Data Warehousing & Business Intelligence

PROJECT

Building and Analyzing a Near-Real-Time Data Warehouse Prototype for METRO Shopping Store in Pakistan

Tashfeen Abbasi
(22I-2041)

Section – D

Date - 11/12/2024
Fall 2024

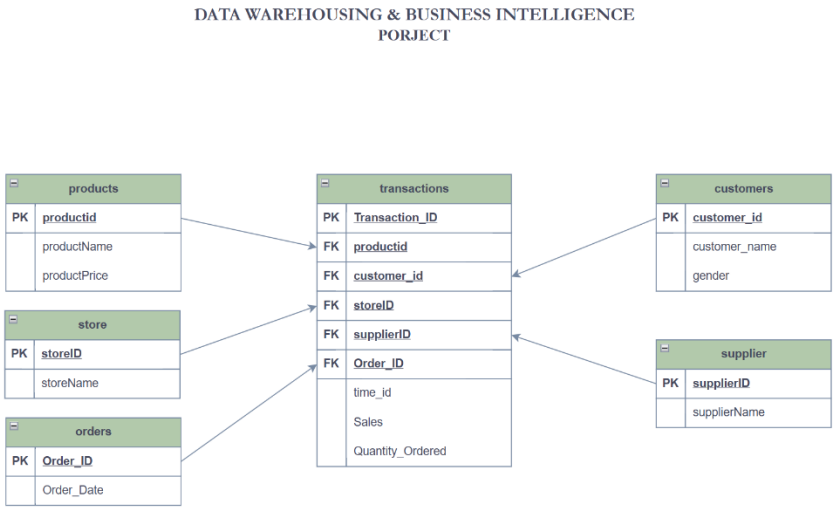
Department of Computer Science

1. Project Overview:

The project makes an improvement in loading of the transactional data into a data warehouse (DW) through the Mesh Join algorithm. The master data base is called tashi_db which is used to store the data after enhancing the data field. The schema namely tashi_dw is used to keep the data after merging of the preprocessed data. Unlike normal data warehousing where extraction, transformation and loading processes are integrated in the data processing mechanism, this organization uses ETL operations to integrate the streaming transactional data with the master data on which the analytic processes are built.

---

2. Schema:



In the current study, the data warehouse schema (tashi_dw) is proposed for analytical applications and comprises the following tables:

- **Customers**:
    - Columns: customer_id, customer_name, gender
    - Stores customer details.
- **Products**:

    - Columns: productid, productName, productPrice
    - Contains product information.
- **Store**:
    - Columns: storeID, storeName
    - Includes details of stores.
- **Supplier**:
    - Columns: supplierID, supplierName
    - Holds supplier details.
- **Orders**:
    - Columns: Order_ID, Order_Date
    - Stores order information.
- **Transactions (Fact Table)**:
    - Columns: Transaction_ID, productid, customer_id, time_id, storeID, supplierID, Order_ID, Sales, Quantity_Ordered
    - Stores transactional data in concert with foreign key references to a dimension table.

3. Mesh Join Algorithm

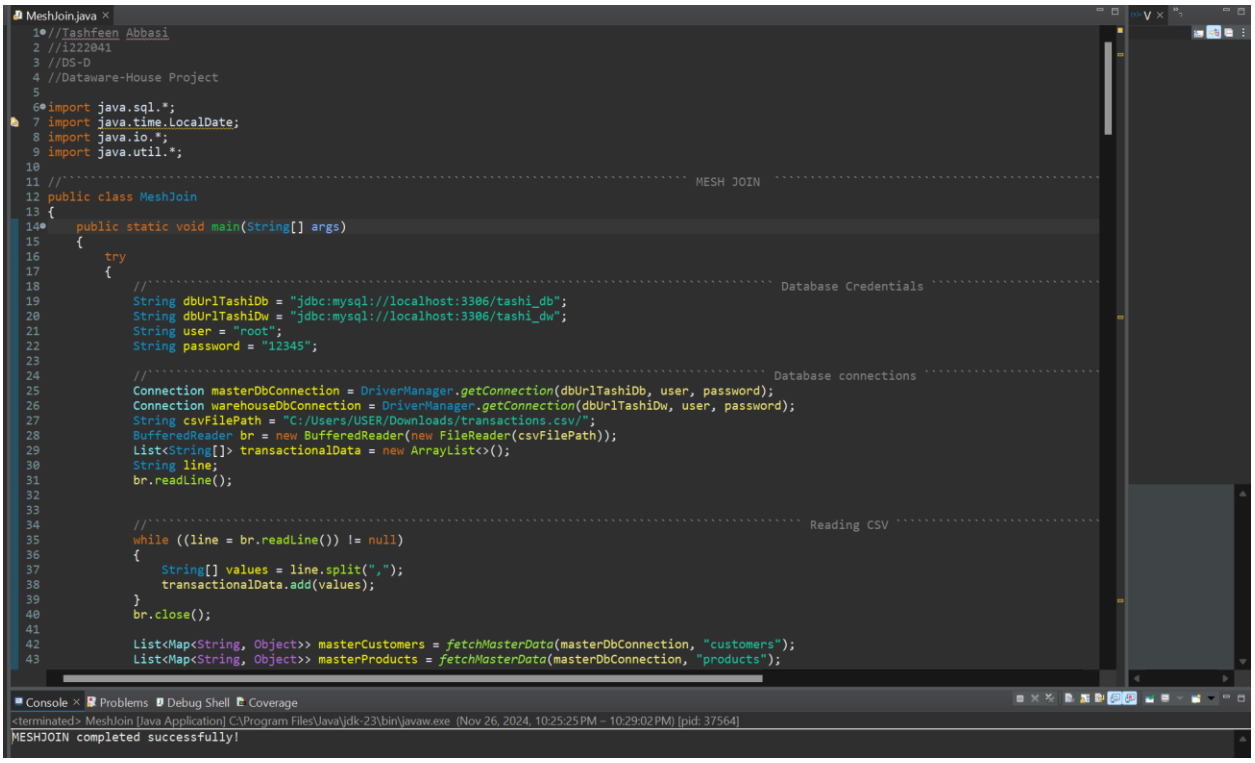
The Mesh Join algorithm allows for a fast join of streaming transactional data with master data in memory for further data enhancement and populating the data warehouse. The following are the steps of practice.

1. **Extract Data**:
    - Reads transactional data from a CSV format.
    - Load master data (tashi_db) from the database into memory.
2. **Match and Enrich**:
    - In each transaction, join productid and customer_id with the respective master data to enhance transaction.
3. **Compute Sales**:

- Derived the sales values:

  Sales = Product Price × Quantity Ordered.

4. **Validation and Deduplication**:
   - Make certain that all the foreign keys exist in the data warehouse.
   - Include only adding new records so that there will be no duplication of records.
5. **Load Data**:
   - Load more informative data into the data warehouse tables.



4. Shortcomings of Mesh Join

1) Memory Usage:

- The algorithm involves using the master data whereby in the case of very large data sets it will enhance a memory overload.

2) Performance Bottlenecks:

- When handling large streams there is a possible added stream processing latency because of lookups and data augmentation.

3) Data Inconsistencies:

- If master data is not in synchronization with real-time updates, then and only then the algorithm may give out bad results.

5. What I Learned from the Project

1) Data Integration:

- Acquired a working knowledge in integrating real-time processing of transactional data with the master data used for business intelligence.

2) Algorithm Design:

- Got to know about how Mesh Join algorithm works and its use in data warehousing projects.

3) ETL Pipeline Optimization:

- About approaches for improving ETL performance and reducing latency.

4) SQL and Query Writing:

- Deeper understanding of schema design, SQL query and the different ways to analyze data to derive meaning.