**Name: Rabia Mustafa**

**Roll number: 20i-1853**

**Section: DS-N**

**Course Name: Data Warehousing & Business Intelligence Semester-Fall, 2022**

**Building and Analyzing a Near-Real-Time Data Warehouse Prototype for METRO Shopping Store in Pakistan**

1. **project overview**

In this project we have designed a Warehouse for Metro which is one of the biggestsuperstores chains in Pakistan. Purpose of the Data Warehouse is to analyze the shopping Behavior of the customers.

For that purpose, a streaming Data (include transactions table) is given.

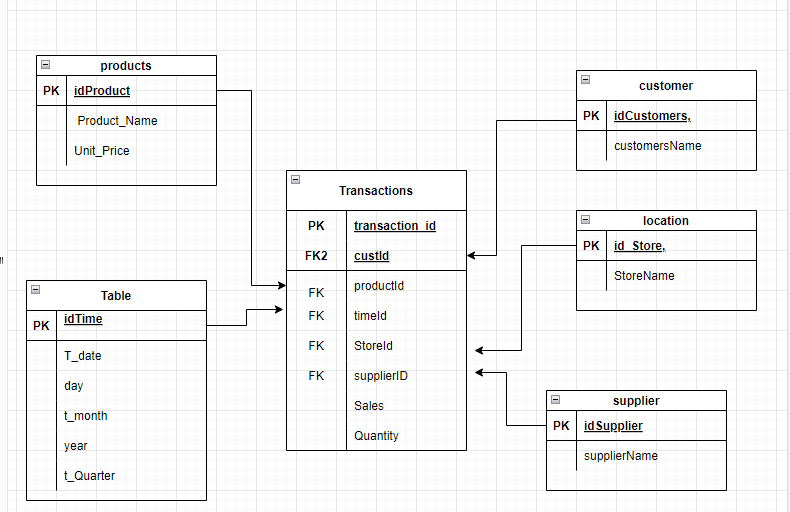
Other than that Master Data is also provides and it includes, Customer table and products table

I have implemented Mesh Join Algorithm to Send the data in the Data Warehouse

ETL (Monitoring, Extraction, Transformation, Loading) phase is also implemented before sending data into mesh join.

The algorithm is implemented in Java SE 8 and can be run using the Eclipse IDE.

1. **create a star-schema for DW**



1. **MESHJOIN algorithm**

**Algorithm** Mesh Join

**Input:** Transactions Table, Master Data (Customer table and product table)

**Output:** Data Warehouse is created

**Method:**

Number of rows are calculated for Transactions Table, Master Data

Then the partition calculation is done

For loop on transactions Data

Read first partition from transactions table

Add partition pointer in queue

add the partition in the HashMap

Load the Master table data

Compare Transactions Table Data with Master Table Data

If tuples match {

Send the data in Data Warehouse

}

Deque the Queue

Empty the HashMap after each comparison

// this process will continue until the transaction data is loaded in DWH

1. **Any three shortcomings in Mesh Join**

**Here are some of the short comings of mesh join**

**LOWER EFFICIENCY:**

During implementation I observed that data must be fetched from the hard drive repeatedly and we don’t use cache to store common data temporarily, it is affecting Efficiency

**FREQUENTLY SOLD PRODUCTS:**

Some products are sold more frequently. For example, in a store egg milk and butter are stored at highest rate, so ideally, we should store the master data related to egg milk and butter such data into memory, so then it can reduce disk I/O operation so there will be less chance of delay in sending to data to DWH.

**SEQUENTIALLY LOADING FROM DISK:**

In Mesh join Master data is loaded sequentially from the disk for every partition of the transactions table So there might be a chance thatchunk of tuples that is loaded from master data do not join with any transaction table tuples, so we should apply algorithm such that data is not loaded sequentially, and we should load only useful part from the disk

**F) what did you learn from the project?**

1)implementation of Mesh Join algorithm

2)How to transform the data before enrichment of data in DWH

3) How to analyze data in DWH and How to perform operations such as ROLLUP, Drilldown etc.

4) How we can create materialized view in a data warehouse

5)It was a good practice of implementation real-time data warehouses and it make me clear about a lot of concepts of DWH.