Project Report: Near-Real-Time Data Warehouse for METRO Shopping Store

1. Project Overview

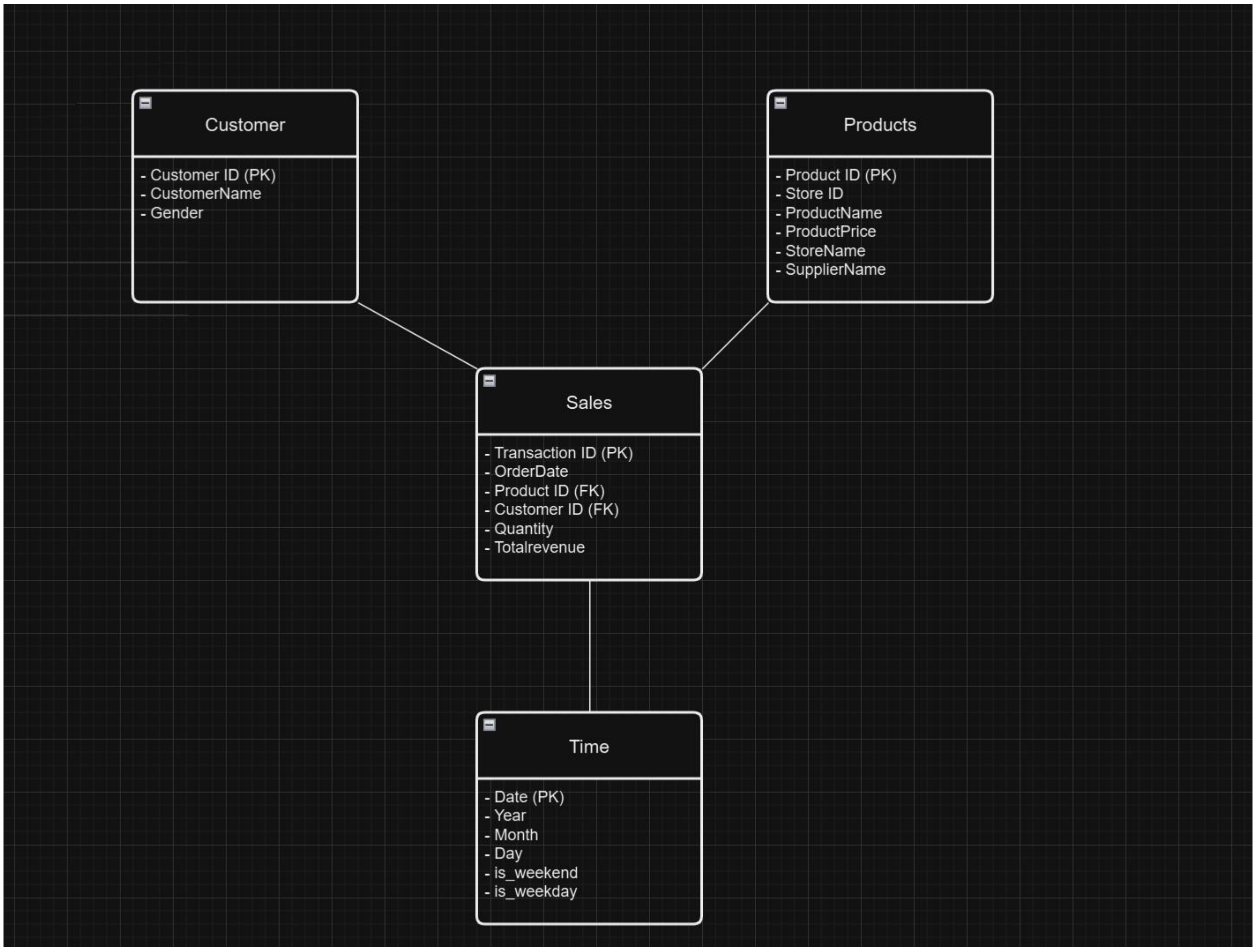
This project was about designing, building, and analyzing a near-real-time Data Warehouse (DW) for METRO, a major superstore chain in Pakistan. The goal was to create a system that could provide near-real-time insights into customer shopping behavior, helping optimize sales strategies like product promotions. We used a star schema, implemented the MESHJOIN algorithm in Java, and ran OLAP queries to analyze the data.

The DW aggregates customer transactions as they happen, offering insights that can help METRO's management make better decisions. By building a near-real-time ETL (Extraction, Transformation, Loading) process, we collected, transformed, enriched, and loaded data from various sources to support multidimensional analysis.

2. Data Warehouse Schema

The Data Warehouse was designed using a star schema. This model consists of a central fact table that records sales transactions and several dimension tables that provide context for those facts, such as product details, customer information, store locations, and time periods.

Star Schema Components:



3. MESHJOIN Algorithm

The MESHJOIN algorithm was implemented to enable the stream-relation join needed for enriching transactional data. The algorithm works by continuously loading incoming customer transactions and joining them with the master data in a cyclic fashion.

Main Components of MESHJOIN:

- Disk Buffer: Loads and stores master data partitions for memory processing.
- Hash Table: Keeps customer transactions in memory for processing.
- Queue: Manages customer transactions based on their arrival times, ensuring each transaction gets enriched with all master data before being loaded into the DW.

The meshjoin2.java code applies this algorithm to process customer transactions in batches, enriching each transaction using data from the ‘Products’ and ‘Customers’ tables.

4. Shortcomings in MESHJOIN Algorithm

1. High Memory Usage: The cyclic loading of master data partitions requires a lot of memory, especially with large datasets.

2. Latency in Processing: Although it's near-real-time, the batch processing causes some delay, which can be an issue for high-frequency transactions.

3. Complexity in Implementation: Managing the queue, hash table, and disk buffer adds complexity to the code, making it harder to maintain.

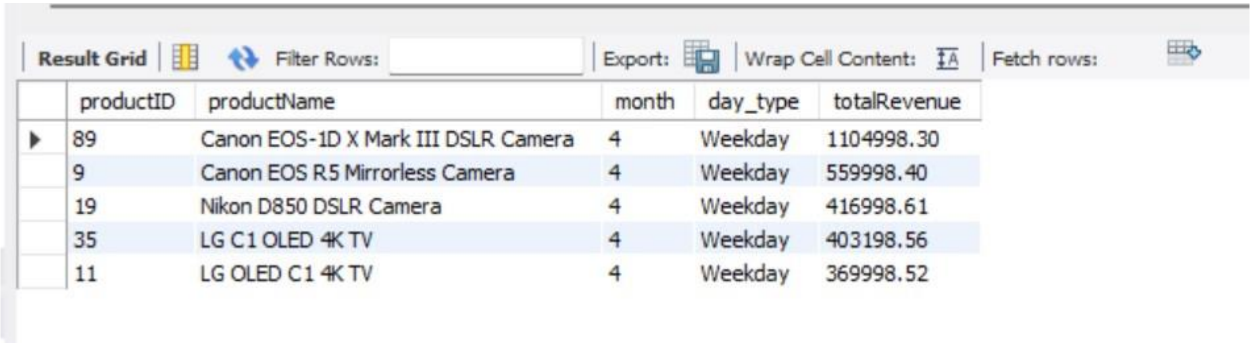
5. Lessons Learned

- Data Warehouse Design: Designing a star schema for multidimensional analysis provides a flexible and efficient structure to support various business analytics.

- ETL Process: Implementing near-real-time ETL is challenging, particularly in balancing latency and performance.

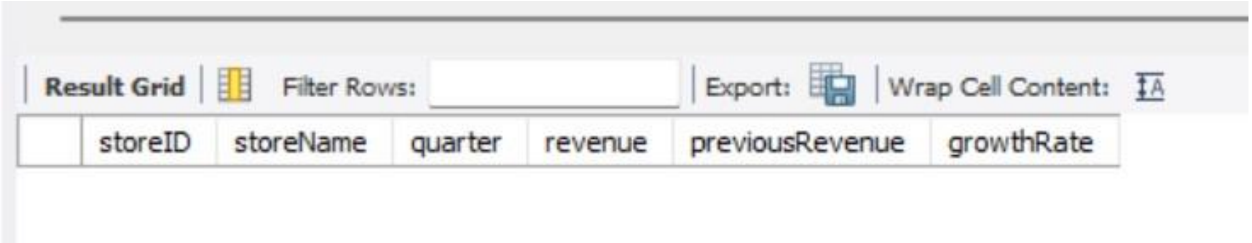
- MESHJOIN Algorithm: Understanding stream-relation joins is crucial when building systems that integrate streaming data with existing sources for real-time analysis.

6. OLAP Queries output Q1:



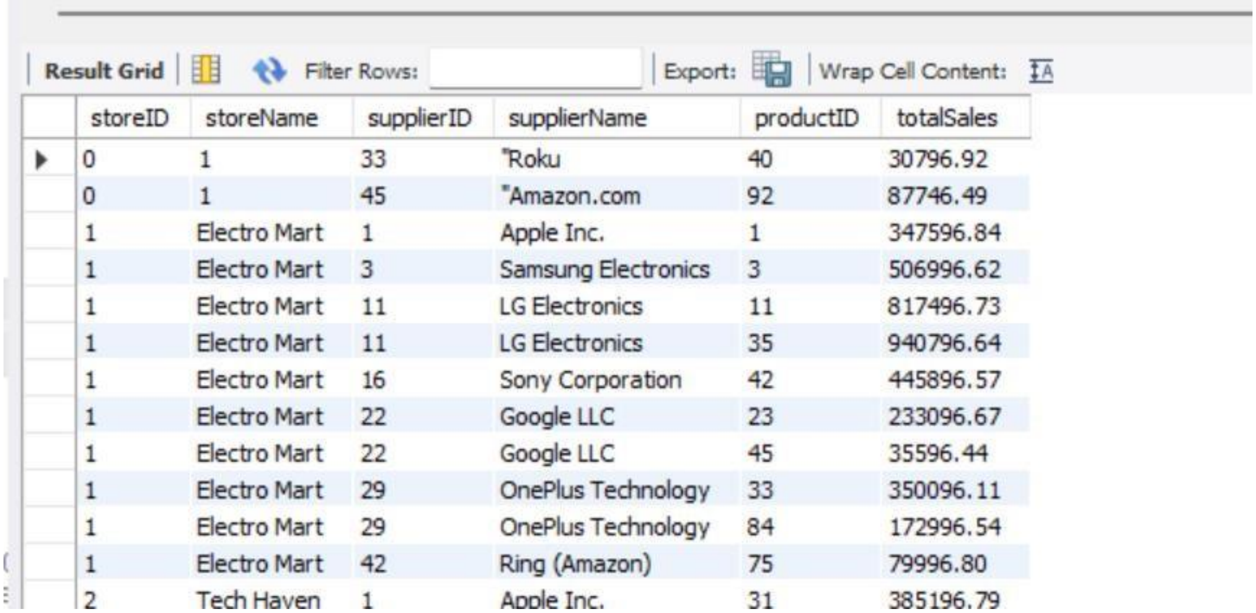
productID	productName	month	day_type	totalRevenue
89	Canon EOS-1D X Mark III DSLR Camera	4	Weekday	1104998.30
9	Canon EOS R5 Mirrorless Camera	4	Weekday	559998.40
19	Nikon D850 DSLR Camera	4	Weekday	416998.61
35	LG C1 OLED 4K TV	4	Weekday	403198.56
11	LG OLED C1 4K TV	4	Weekday	369998.52

Q2:



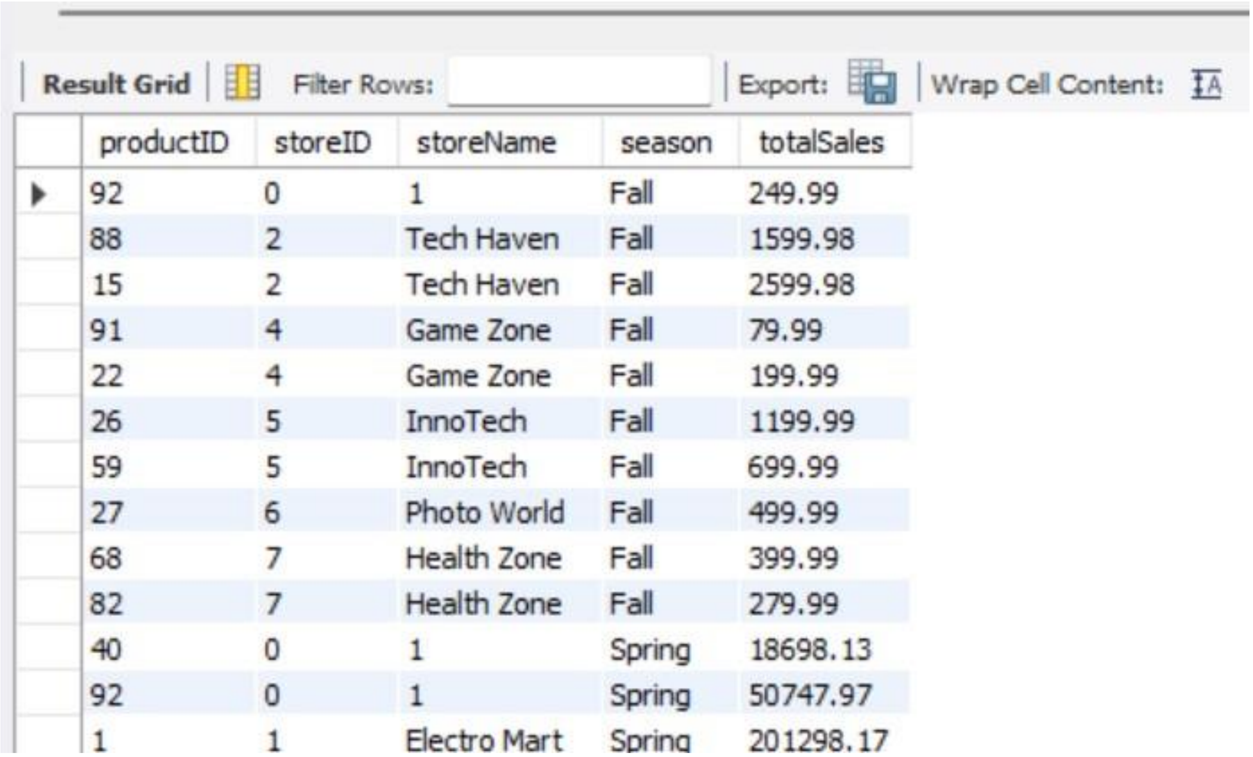
storeID	storeName	quarter	revenue	previousRevenue	growthRate

Q3:



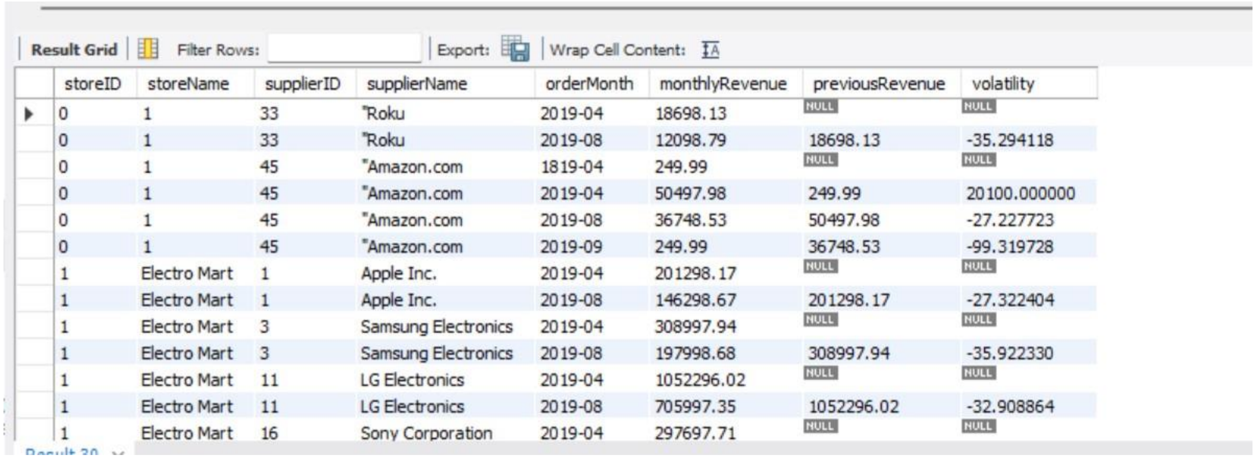
storeID	storeName	supplierID	supplierName	productID	totalSales
0	1	33	"Roku	40	30796.92
0	1	45	"Amazon.com	92	87746.49
1	Electro Mart	1	Apple Inc.	1	347596.84
1	Electro Mart	3	Samsung Electronics	3	506996.62
1	Electro Mart	11	LG Electronics	11	817496.73
1	Electro Mart	11	LG Electronics	35	940796.64
1	Electro Mart	16	Sony Corporation	42	445896.57
1	Electro Mart	22	Google LLC	23	233096.67
1	Electro Mart	22	Google LLC	45	35596.44
1	Electro Mart	29	OnePlus Technology	33	350096.11
1	Electro Mart	29	OnePlus Technology	84	172996.54
1	Electro Mart	42	Ring (Amazon)	75	79996.80
2	Tech Haven	1	Apple Inc.	31	385196.79

Q4:



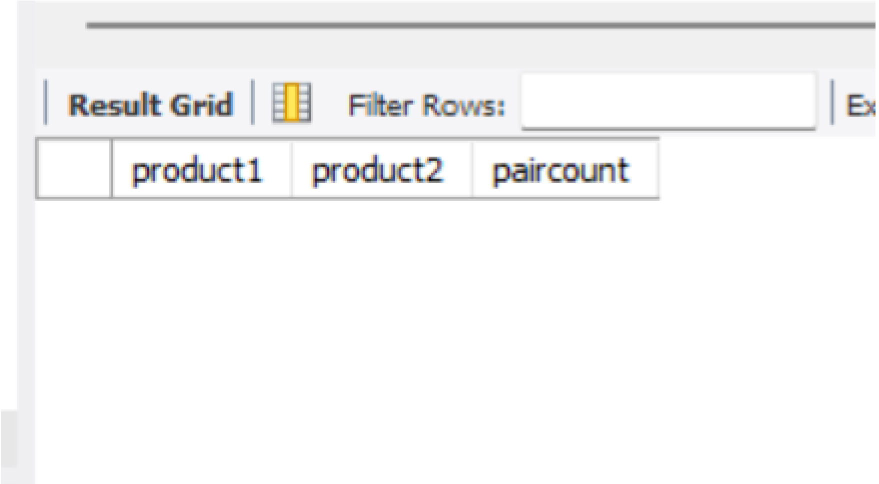
productID	storeID	storeName	season	totalSales
92	0	1	Fall	249.99
88	2	Tech Haven	Fall	1599.98
15	2	Tech Haven	Fall	2599.98
91	4	Game Zone	Fall	79.99
22	4	Game Zone	Fall	199.99
26	5	InnoTech	Fall	1199.99
59	5	InnoTech	Fall	699.99
27	6	Photo World	Fall	499.99
68	7	Health Zone	Fall	399.99
82	7	Health Zone	Fall	279.99
40	0	1	Spring	18698.13
92	0	1	Spring	50747.97
1	1	Electro Mart	Spring	201298.17

Q5:



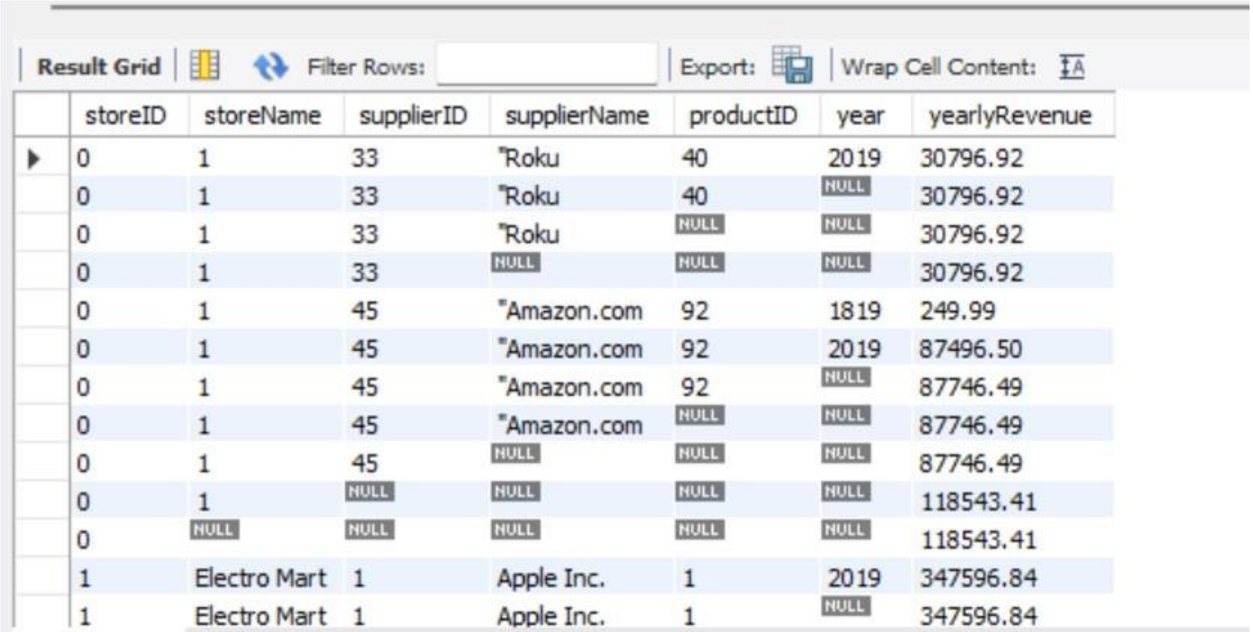
storeID	storeName	supplierID	supplierName	orderMonth	monthlyRevenue	previousRevenue	volatility
0	1	33	"Roku	2019-04	18698.13	NULL	NULL
0	1	33	"Roku	2019-08	12098.79	18698.13	-35.294118
0	1	45	"Amazon.com	1819-04	249.99	NULL	NULL
0	1	45	"Amazon.com	2019-04	50497.98	249.99	20100.000000
0	1	45	"Amazon.com	2019-08	36748.53	50497.98	-27.227723
0	1	45	"Amazon.com	2019-09	249.99	36748.53	-99.319728
1	Electro Mart	1	Apple Inc.	2019-04	201298.17	NULL	NULL
1	Electro Mart	1	Apple Inc.	2019-08	146298.67	201298.17	-27.322404
1	Electro Mart	3	Samsung Electronics	2019-04	308997.94	NULL	NULL
1	Electro Mart	3	Samsung Electronics	2019-08	197998.68	308997.94	-35.922330
1	Electro Mart	11	LG Electronics	2019-04	1052296.02	NULL	NULL
1	Electro Mart	11	LG Electronics	2019-08	705997.35	1052296.02	-32.908864
1	Electro Mart	16	Sony Corporation	2019-04	297697.71	NULL	NULL

Q6:



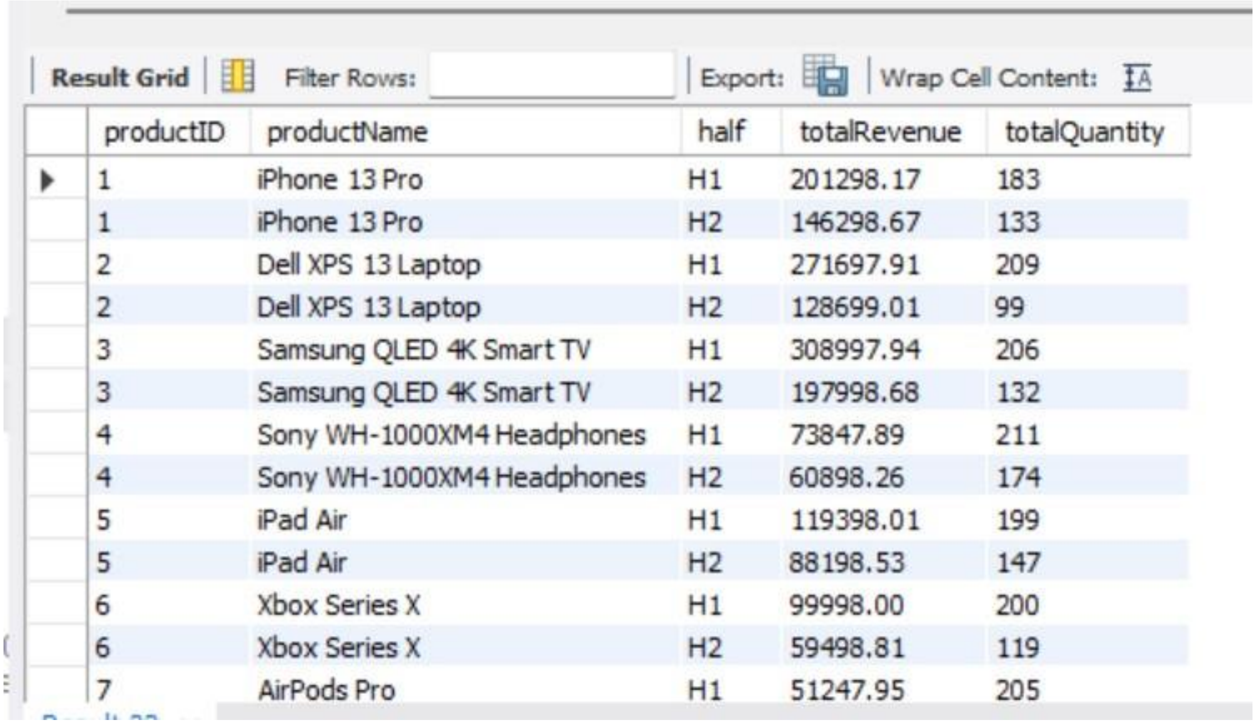
	product1	product2	paircount

Q7:



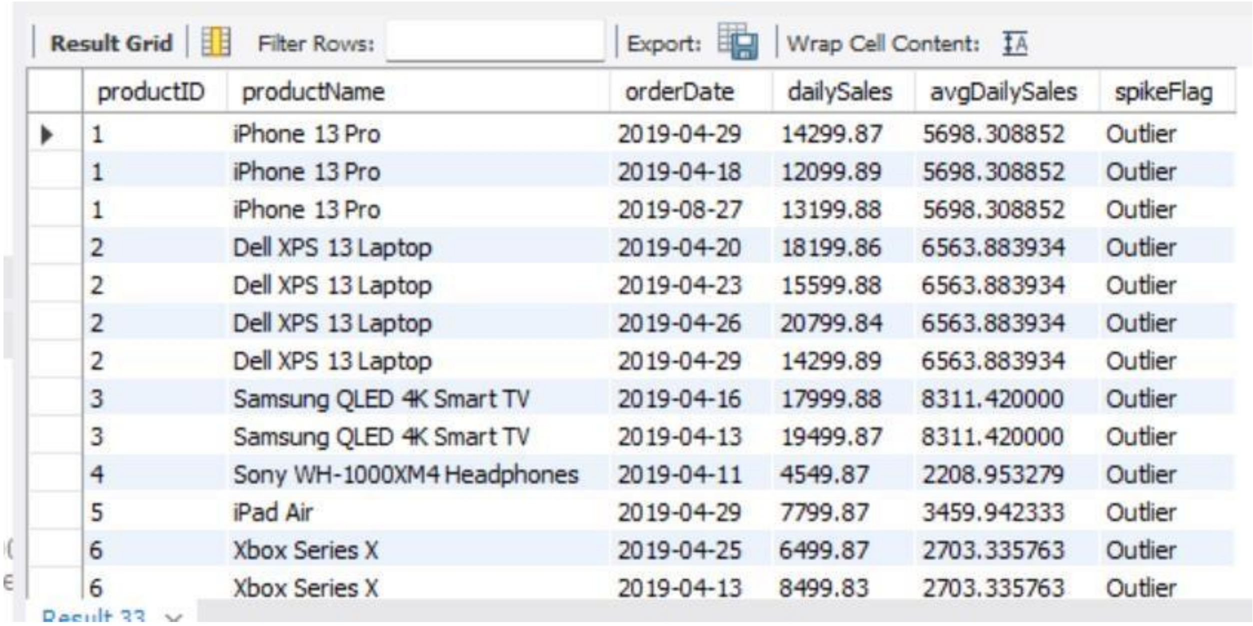
	storeID	storeName	supplierID	supplierName	productID	year	yearlyRevenue
▶	0	1	33	"Roku	40	2019	30796.92
	0	1	33	"Roku	40	NULL	30796.92
	0	1	33	"Roku	NULL	NULL	30796.92
	0	1	33	NULL	NULL	NULL	30796.92
	0	1	45	"Amazon.com	92	1819	249.99
	0	1	45	"Amazon.com	92	2019	87496.50
	0	1	45	"Amazon.com	92	NULL	87746.49
	0	1	45	"Amazon.com	NULL	NULL	87746.49
	0	1	45	NULL	NULL	NULL	87746.49
	0	1	NULL	NULL	NULL	NULL	118543.41
	0	NULL	NULL	NULL	NULL	NULL	118543.41
	1	Electro Mart	1	Apple Inc.	1	2019	347596.84
	1	Electro Mart	1	Apple Inc.	1	NULL	347596.84

Q8:



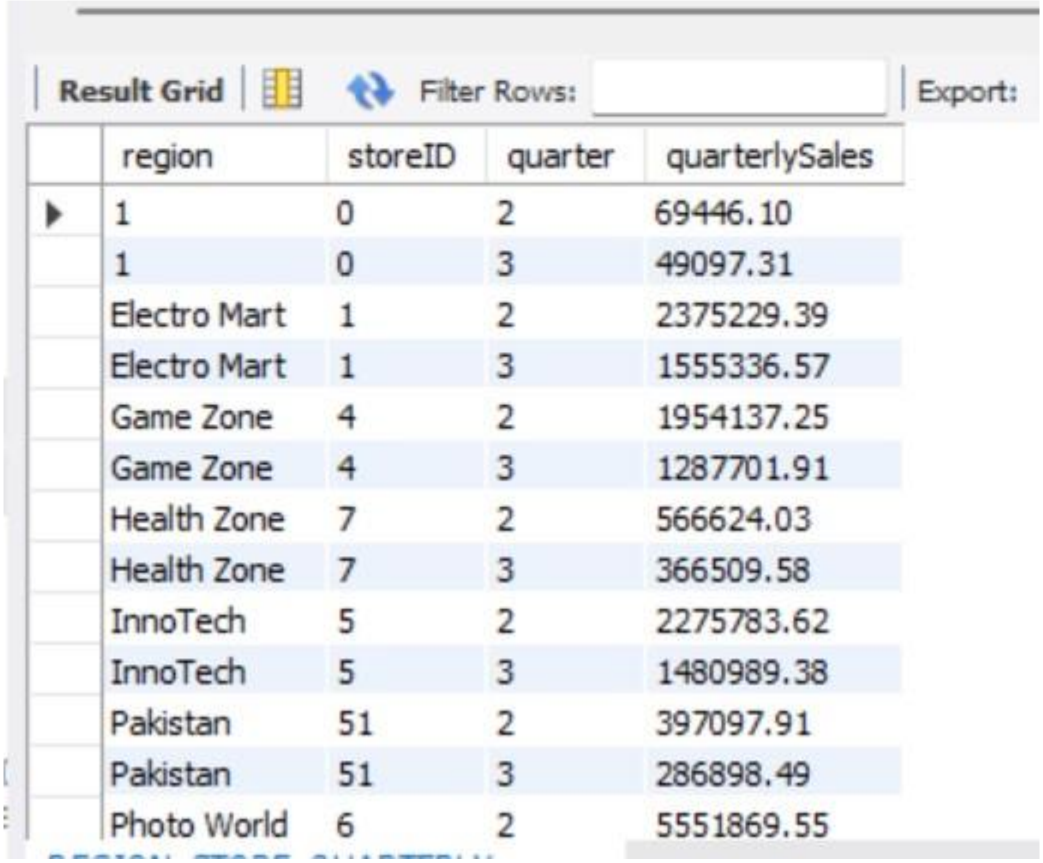
productID	productName	half	totalRevenue	totalQuantity
1	iPhone 13 Pro	H1	201298.17	183
1	iPhone 13 Pro	H2	146298.67	133
2	Dell XPS 13 Laptop	H1	271697.91	209
2	Dell XPS 13 Laptop	H2	128699.01	99
3	Samsung QLED 4K Smart TV	H1	308997.94	206
3	Samsung QLED 4K Smart TV	H2	197998.68	132
4	Sony WH-1000XM4 Headphones	H1	73847.89	211
4	Sony WH-1000XM4 Headphones	H2	60898.26	174
5	iPad Air	H1	119398.01	199
5	iPad Air	H2	88198.53	147
6	Xbox Series X	H1	99998.00	200
6	Xbox Series X	H2	59498.81	119
7	AirPods Pro	H1	51247.95	205

Q9:



productID	productName	orderDate	dailySales	avgDailySales	spikeFlag
1	iPhone 13 Pro	2019-04-29	14299.87	5698.308852	Outlier
1	iPhone 13 Pro	2019-04-18	12099.89	5698.308852	Outlier
1	iPhone 13 Pro	2019-08-27	13199.88	5698.308852	Outlier
2	Dell XPS 13 Laptop	2019-04-20	18199.86	6563.883934	Outlier
2	Dell XPS 13 Laptop	2019-04-23	15599.88	6563.883934	Outlier
2	Dell XPS 13 Laptop	2019-04-26	20799.84	6563.883934	Outlier
2	Dell XPS 13 Laptop	2019-04-29	14299.89	6563.883934	Outlier
3	Samsung QLED 4K Smart TV	2019-04-16	17999.88	8311.420000	Outlier
3	Samsung QLED 4K Smart TV	2019-04-13	19499.87	8311.420000	Outlier
4	Sony WH-1000XM4 Headphones	2019-04-11	4549.87	2208.953279	Outlier
5	iPad Air	2019-04-29	7799.87	3459.942333	Outlier
6	Xbox Series X	2019-04-25	6499.87	2703.335763	Outlier
6	Xbox Series X	2019-04-13	8499.83	2703.335763	Outlier

Q10:



region	storeID	quarter	quarterlySales
1	0	2	69446.10
1	0	3	49097.31
Electro Mart	1	2	2375229.39
Electro Mart	1	3	1555336.57
Game Zone	4	2	1954137.25
Game Zone	4	3	1287701.91
Health Zone	7	2	566624.03
Health Zone	7	3	366509.58
InnoTech	5	2	2275783.62
InnoTech	5	3	1480989.38
Pakistan	51	2	397097.91
Pakistan	51	3	286898.49
Photo World	6	2	5551869.55

7. Conclusion

This project successfully demonstrated the implementation of a near-real-time Data Warehouse for METRO. By combining transactional data with enriched master data through MESHJOIN and using a flexible star schema, the DW supports insightful analyses that can improve business strategies. This prototype lays the groundwork for real-time customer behavior analytics, potentially boosting sales and enhancing customer experience for METRO.