understand what goes behind the ETL phase.

The algorithm that we have implemented i.e., MeshJoin though have shortcomings as discussed above but it does serve the purpose of reflecting the near real-time ETL.

To analyze the Metro data warehouse, one can now write queries to perform any kind of promotion on products or analyze the sales that too with respect to time generated by Metro, so they can understand the purchasing behavior of customers.

As far as it is concerned with the implementation of MeshJoin in Java, it was a bit challenging since I didn’t have grasp on Java. Though the libraries provided by Java are really helpful in performing the job I intended to do.

After working on this project, I now understand how warehouse is built in actual and how the ETL phase works. That will surely be beneficial in my professional career.