

**Title:** Project Report

**Name:** Umm-e-Hani

**Roll No.:** 21I-1715

**Section:** M

**Submitted to:**

Dr. Asif Naeem

# 

# **Project Overview:**

In this project, our goal was to implement a near real time data warehouse that will get data from multiple sources and process it to store it in a dynamic warehouse. We used HYBRID JOIN Algorithm to implement ETL and process data before putting it in the electronia-dw.

OLAP queries are applied and analysis is performed.

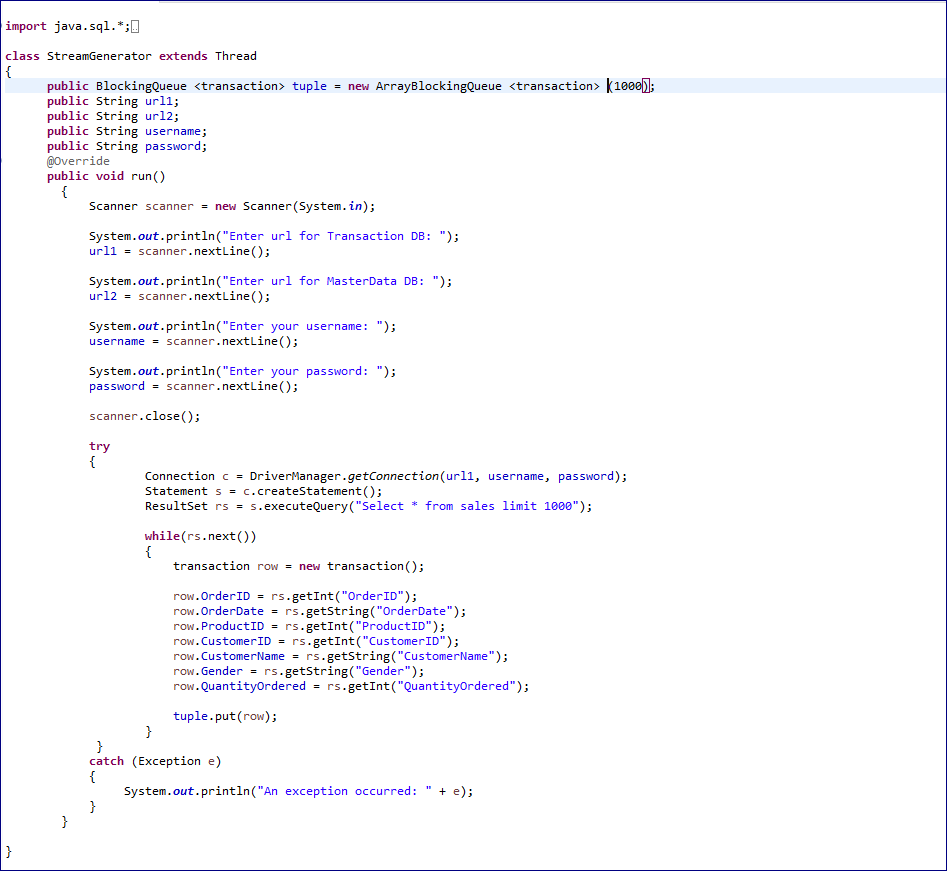
# **Star schema:**

Star schema is used to design the Electronica data warehouse inorder to facilitate the storage and analysis of sales-related data. It allows for detailed examination of sales transactions from different perspectives, such as customer behavior, product performance, supplier engagement, and temporal patterns, enabling data-driven decision-making and business optimization within Electronica's operations.

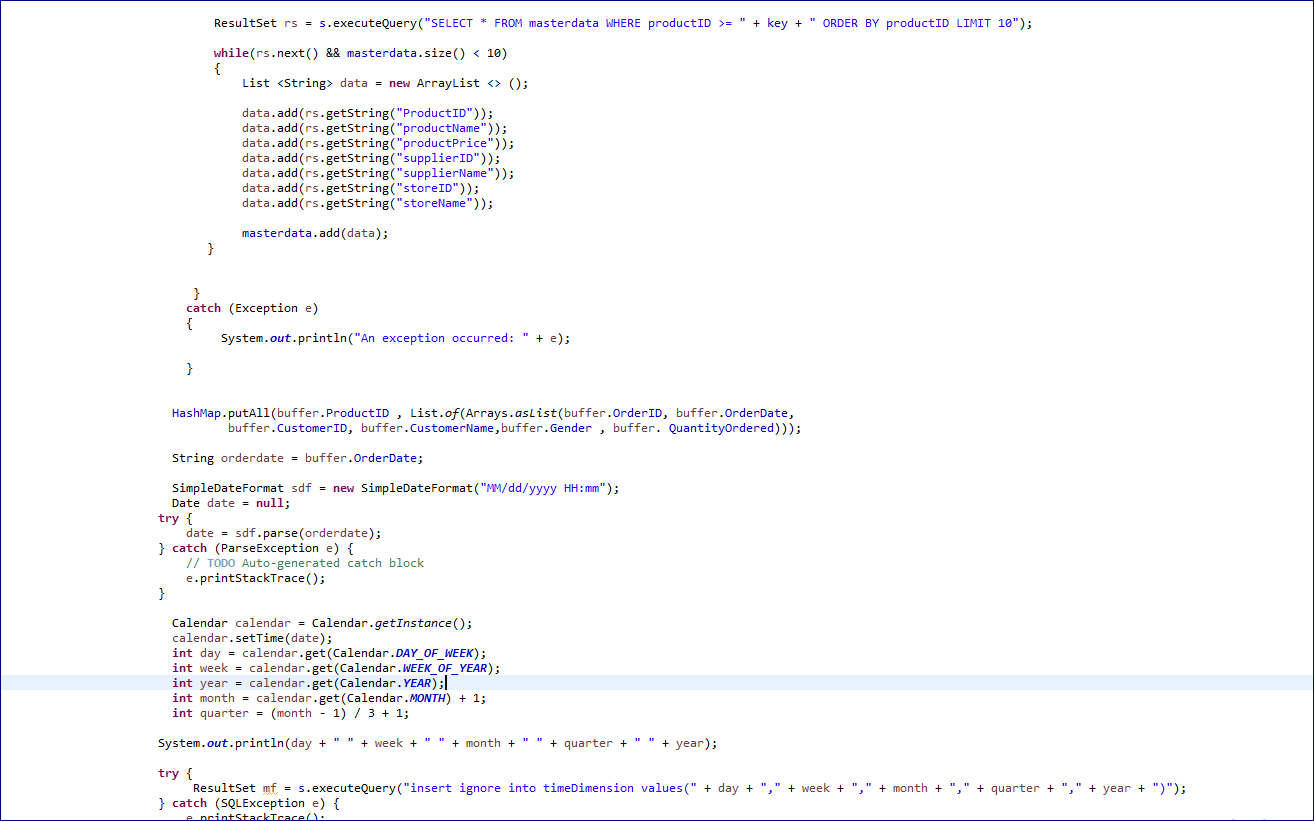
# 

# **Hybrid Join:**

I have created a streamGenertor thread that takes input of the url, username and password. Then it generates a stream of transaction data storing it in an ArrayBlockingQueue of size 1000 then passing it onto the next thread hybrid join. In the HybridJoin thread the streamGenerator thread is started and the incoming stream is stored in a buffer the product ids from the buffer are inserted into a queue and also in a hashmap containing productID’s as keys and rest of the tuple as value. The masterData is fetched and stored in a buffer that can store approximately 10 rows then the values are compared and matched and inserted into HashMap. I couldn’t do more than this as I faced a lot of trouble inserting the data into the data warehouse.



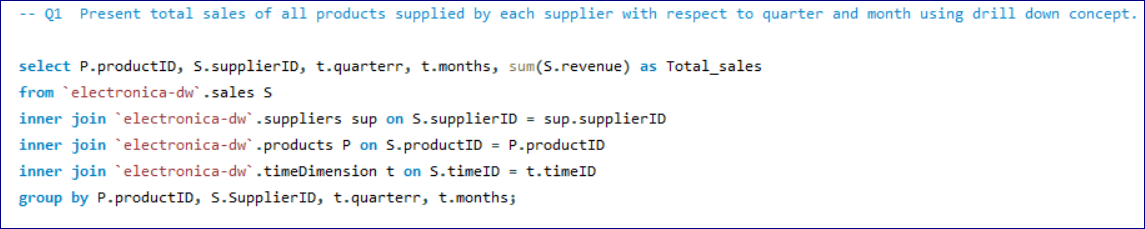
# 





# **OLAP QUERIES:**

**1) Present total sales of all products supplied by each supplier with respect to quarter and month using drill down concept.**

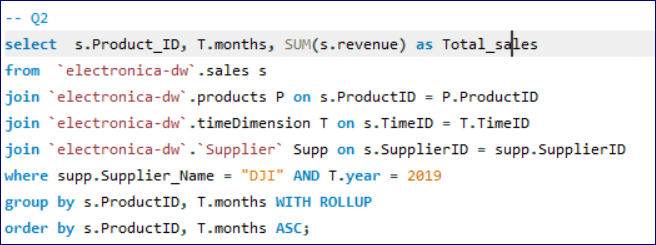
****

**2) Find total sales of product with respect to month using feature of rollup on month and**

**feature of dicing on supplier with name "DJI" and Year as "2019". You will use the**

**grouping sets feature to achieve rollup. Your output should be sequentially ordered**

**according to product and month.**

****

**3) Find the 5 most popular products sold over the weekends.**

**4) Present the quarterly sales of each product for 2019 along with its total yearly sales.**

**Note: each quarter sale must be a column and yearly sale as well as Order result according to product.**

**5) Find an anomaly in the data warehouse dataset. write a query to show the anomaly and explain the anomaly in your project report.**

**6) Create a materialized view with the name “STOREANALYSIS\_MV” that presents the**

**product-wise sales analysis for each store.**

**7) Use the concept of Slicing to calculate the total sales for the store “Tech Haven”and product combination over the months.**

**8) Create a materialized view named "SUPPLIER\_PERFORMANCE\_MV" that presents the monthly performance of each supplier.**

**9) Identify the top 5 customers with the highest total sales in 2019, considering the number of unique products they purchased.**

**10) Create a materialized view named "CUSTOMER\_STORE\_SALES\_MV" that presents the monthly sales analysis for each store and then customers wise.**

# **Shortcomings of Hybrid Join:**

* Following are the shortcomings according to me:
* **High Memory Usage:** HYBRID JOIN needs a lot of memory, especially with large data. This can be a problem if there's not enough memory available for processing.
* **Slower Performance with Disk Access:** Even though HYBRIDJOIN is optimized for fast memory operations, it still relies on reading data from the disk. This might slow things down, especially if the disk is slow or handling a lot of data at once.

# 

# **Learning Out Come of the Project:**

Although the project was quite tough and I couldn't complete all of it, I did learn a lot from it The project highlighted how I can effectively and efficiently process TBs of data in a near real time warehouse.I learned the hybrid join algorithm to some extent as i couldn’t complete all of it I had trouble inserting the data in warehouse apart from that i also learned how it makes it manageable to execute business related OLAP analysis when using a warehouse rather than a database.