Data Warehousing & Business Intelligence

Project Report

FALL 2022

Name: Abdullah Basharat

Roll No: 20i-2446

Submitted to: Dr Asif Naeem

# Overview

In this project, we made a warehouse of a renowned retail store ‘Metro – Cash and Carry’ through ETL process. It demands certain OLAP queries which require a Warehouse. To Implement a warehouse, we were given a Sample Data Stream of transactions which are done at the retail store and two tables of master data. We had to perform its MESH-JOIN with Master Data and send all the data to warehouse.

The Mesh-Join part was implemented on Eclipse through JAVA.

The Warehouse was made using SQL Workbench

# Schema

We have implemented a **Star Schema** for Data Warehouse

## Dimension Tables:

There are 5 Dimension Tables:

1. Customers

* **Customers** (CUSTOMER\_ID, CUSTOMER\_NAME)

1. Products

* **Products** (PRODUCT\_ID, PRODUCT\_NAME)

1. Stores

* **Stores** (STORE\_ID, STORE\_NAME)

1. Suppliers

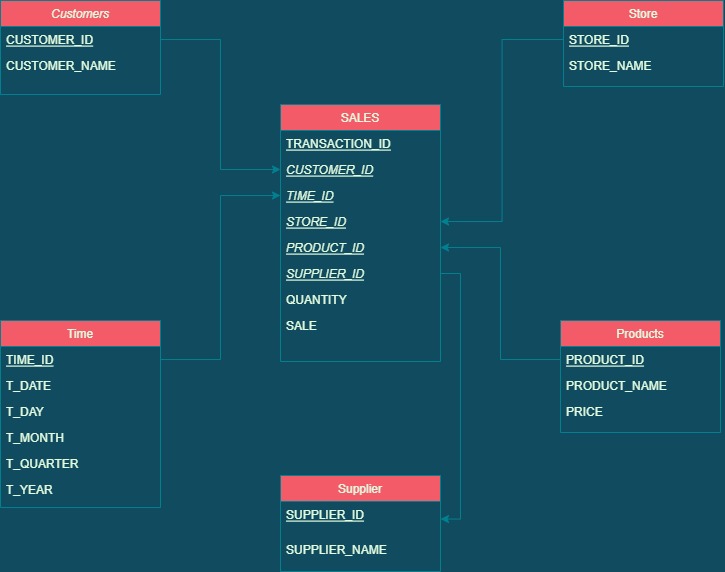
* **Supplier** (SUPPLIER\_ID, SUPPLIER\_NAME)

1. Time

* **Time** (TIME\_ID, TIME\_DATE, T\_DAY, T\_MONTH, T\_QUARTER, T\_YEAR)

1. Main Fact Table:

**Transactions** (TRANSACTION\_ID, *PRODUCT\_ID*, *CUSTOMER\_ID*, *TIME\_ID*, *STORE\_ID*, *SUPPLIER\_ID*, QUANTITY, SALE,)



# MESHJOIN Algorithm

Get partitions

For ( i=0 to No of Partitions of Stream/Transactional Data)

{

Map <List<String>, List<String> > hm = new HashMap //List of Key and Value Pairs

S <- Partition [i] // Add data of Partition [i] to HashMap

M <-ReadMasterData1() //Add Master Data

While true:

if (Partition\_Tuple.Key in M)

Add M’s values (or joinable attribute) into HashMap

Thread insertion (HashMap)

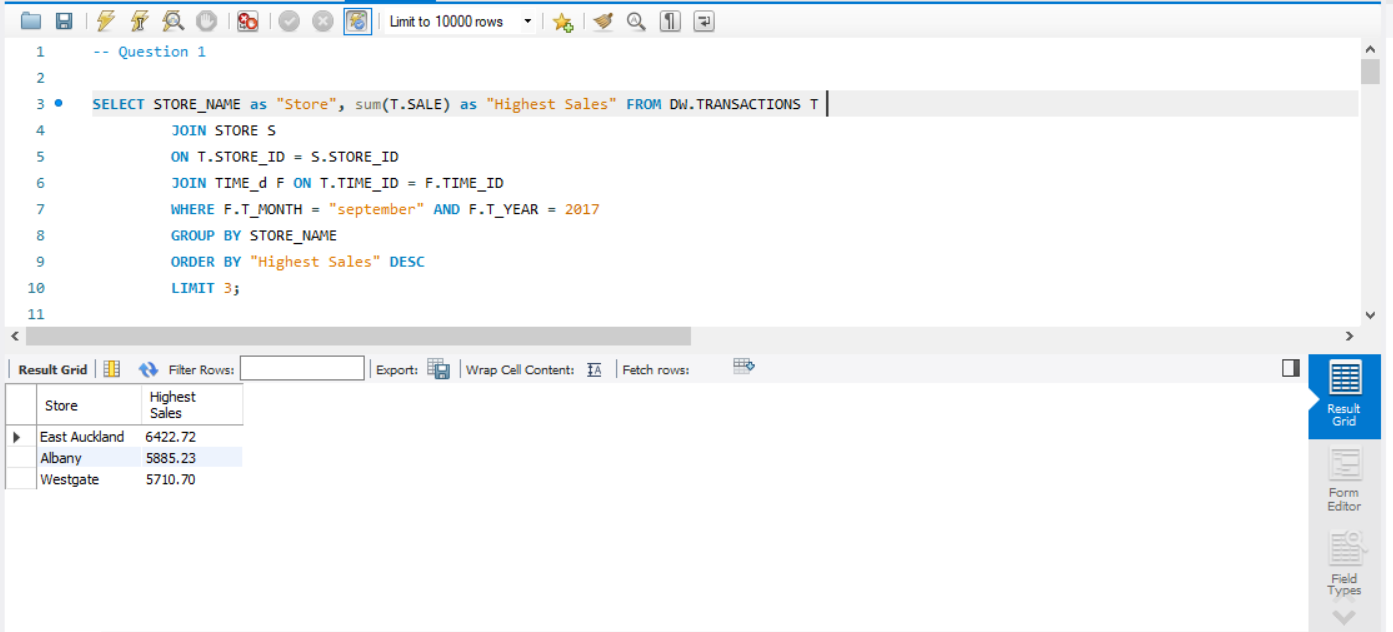
}

# Short Comings of Mesh Join

* The Retail Data is not normally distributed among the Master data tables. So, if there’s any way, we know which product has more sales so during Mesh Join Comparisons we know which product is going to have more frequent matching we can reduce the I/O operation which reduces the chances of delay in streaming.
* All the Master Data is loaded sequentially and there’s high probability that tuple loaded from Master Data does not join with any stream tuple present in HashMap so a mechanism should be in place so Only useful part of Master Data is loaded from disk and checked.

# Outputs:

## Q1



## Q2

**Explanation:** We can forecast the next week supplier by analyzing the top suppliers of previous weekends all over the month or years.

Graphical user interface, text, application

Description automatically generated

## Q3

Graphical user interface, text, application

Description automatically generated

## Q4

Graphical user interface, text, application

Description automatically generated

## Q5

Graphical user interface, text, application

Description automatically generated

## Q6

Graphical user interface, text, application

Description automatically generated

## Q7

**Explanation:** The query gives the total sales considering all the scenarios; each store, supplier and product. The Roll up gives Total Sale with respect to

* Each Store
* Each Store and Supplier
* Each Store, Supplier and Product

Graphical user interface, text, application, email

Description automatically generated

## Q9

Graphical user interface, text, application

Description automatically generated

## Q10

Graphical user interface, text, application

Description automatically generated

# What did you learn from this project?

I have learned about how a Fast Data Stream can be transformed into useful Data Warehouse. This project has enhanced my understanding and concepts and got the knowledge about underlying challenges in Near Real time data Streams.