## **DATA WAREHOUSE PROJECT REPORT**

**PROJECT REPORT**

**Name : Nooran Ishtiaq**

**Roll no : 22i-2010**

**Section :DS-B**

**PROJECT OVERVIEW :**

The goal of this project is to design the prototype of metro sales near real-time Datawarehouse .The project aims to extract stream of transactional data , transform it and store it in the data warehouse using **MeshJoin Algorithm.** The working of this project involves creating star-schema for this near-real time data-warehouse , applying meshjoin and then loading the transformed data into fact table and dimensions and also support the working of OLAP queries on the datawarehouse .

**SCHEMA FOR DATAWAREHOUSE :**

Star-schema has been for storage in data warehouse.

**Fact table** :

Metro\_Sales\_Fact

* ORDER\_ID
* ORDER\_DATE
* TIME\_ID
* PRODUCT\_ID
* CUSTOMER\_ID
* QUANTITY
* SALE

**Dimensions :**

TIME\_DIMENSION

* TIME\_ID
* Orderdate
* day
* month
* quarter
* year
* week
* day\_of\_week

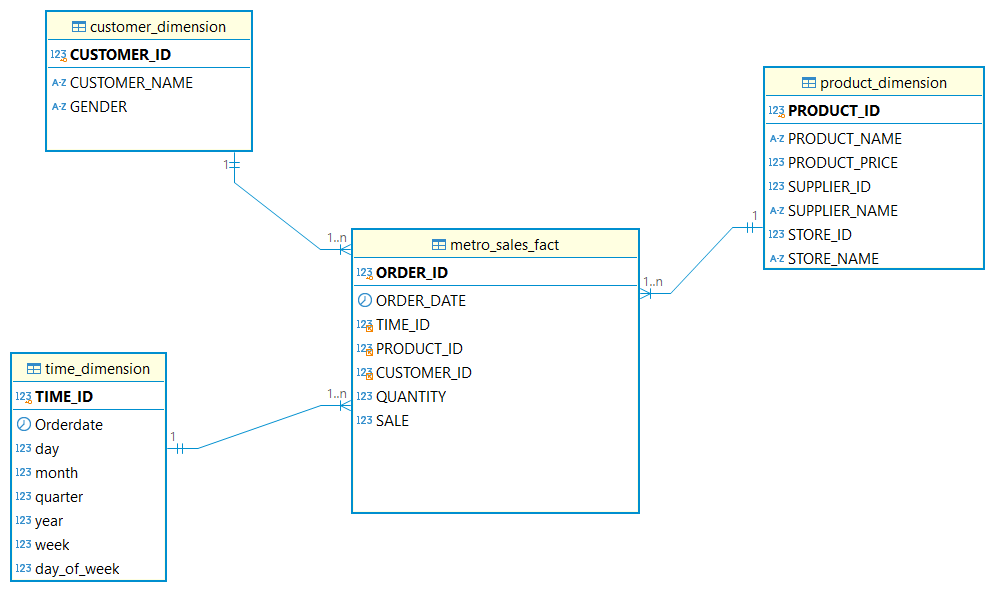
PRODUCT\_DIMENSION

* PRODUCT\_ID
* PRODUCT\_NAME
* PRODUCT\_PRICE
* SUPPLIER\_ID
* SUPPLIER\_NAME
* STORE\_ID
* STORE\_NAME

CUSTOMER\_DIMENSION(

* CUSTOMER\_ID
* CUSTOMER\_NAME
* GENDER

**Schema DIAGRAM:**

****

**Technologies Used:**

**Database:** MySQL

**Framework:** Java-based parallel processing for efficient working

**Algorithm Implemented** : MESHJOIN ALGORITHM

**Data Processing:** Partitioning, parallel processing, stream buffer , disk buffer and stream queue and hash table.

**MESHJOIN ALGORITHM :**

Meshjoin algorithm helps in optimizing the join of larger datasets from multiple tables.The meshjoin algorithm reduces the overall computational load by loading data in chunks and then apply join parallelly.

Steps in Meshjoin:

Step 1: Partitioning of transactional data and storing it in stream buffer.

Step 2: Stream buffer will extract foreign keys from specific chunk size and load it into the queue.

Step 3: The queue will keep track either the chunk has matched with all the partitions from both master tables or not .

Step 4: The foreign keys are also stored in the hash table keys .The values of hash table contain remaining data .

Step 5: The join will be performed by loading both master data (product and customer ) partitions from the disk buffer cyclically .If the join is successful that is the case when the keys in the hash table are matched with the keys in the disk buffer , the additional information will be enriched into the hash table values.

Step 6: After that , the transformed data will be loaded Into the the fact table and dimensions in the database .

**THREE SHORT COMINGS:**

1. Meshjoin algorithm has high memory consumption when dealing with larger datasets because it uses buffers and hashtables to store temporary results that can be inefficient.
2. The size of larger datasets slows down the performance of algorithm. It can take a lot of time in transformation , the scenario becomes worst when we have to deal with data update .
3. The unbalancing of data in the partitions can decrease the processing time.

**Learning From Project:**

The project enhanced by knowlwdge in datawarehouse field on how to handle large stream of data , process it and perform analytical queries in real-time.

This project helped me on how to manage the challenges ofmanaging large data i.e memory management and performance optimization.

I have learned how to use data warehouse to perform OLAP queries and derive meaningful insights from it.

To conclusion, I have gained a lot of knowledge from this project how can we make datawarehouse and perfrom olap queries in real time.