WAREHOUSE

DATAWAREHOUSE
PROJECT

26TH NOVEMBER 2024

PROJECT OVERVIEW

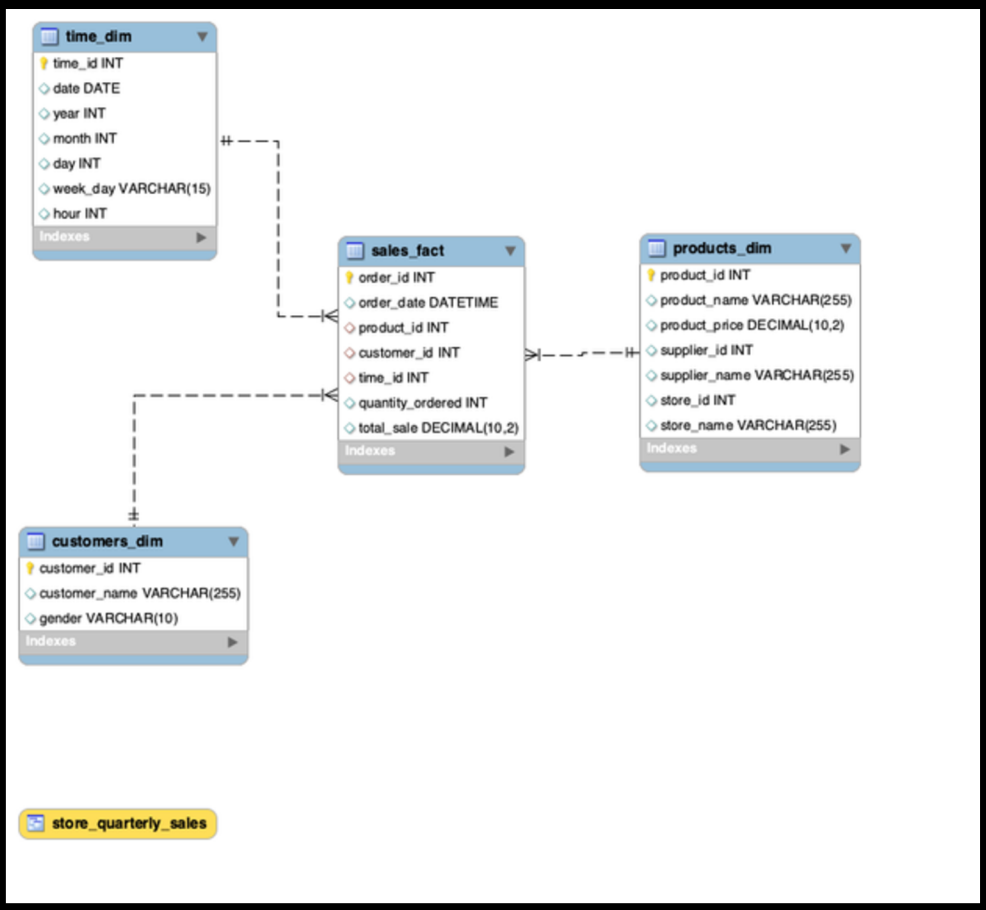
This project aims to develop a near-real-time Data Warehouse (DW) for METRO Shopping Store in Pakistan to support real-time analysis of customer shopping behavior and optimize sales strategies.

The system integrates transactional data with master data (e.g., products and customers) using a customized MESHJOIN algorithm, enabling enriched data analysis.

The DW facilitates slicing, dicing, and OLAP-based insights, such as identifying top-performing products, trend analysis, and sales forecasting.

SCHEMA

The DW is modeled as a star schema, optimized for querying and analytical tasks.



MESHJOIN

The MESHJOIN algorithm was implemented to handle the ETL process for integrating transactional streams with master data. This ensures near-real-time data enrichment and loading into the DW.

1. **Stream Extraction:** Read a segment of transactions (TRANSACTIONS table) into memory (hash table).
2. **Master Data Loading:** Cyclically load partitions of master data (products_dim and customers_dim) into a disk buffer.
3. **Joining Data:** Perform a stream-relation join, enriching transactions with details like product price, customer name, and store name.
4. **Loading to DW:** Insert enriched data into the sales_fact table, ensuring no duplicate records.

SHORTCOMINGS

Performance Bottlenecks:
Loading large master data partitions cyclically can slow the process, especially for high-frequency transactions.

Memory Overhead:
Maintaining hash tables and queues in memory increases resource usage, limiting scalability for very large datasets.

Latency in Real-Time Processing:
The chunk-based processing introduces slight delays, making it “near-real-time” rather than true real-time.

LEARNINGS

This project offered valuable insights into data warehousing and business intelligence practices:

Star Schema Design:
Learned to model multidimensional data efficiently, ensuring query performance and data integrity.

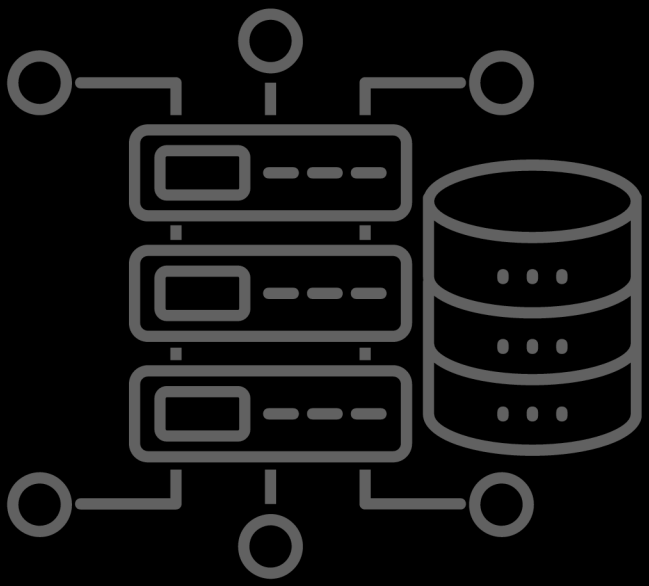
ETL Processes with MESHJOIN:
Gained experience in implementing and extending stream-relation joins to enrich transactional data dynamically.

OLAP Analysis:
Developed SQL queries for trend analysis, product affinity insights, and identifying sales outliers, demonstrating the power of DW in decision-making.

Practical Challenges:
Understood the trade-offs between performance, scalability, and real-time capabilities in ETL systems.

CONCLUSION

This project successfully developed a near-real-time DW prototype for METRO Shopping Store. By leveraging a well-designed star schema, the MESHJOIN algorithm, and OLAP analysis, the system provides actionable insights to optimize business operations.



DATAWAREHOUSE