Project Report

Building and Analyzing a Near-Real-Time Data Warehouse

Name: Moiz Tanvir

Roll no: 22I-1932

Course: Data Warehouse and BI

Submitted To: Sir Asif Naeem

Project Overview

This project aimed to develop a near-real-time Data Warehouse (DW) prototype for METRO Shopping Store in Pakistan. The DW analyzes customer shopping behavior to enable data-driven decision-making. Key features include:

- Integration of transactional and master data using the **MESHJOIN algorithm**.
- A **star schema** to support OLAP (Online Analytical Processing) queries.
- Implementation of insightful queries to analyze trends, product performance, and supplier contributions.

1. Schema for Data Warehouse

Identified Tables and Attributes

1. **Fact Table**:

    - **FACT_TRANSACTIONS:** Stores transactional data and computed sales metrics.
    - **Attributes:** ORDER_ID, ORDER_DATE, PRODUCT_ID, CUSTOMER_ID, CUSTOMER_NAME, GENDER, PRODUCT_NAME, PRODUCT_PRICE, SUPPLIER_ID, SUPPLIER_NAME, STORE_ID, STORE_NAME, QUANTITY, SALE.

2. **Dimension Tables**:

    - **CUSTOMERS**: Contains customer details.

      Attributes: CUSTOMER_ID, CUSTOMER_NAME, GENDER.

- **PRODUCTS**: Contains product information.

  Attributes: PRODUCT_ID, PRODUCT_NAME, PRODUCT_PRICE, SUPPLIER_ID, SUPPLIER_NAME, STORE_ID, STORE_NAME.

- **TRANSACTIONS**: Contains product information.

  Attributes: ORDER_ID, PRODUCT_ID, CUSTOMER_ID, QUANTITY, ORDER_DATE.

Star Schema

The star schema is designed to facilitate multidimensional analysis. The **FACT_TRANSACTIONS** table serves as the central fact table linked to dimension tables via foreign keys. This ensures efficient querying.

Primary and Foreign Keys:

- Fact Table:
  - Primary Key: ORDER_ID.
  - Foreign Keys: CUSTOMER_ID, PRODUCT_ID.
- Dimensions:
  - Primary Keys: CUSTOMER_ID (CUSTOMERS), PRODUCT_ID (PRODUCTS).

2. Implementation of MESHJOIN Algorithm

The MESHJOIN algorithm was implemented in Java using Eclipse. Key steps include:

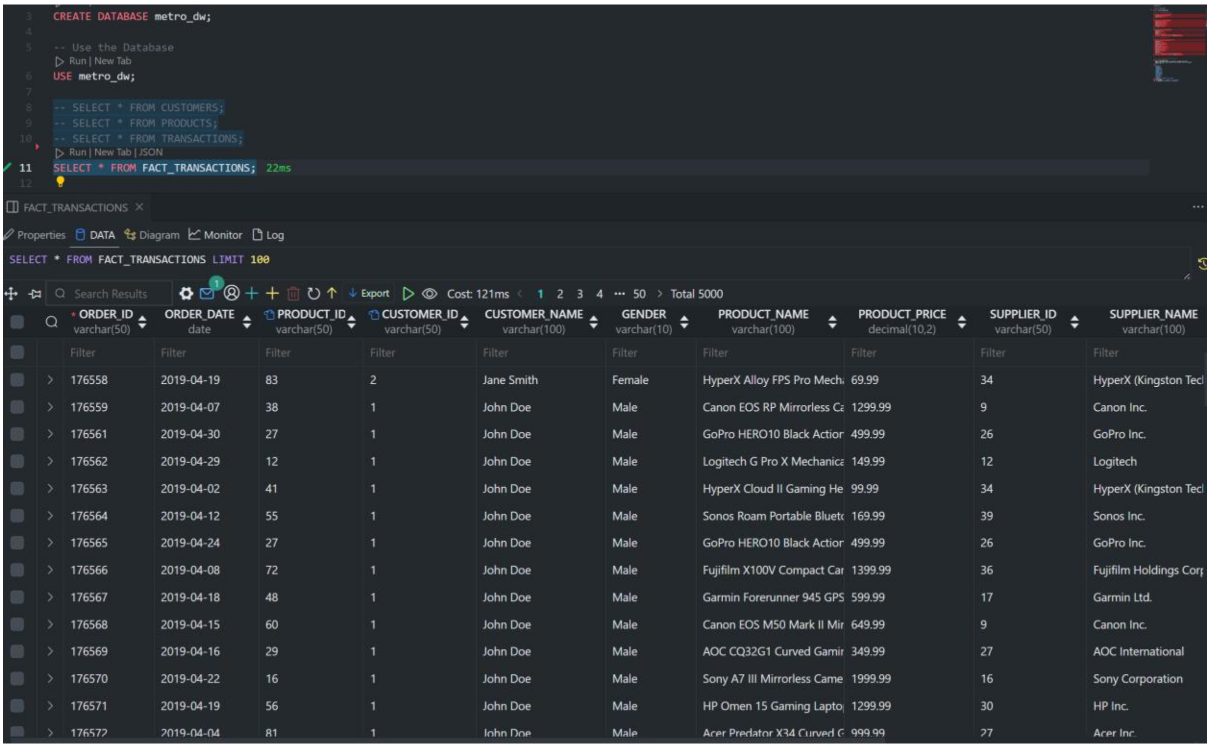
1. **Reading Stream Data**:
   - Imported transactional data (TRANSACTIONS) into memory in chunks.
   - Organized using a queue for staggered processing.
2. **Partitioning Master Data**:
   - Master data (CUSTOMERS, PRODUCTS) was divided into memory-efficient partitions.
   - Cyclical traversal ensured all data was joined.
3. **Joining and Enriching Data**:

- Transactions were joined with master data using hash tables.
- Computed derived attributes like SALE = QUANTITY × PRODUCT_PRICE.
- Deduplicated records before loading into the DW.

4. **Loading Data**:

   - Joined and enriched data was inserted into the **FACT_TRANSACTIONS** table.
   - Data integrity ensured through SQL constraints.

After loading the data into FACT_TRANSACTIONS table:



```
CREATE DATABASE metro_dw;

-- Use the Database
USE metro_dw;

-- SELECT * FROM CUSTOMERS;
-- SELECT * FROM PRODUCTS;
-- SELECT * FROM TRANSACTIONS;
SELECT * FROM FACT_TRANSACTIONS; 22ms
```

FACT_TRANSACTIONS

```
SELECT * FROM FACT_TRANSACTIONS LIMIT 100
```

Cost: 121ms 1 2 3 4 ··· 50 Total 5000

ORDER_ID varchar(50)	ORDER_DATE date	PRODUCT_ID varchar(50)	CUSTOMER_ID varchar(50)	CUSTOMER_NAME varchar(100)	GENDER varchar(10)	PRODUCT_NAME varchar(100)	PRODUCT_PRICE decimal(10,2)	SUPPLIER_ID varchar(50)	SUPPLIER_NAME varchar(100)
176558	2019-04-19	83	2	Jane Smith	Female	HyperX Alloy FPS Pro Mech	69.99	34	HyperX (Kingston Tec
176559	2019-04-07	38	1	John Doe	Male	Canon EOS RP Mirrorless C	1299.99	9	Canon Inc.
176561	2019-04-30	27	1	John Doe	Male	GoPro HERO10 Black Actior	499.99	26	GoPro Inc.
176562	2019-04-29	12	1	John Doe	Male	Logitech G Pro X Mechanica	149.99	12	Logitech
176563	2019-04-02	41	1	John Doe	Male	HyperX Cloud II Gaming He	99.99	34	HyperX (Kingston Tec
176564	2019-04-12	55	1	John Doe	Male	Sonos Roam Portable Bluet	169.99	39	Sonos Inc.
176565	2019-04-24	27	1	John Doe	Male	GoPro HERO10 Black Actior	499.99	26	GoPro Inc.
176566	2019-04-08	72	1	John Doe	Male	Fujifilm X100V Compact Car	1399.99	36	Fujifilm Holdings Cor
176567	2019-04-18	48	1	John Doe	Male	Garmin Forerunner 945 GPS	599.99	17	Garmin Ltd.
176568	2019-04-15	60	1	John Doe	Male	Canon EOS M50 Mark II Mir	649.99	9	Canon Inc.
176569	2019-04-16	29	1	John Doe	Male	AOC CQ32G1 Curved Gamir	349.99	27	AOC International
176570	2019-04-22	16	1	John Doe	Male	Sony A7 III Mirrorless Came	1999.99	16	Sony Corporation
176571	2019-04-19	56	1	John Doe	Male	HP Omen 15 Gaming Lapto	1299.99	30	HP Inc.
176572	2019-04-04	81	1	John Doe	Male	Acer Predator X34 Curved G	999.99	27	Acer Inc.

3. OLAP Queries

The OLAP queries in olap_queries.sql provided actionable insights:

1. **Top Revenue-Generating Products**: Identified products with the highest sales.
2. **Revenue Growth Trends**: Analyzed quarterly revenue growth per store.
3. **Supplier Contributions**: Showed sales contribution by suppliers.
4. **Seasonal Analysis**: Compared product sales across seasonal periods.
5. **Revenue Volatility**: Measured monthly revenue fluctuations by store and supplier.
6. **Product Affinity**: Identified frequently purchased product pairs.
7. **Yearly Trends**: Aggregated yearly revenue by store, supplier, and product.
8. **Revenue Analysis**: Compared sales in the first and second halves of the year.
9. **High Revenue Spikes**: Identify High Spikes in Product Sales and Highlight Outliers.
10. **View Creation**: Create a View STORE_QUARTERLY_SALES for Optimized Sales Analysis.

All queries ran successfully, leveraging the star schema's design for efficient performance.

```
OUTPUT OF OLAP QUERIES

1) Top 5 Revenue Generating Products:
Canon EOS-1D X Mark III DSLR Camera: $1.955196992E7
Canon EOS R5 Mirrorless Camera: $8385976.04
Nikon D850 DSLR Camera: $6776977.41
MSI GS66 Stealth Gaming Laptop: $6139969.3
LG C1 OLED 4K TV: $5599980.0


2) Revenue Growth Rate Quarterly for Year 2017:
No Data for Year 2017

3) Supplier Sales Contribution by Store and Product Name
Total Sales from All Suppliers: $2.0686885997E8


4) Present Total Sales for Products
Total Sales Across All Seasons: 2.0686885997E8


5) Monthly Revenue Volatility
Average Revenue Volatility: -98.98688325%


6) Top 5 Products Purchased Together
Most Frequently Purchased Together Count: 0


7) Yearly Revenue Trends by Store, Supplier, and Product
Total Revenue for the Year: 2.0686211024E8


8) Sales Analysis for Products for H1 and H2
Total Sales for H1: $2.0686885997E8
Total Sales for H2: $0.0


9) High Revenue Spikes
Number of High Revenue Spikes: 194


10) Create a View STORE_QUARTERLY_SALES for Optimized Sales Analysis
```

Shortcomings of MESHJOIN Algorithm

1. **Memory Constraints**:
    - Handling large master datasets in memory partitions requires significant resources.
    - Scales poorly with increased data size.
2. **Latency in Cyclical Processing**:
    - Staggered processing introduces delays as transactions wait for all partitions to be processed.

What I Learned

1. **Designing a Star Schema**:
    - Mapping business operations to relational models using dimensions and fact tables.
    - Ensuring primary and foreign key relationships for robust data integrity.
2. **Implementing ETL Pipelines**:
    - Leveraged the MESHJOIN algorithm to process and enrich streaming data.
    - Addressed challenges in partitioning, joining, and loading data efficiently.
3. **OLAP Query Development**:
    - Crafted complex queries for revenue, growth, and product affinity analysis.
    - Enhanced understanding of analytical frameworks like slicing, dicing, and rollups.
4. **Real-Time Data Integration**:
    - Learned techniques for integrating streaming data with static datasets in near-real-time.

Conclusion

The project successfully delivered a functional DW prototype for METRO Shopping Store. It demonstrated the practical application of data warehousing, ETL pipelines, and OLAP analytics, providing valuable insights for business decisions. The MESHJOIN algorithm proved effective for near-real-time data integration, despite its limitations. Overall, this project deepened my knowledge of data engineering and analytics in real-world scenarios.