Report on the Data Warehouse Project

1. Introduction

Developing a robust data warehouse system for Metro Pakistan using MySQL and Java was the idea behind this project. Master data tables were to be developed with the capability to stage raw data into transformation. MESHJOIN algorithm support to enable data integration, star schema application for OLAP optimization, and finally analytical queries for insights would be executed.

---

2. Data Preparation and Master Data Creation

2.1 Database and Tables Setup

Starting from the basic tables and MetroDW database, lay down the ground of structuring product and customer data. The three tables in basic structure are as follows:

1. **Products Table**: Stored product information that may include the id, name, price, supplier, store, etc.
2. **Customers Table**: Managed customer documentation, which covered the identification number, name, and gender.
3. **Staging Table**: It created a staging table, that is, products_staging to preprocess raw product data received in CSV files. Thus, this staging table allowed some cleansing and transformation steps like removing currency symbols and altering data types to be got inserted into the master products table.

2.2 Data Loading and Transformation

Raw data imported from CSV files into staging and master tables. It was cleansed and transformed in the process of insertion.

- Prices that were in string format have been represented in numeric format.
- The anomalies in the source data were treated for maintaining consistency.

Once the data was uploaded to the primary table, the staging table was deleted to optimize the schema.

---

3. Star Schema Design

It designed a star schema to enable effective OLAP analysis. This schema was basically central to a fact table, surrounded by dimension tables, and it had transaction information recorded in the fact table by incorporating sales data whereas entities-customer, product, supplier, and store-made up the dimension tables.

Major characteristics of the star schema:

- **Fact Table**: The order details along with quantities and overall sales will be stored.
- **Dimension Tables**: Provided descriptive attributes for filtering and aggregation of data.

Indexes were created on key columns to improve query performance,
enabling relatively quicker data access in complex computations.

---

4. Data Integration Using the MESHJOIN Algorithm

For inputs of data coming from the source, the MESHJOIN algorithm was applied with implementation in Java, this will easily scale up for large dataset processing when broken up into manageable batches.

4.1 Workflow of MESHJOIN

1. **Partitions Loading**:
    - Product and customer tables were loaded in memory as discrete records.
    - Transactions fed from the extract file in batches.
2. **Join Logic**:
    - All the transactions have been enriched by aligning them with relevant data acquired both from customer and product
    partitions, through common identifiers such as customer_id and ProductID.
    - Supplementary characteristics such as the customer's name, gender, product pricing, and supplier details were incorporated into the transaction.
    - Then, total sales were calculated by multiplying the product price with the quantity ordered.
3. **Data Writing**:
    - To enhance the performance, enriched records were input in batches into the CombinedData table systematically.
4. **Resource Management**:
    - This would ensure proper utilization of memory by flushing the completed batches while it s ettled other transactions before close connections.

It reduced the memory consumption when working with large datasets and ensured data
consistency when joined.

---

5. Transition to Star Schema

The data of the CombinedData table was then loaded into the star_schema_transactions fact through the MESHJOIN procedure. It was the last step to this integration, and the data had been structured for analytical processing.

6. OLAP Analysis

The advanced OLAP queries used the star schema to give actionable insights. The analytical tasks included:

- ***Sales Performance***: Adding up total sales by product, supplier, or store.
- ***Consumer Analysis***: Ascertaining purchase characteristics in line with demographics, including customer-wise and gender-wise spending patterns.
- ***Time Analysis***: The sales trends over time, such as monthly or yearly trends.

These queries used indexed schema to ensure computation was fast. This made the system suitable for dynamic and ad-hoc reporting.

7. Project Achievements

It successfully showcased the entire process that takes place in a data warehouse system:

1. 1. Staging and transformation for the integration of multiple data.
2. Scalable Data Merging with the MESHJOIN Algorithm 2.
3. 3. Star schema Implementations with Optimized Analytical Processing.
4. 4. Business intelligence through efficient OLAP query execution.

8. Conclusion

The project focused on the integration of database management and algorithmic efficiency into a high-performance data warehouse design. The star schema, coupled with OLAP capabilities, can do well in extracting some meaningful insights related to business and thus supports decision-making processes. Integrating real-time data pipelines along with the ability of analysis towards more dimensions or metrics might be the future enhancements.