Data Warehouse

Metro

Project Report

21i-2654 Muhammad Zoraib Qadir

Submitted to: Muhammad Asif Naeem

Table of Content

INTRODUCTION

In the fast-paced retail industry, understanding customer behavior is critical for optimizing sales strategies, product offerings, and promotions. METRO, one of the largest supermarket chains in Pakistan, deals with a vast amount of transactional data generated daily by its customers. To effectively analyze this data and gain valuable insights, METRO needs a system that allows for the real-time processing and analysis of customer transactions. This project addresses this need by designing, implementing, and analyzing a near-real-time Data Warehouse (DW) prototype.

The project focuses on building a near-real-time DW that consolidates data from multiple sources, transforming and enriching it to create a comprehensive view of customer shopping behavior. To achieve this, an ETL (Extract, Transform, Load) process is implemented, where:

- **Extract**: The raw transactional data is extracted from various Data Sources (DSs).
- **Transform**: The extracted data is transformed to include missing or enriched information.
- **Load**: The transformed data is then loaded into the DW for analysis.

A critical feature of the transformation phase is the enrichment of the data using **Master Data (MD)**, such as customer details, product information, and supplier data. This enrichment ensures that the data in the DW reflects a complete and up-to-date view of

customer shopping behavior.

To ensure efficient and accurate data transformation, this project employs the **MESHJOIN** algorithm, a stream-relation join operator, which facilitates real-time data integration. By using this algorithm, the system is capable of joining transactional data with the master data sources, providing enriched data that reflects the most up-to-date information available.

This project aims to enable METRO to leverage the power of **real-time analytics** for better decision-making. By developing and implementing the near-real-time DW prototype, METRO will be able to:

- Understand customer preferences
- Identify emerging trends
- Implement personalized promotions and sales strategies in a timely manner

The ability to perform this analysis in near real-time is key to staying competitive in the retail market, ensuring that METRO's operations are responsive, data-driven, and optimized for customer satisfaction.

Project Overview

This project involves a **data warehouse** implementation using a **star schema** for efficient querying and analysis of transactional data. It includes a **Java-based ETL process** that extracts, transforms, and loads data into a database. The process also incorporates **OLAP (Online Analytical Processing) queries** to analyze the transformed data stored in the star schema.

The project has three major components:

- **Java ETL Project**: Extracts, transforms, and loads data using mesh join transformations.
- **Star Schema Creation**: Defines the database schema where the transformed data is stored.
- **OLAP Queries**: Queries executed to analyze the data from the star schema.

Star Schema

The **star schema** is a fundamental part of the data warehouse model and supports fast querying and analysis. It consists of:

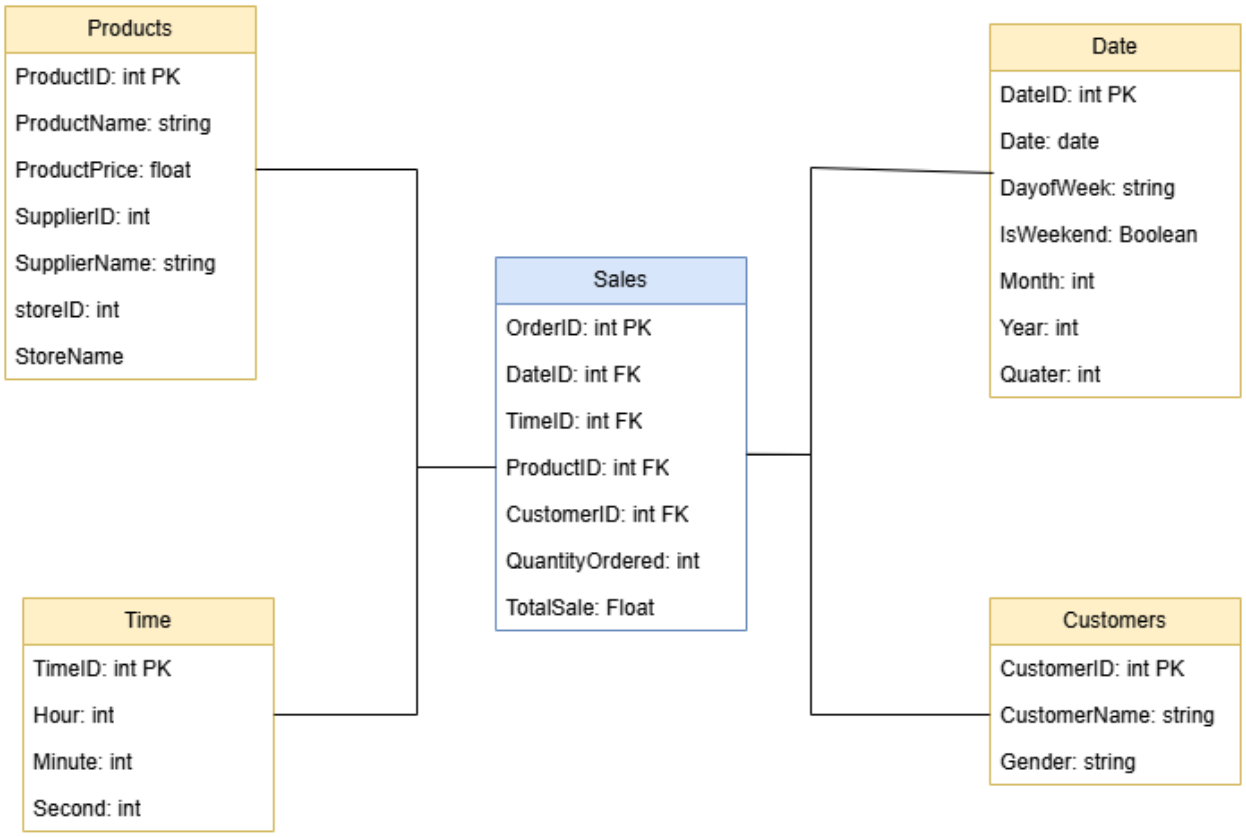
- **Fact Table**: Contains transactional or quantitative data.
- **Dimension Tables**: Contain descriptive information related to the fact data.

Fact Table:

1. Sales

Dimensions:

1. Products
2. Date
3. Time
4. Customers



Mesh Join Algorithm

The Mesh Join is a technique used to combine data from multiple datasets based on their relationships. In this project, the Mesh Join operation is performed between the product and transaction datasets, allowing for the aggregation of sales data related to each product.

Mesh Join Steps:

- **Loading Data**:

  Each loop step reads a new chunk of customer transactions into the hash table (main memory). Simultaneously, a partition of the MD is loaded into the disk-buffer, replacing the previous MD partition.

- **Joining Process**:

  Once the MD partition is in the disk-buffer, the algorithm probes each tuple in the disk-buffer against the hash table (which holds customer transactions). If a matching tuple is found, the join output is generated.

- **Queue Management**:

  The queue holds customer transactions in the order they arrive, with each chunk being one element in the queue. As each chunk remains in memory for the duration of one full cycle of the MD, the processing of each chunk overlaps with the processing of other chunks. After each iteration, the algorithm removes the oldest chunk of customer transactions from the hash table, along with its pointers from the queue. These chunks are considered fully processed once they have been joined with the entire MD partition.

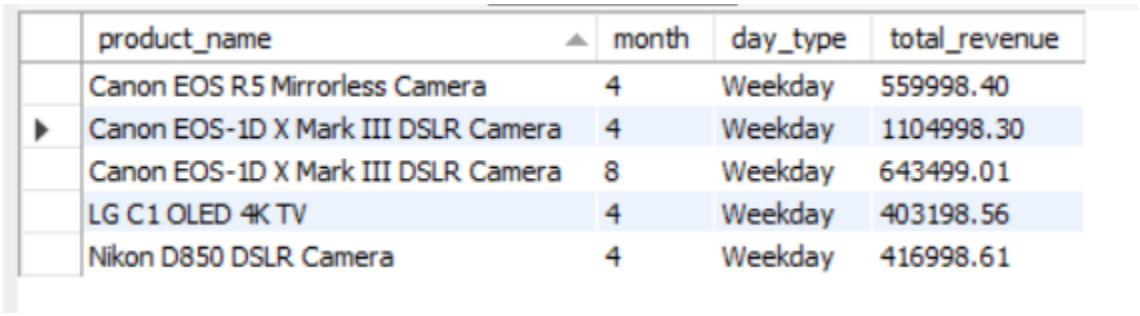
- **Staggered Processing**:

  The algorithm's staggered processing ensures that incoming data is handled efficiently, with each chunk staying in memory for an overlapping period. This overlap allows the system to maintain real-time performance and quickly process

new data while continuously joining it with the master data.

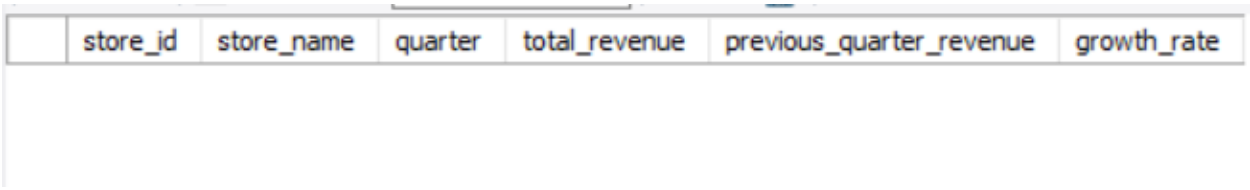
DataWarehouse Analysis

Q1. Top Revenue-Generating Products on Weekdays and Weekends with Monthly Drill-Down



product_name	month	day_type	total_revenue
Canon EOS R5 Mirrorless Camera	4	Weekday	559998.40
Canon EOS-1D X Mark III DSLR Camera	4	Weekday	1104998.30
Canon EOS-1D X Mark III DSLR Camera	8	Weekday	643499.01
LG C1 OLED 4K TV	4	Weekday	403198.56
Nikon D850 DSLR Camera	4	Weekday	416998.61

Q2. Trend Analysis of Store Revenue Growth Rate Quarterly for 2017



store_id	store_name	quarter	total_revenue	previous_quarter_revenue	growth_rate

No Data Available for 2017

Q3. Detailed Supplier Sales Contribution by Store and Product Name

store_id	store_name	supplier_id	supplier_name	product_name	total_sales_contribution
1	Electro Mart	1	Apple Inc.	iPhone 13 Pro	347596.84
1	Electro Mart	22	Google LLC	Google Nest Hub (2nd Gen)	35596.44
1	Electro Mart	22	Google LLC	Google Pixel 6	233096.67
1	Electro Mart	11	LG Electronics	LG C1 OLED 4K TV	940796.64
1	Electro Mart	11	LG Electronics	LG OLED C1 4K TV	817496.73
1	Electro Mart	29	OnePlus Technology	OnePlus 9 Pro	350096.11
1	Electro Mart	29	OnePlus Technology	OnePlus Nord 2	172996.54
1	Electro Mart	42	Ring (Amazon)	Ring Video Doorbell Pro 2	79996.80
1	Electro Mart	3	Samsung Electronics	Samsung QLED 4K Smart TV	506996.62
1	Electro Mart	16	Sony Corporation	Sony Xperia 1 III	445896.57
2	Tech Haven	25	Acer Inc.	Acer Aspire 5 Laptop	187546.59
2	Tech Haven	1	Apple Inc.	Apple iPad Pro (12.9-inch)	338796.92
2	Tech Haven	5	Apple Inc.	iPad Air	207596.54
2	Tech Haven	1	Apple Inc.	MacBook Air M2	385196.79
2	Tech Haven	1	Apple Inc.	MacBook Pro M2	639996.80
2	Tech Haven	2	Dell Technologies	Dell Inspiron 14 Laptop	244796.94
2	Tech Haven	2	Dell Technologies	Dell XPS 13 Laptop	400396.92
2	Tech Haven	30	HP Inc.	HP Pavilion 27xw Monitor	82746.69
2	Tech Haven	30	HP Inc.	HP Spectre x360 Laptop	464796.68
2	Tech Haven	32	Lenovo Group	Lenovo ThinkPad X1 Carbon Gen 9 Laptop	644096.61

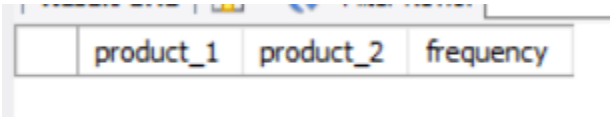
Q4. Seasonal Analysis of Product Sales Using Dynamic Drill-Down

product_name	season	total_sales
Acer Aspire 5 Laptop	Spring	110547.99
Acer Aspire 5 Laptop	Summer	76998.60
Acer Predator Helios 300 Gaming Laptop	Spring	230398.08
Acer Predator Helios 300 Gaming Laptop	Summer	145198.79
Acer Predator Helios 300 Gaming Laptop	Fall	1199.99
Acer Predator X34 Curved Gaming Monitor	Spring	213997.86
Acer Predator X34 Curved Gaming Monitor	Summer	117998.82
Acer Predator XB271HU Gaming Monitor	Spring	119398.01
Acer Predator XB271HU Gaming Monitor	Summer	75598.74
AirPods Pro	Spring	51247.95
AirPods Pro	Summer	33748.65
Alienware Aurora Gaming PC	Spring	307798.29
Alienware Aurora Gaming PC	Summer	262798.54
Alienware AW2521HFL Gaming Monitor	Spring	104997.90
Alienware AW2521HFL Gaming Monitor	Summer	67498.65
Anker Soundcore Flare+ Portable Speaker	Spring	20797.92
Anker Soundcore Flare+ Portable Speaker	Summer	16798.32
Anker Soundcore Liberty Air 2 Pro Earbuds	Spring	28467.81
Anker Soundcore Liberty Air 2 Pro Earbuds	Summer	15598.80
AOC CQ32G1 Curved Gaming Monitor	Spring	78747.75
AOC CQ32G1 Curved Gaming Monitor	Summer	50398.56
Apple AirPods (3rd generation)	Spring	36177.99
Apple AirPods (3rd generation)	Summer	25018.61
Apple AirPods Max	Spring	111647.97

Q5. Store-Wise and Supplier-Wise Monthly Revenue Volatility

store_name	supplier_name	month	monthly_revenue	previous_month_revenue	volatility
Electro Mart	Apple Inc.	4	201298.17	NULL	NULL
Electro Mart	Apple Inc.	8	146298.67	201298.17	-27.322404
Electro Mart	Google LLC	4	153995.88	NULL	NULL
Electro Mart	Google LLC	5	99.99	153995.88	-99.935070
Electro Mart	Google LLC	8	114597.24	99.99	114508.700870
Electro Mart	LG Electronics	4	1052296.02	NULL	NULL
Electro Mart	LG Electronics	8	705997.35	1052296.02	08864
Electro Mart	OnePlus Technology	4	312095.63	NULL	
Electro Mart	OnePlus Technology	8	210997.02	312095.63	-32.393472
Electro Mart	Ring (Amazon)	4	48748.05	NULL	NULL
Electro Mart	Ring (Amazon)	8	31248.75	48748.05	-35.897436
Electro Mart	Samsung Electronics	4	308997.94	NULL	NULL
Electro Mart	Samsung Electronics	8	197998.68	308997.94	-35.922330
Electro Mart	Sony Corporation	4	297697.71	NULL	NULL
Electro Mart	Sony Corporation	8	148198.86	297697.71	-50.218341
Game Zone	Acer Inc.	4	332795.88	NULL	NULL
Game Zone	Acer Inc.	5	599.99	332795.88	-99.819712
Game Zone	Acer Inc.	8	193597.56	599.99	32166.797780
Game Zone	Alienware (Dell)	4	348095.88	NULL	NULL
Game Zone	Alienware (Dell)	5	499.99	348095.88	-99.856364
Game Zone	Alienware (Dell)	8	252297.11	499.99	50360.431209
Game Zone	AOC International	4	78747.75	NULL	NULL
Game Zone	AOC International	8	50398.56	78747.75	-36.000000
Game Zone	ASUS	4	195996.30	NULL	NULL

Q6. Top 5 Products Purchased Together Across Multiple Orders (Product Affinity Analysis)



Data not Provided

Q7. Yearly Revenue Trends by Store, Supplier, and Product with ROLLUP

store_name	supplier_name	product_name	total_sales
Electro Mart	Apple Inc.	iPhone 13 Pro	347596.84
Electro Mart	Google LLC	NULL	268693.11
Electro Mart	Google LLC	Google Nest Hub (2nd Gen)	35596.44
Electro Mart	Google LLC	Google Pixel 6	233096.67
Electro Mart	LG Electronics	NULL	1758293.37
Electro Mart	LG Electronics	LG C1 OLED 4K TV	940796.64
Electro Mart	LG Electronics	LG OLED C1 4K TV	817496.73
Electro Mart	OnePlus Tech...	NULL	523092.65
Electro Mart	OnePlus Tech...	OnePlus 9 Pro	350096.11
Electro Mart	OnePlus Tech...	OnePlus Nord 2	172996.54
Electro Mart	Ring (Amazon)	NULL	79996.80
Electro Mart	Ring (Amazon)	Ring Video Doorbell Pro 2	79996.80
Electro Mart	Samsung Elec...	NULL	506996.62
Electro Mart	Samsung Elec...	Samsung QLED 4K Smart TV	506996.62
Electro Mart	Sony Corpora...	NULL	445896.57
Electro Mart	Sony Corpora...	Sony Xperia 1 III	445896.57
Game Zone	NULL	NULL	3241839.16
Game Zone	Acer Inc.	NULL	526993.43
Game Zone	Acer Inc.	Acer Predator X34 Curved...	331996.68
Game Zone	Acer Inc.	Acer Predator XB271HU G...	194996.75
Game Zone	Alienware (Dell)	NULL	600892.98
Game Zone	Alienware (Dell)	Alienware AW2521HFL Ga...	172496.55
Game Zone	Alienware (Dell)	Dell Alienware AW3420DW...	428396.43
Game Zone	AOC Internat...	NULL	129146.31

Q8. Revenue and Volume-Based Sales Analysis for Each Product for H1 and H2

product_name	H1_revenue	H2_revenue	H1_quantity	H2_quantity	total_revenue	total_quantity
AirPods Pro	51247.95	33748.65	205	135	84996.60	340
Alienware Aurora Gaming PC	307798.29	262798.54	171	146	570596.83	317
Alienware AW2521HFL Gaming Monitor	104997.90	67498.65	210	135	172496.55	345
Anker Soundcore Flare+ Portable Speaker	20797.92	16798.32	208	168	37596.24	376
Anker Soundcore Liberty Air 2 Pro Earbuds	28467.81	15598.80	219	120	44066.61	339
AOC CQ32G1 Curved Gaming Monitor	78747.75	50398.56	225	144	129146.31	369
Apple AirPods (3rd generation)	36177.99	25018.61	201	139	61196.60	340
Apple AirPods Max	111647.97	66548.79	203	121	178196.76	324
Apple HomePod Mini	20997.90	13198.68	210	132	34196.58	342
Apple iPad Pro (12.9-inch)	204598.14	134198.78	186	122	338796.92	308
Apple Watch SE	51798.15	43678.44	185	156	95476.59	341
Apple Watch Series 7	71998.20	51998.70	180	130	123996.90	310
ASUS ROG Swift PG279Q Gaming Monitor	133698.09	102898.53	191	147	236596.62	338
ASUS TUF Gaming VG279QM Monitor	63698.18	44448.73	182	127	108146.91	309
Beats Powerbeats Pro Wireless Earphones	55247.79	35748.57	221	143	90996.36	364
Bose QuietComfort 35 II Wireless Headph...	68697.71	49498.35	229	165	118196.06	394
Bose SoundLink Revolve+ Bluetooth Spea...	62697.91	36598.78	209	122	99296.69	331
Canon EOS 90D DSLR Camera	209998.25	127198.94	175	106	337197.19	281
Canon EOS M50 Mark II Mirrorless Camera	115048.23	84498.70	177	130	199546.93	307
Canon EOS R5 Mirrorless Camera	787497.75	479498.63	225	137	1266996.38	362
Canon EOS RP Mirrorless Camera	248298.09	165098.73	191	127	413396.82	318
Canon EOS-1D X Mark III DSLR Camera	1475497.73	857998.68	227	132	2333496.41	359
Corsair K95 RGB Platinum XT Mechanical G...	39398.03	28998.55	197	145	68396.58	342
Corsair Virtuoso RGB Wireless Gaming He...	38877.84	24118.66	216	134	62996.50	350

Q9. Identify High Revenue Spikes in Product Sales and Highlight Outliers

product_id	product_name	date	daily_sales	daily_avg_sales	status
1	iPhone 13 Pro	2019-04-15	4399.96	5698.308852	Normal
1	iPhone 13 Pro	2019-04-16	8799.92	5698.308852	Normal
1	iPhone 13 Pro	2019-04-17	4399.96	5698.308852	Normal
1	iPhone 13 Pro	2019-04-18	12099.89	5698.308852	Outlier
1	iPhone 13 Pro	2019-04-19	2199.98	5698.308852	Normal
1	iPhone 13 Pro	2019-04-20	6599.94	5698.308852	Normal
1	iPhone 13 Pro	2019-04-21	3299.97	5698.308852	Normal
1	iPhone 13 Pro	2019-04-22	7699.93	5698.308852	Normal
1	iPhone 13 Pro	2019-04-23	3299.97	5698.308852	Normal
1	iPhone 13 Pro	2019-04-24	2199.98	5698.308852	Normal
1	iPhone 13 Pro	2019-04-25	3299.97	5698.308852	Normal
1	iPhone 13 Pro	2019-04-26	4399.96	5698.308852	Normal
1	iPhone 13 Pro	2019-04-27	8799.92	5698.308852	Normal
1	iPhone 13 Pro	2019-04-28	7699.93	5698.308852	Normal
1	iPhone 13 Pro	2019-04-29	14299.87	5698.308852	Outlier
1	iPhone 13 Pro	2019-04-30	6599.94	5698.308852	Normal
1	iPhone 13 Pro	2019-08-01	2199.98	5698.308852	Normal
1	iPhone 13 Pro	2019-08-02	3299.97	5698.308852	Normal
1	iPhone 13 Pro	2019-08-03	6599.94	5698.308852	Normal
1	iPhone 13 Pro	2019-08-04	4399.96	5698.308852	Normal
1	iPhone 13 Pro	2019-08-05	4399.96	5698.308852	Normal
1	iPhone 13 Pro	2019-08-06	4399.96	5698.308852	Normal
1	iPhone 13 Pro	2019-08-07	3299.97	5698.308852	Normal
1	iPhone 13 Pro	2019-08-08	4399.96	5698.308852	Normal

Q10. Create a View STORE_QUARTERLY_SALES for Optimized Sales Analysis

store_name	quarter	total_sales
Electro Mart	2	2375229.39
Electro Mart	3	1555336.57
Game Zone	2	1954137.25
Game Zone	3	1287701.91
Health Zone	2	566624.03
Health Zone	3	366509.58
InnoTech	2	2275783.62
InnoTech	3	1480989.38
Pakistan	2	397097.91
Pakistan	3	286898.49
Photo World	2	5551869.55
Photo World	3	3523130.65
Sound Zone	2	954326.66
Sound Zone	3	629701.21
Tech Haven	2	3010019.80
Tech Haven	3	1953880.48

Shortcomings of Mesh Join

While the Mesh Join is an effective method for combining datasets, it has several shortcomings:

1. **Memory Consumption:** Mesh Join can consume a significant amount of memory, especially with large datasets, as it requires loading entire datasets into memory.
2. **Performance Issues:** The performance of a Mesh Join can degrade as the size of the datasets increases, leading to longer processing times.
3. **Complexity in Handling Duplicates:** If there are duplicate entries in the datasets, the Mesh Join can produce unexpected results, necessitating additional logic to handle such cases.

Learnings from the Project

1. **Understanding ETL Processes**: The project provided hands-on experience in designing and implementing an ETL process, emphasizing the importance of data extraction, transformation, and loading.
2. **Data Warehousing Concepts:** Gained insights into data warehousing concepts, including star schema design, which facilitates efficient data retrieval and analysis.
3. **OLAP Queries:** Learned to formulate and execute OLAP queries to derive insights from the data warehouse, enabling analytical decision-making.

CONCLUSION

This project effectively combines ETL processing, a star schema database design, and OLAP queries to deliver a powerful data warehouse solution. It enables the efficient extraction, transformation, and loading of data into a database, followed by analytical querying to provide business insights. The **star schema** design ensures optimized data storage and querying, making it ideal for decision-making processes in areas such as sales analysis, product performance, and customer behavior.

If further modifications or additions are required, such as enhancing the ETL process, refining the schema, or expanding the OLAP queries, feel free to reach out.