Project Report

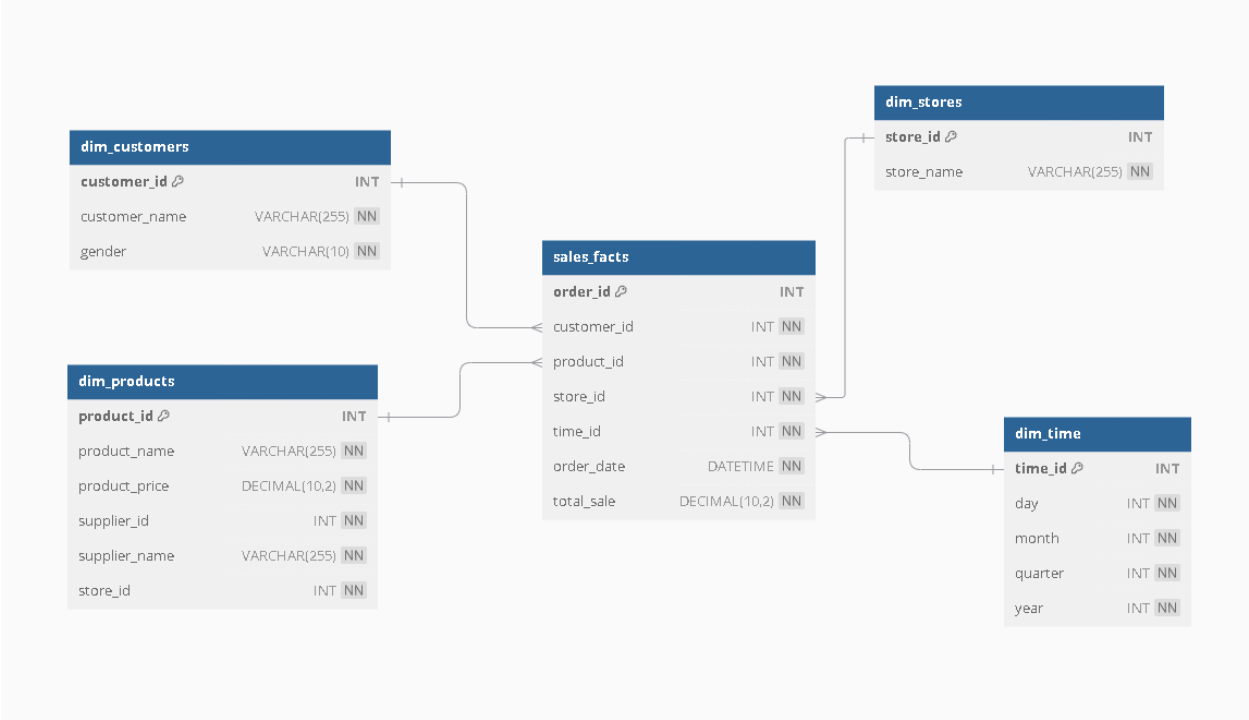
1. Project Report Overview:

This project is focused on building a **Data Warehouse (DW)** for a retail business, which involves the use of **Star Schema** to structure the data efficiently for reporting and analysis. The key objective is to analyze sales data using **OLAP queries** and to implement an ETL process for transforming and loading data into the warehouse. We also explore the **MESHJOIN algorithm** for optimizing the ETL process by processing large datasets in partitions, specifically using **disk-buffer**, **stream-buffer**, **hash-table**, and **queue** concepts.

The project aims to:

- Load sales data into a structured **star schema** data warehouse.
- Analyze sales, product, customer, and store data through SQL queries.
- Implement **MESHJOIN** to optimize data processing during ETL.

2. Data Warehouse Schema (Star Schema):



Fact Table:

- `sales_facts`: Contains transaction data (sales data) with foreign keys linking to the dimension tables.

Dimension Tables:

- `dim_customers`: Contains customer details (ID, name, gender).
- `dim_products`: Contains product details (ID, name, price, supplier).
- `dim_stores`: Contains store details (ID, name).
- `dim_time`: Contains time information (day, month, quarter, year).

3. MeshJoin Algorithm:

The key steps in **MESHJOIN**:

- Load a new chunk of customer transactions into the stream-buffer.
- Replace the product partition in the disk-buffer.
- Join customer transactions with the product partition.
- After each cycle, remove the oldest chunk of customer transactions from the hash table and queue.

How it works:

- **Disk-Buffer**: Loads product data in chunks (partitions) into memory, represented by `productPartitions`.
- **Stream-Buffer**: Holds chunks of customer transaction data in memory, represented by the `transactionQueue`.
- **Hash Table**: Stores customer transactions for quick lookup and joining with product data.
- **Queue**: Ensures each chunk of customer transactions is processed and joined with the entire product data before being removed in a FIFO order.

4. Shortcomings of Meshjoin:

- **Memory Consumption**: Despite the use of buffers, memory consumption can still be high for very large datasets. If the buffer size is not properly optimized, the system may experience performance degradation.
- **Complexity in Managing Buffers**: The cyclic nature of the buffer management can be complex, especially when the data partitions need to be dynamically adjusted. Mismanagement of buffer sizes can lead to data inconsistencies or slower processing.
- **Not Suitable for Small Datasets**: MESHJOIN is designed for large-scale data processing. For smaller datasets, traditional join methods (like hash joins) are more efficient. Therefore, using MESHJOIN on small data might introduce unnecessary overhead.

5. Learning from project:

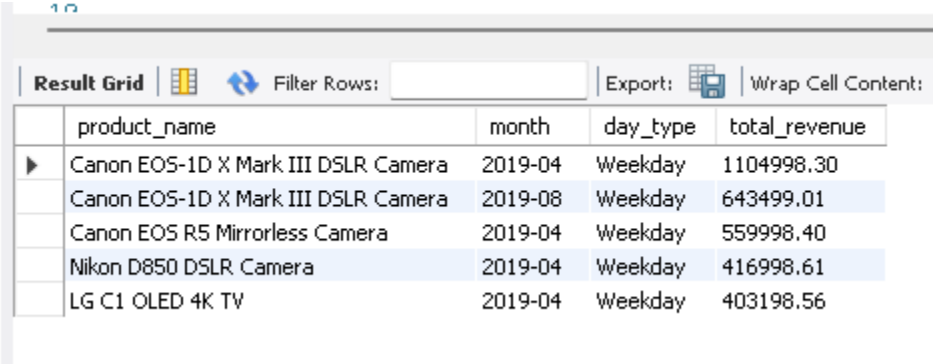
- The importance of **star schema** in organizing data efficiently for business intelligence and reporting.
- How to implement **OLAP queries** for complex data analysis, enabling meaningful insights from large transactional datasets.
- The **MESHJOIN algorithm** and its practical implementation to optimize data processing in ETL systems, especially for large-scale data in a multidimensional context.
- The key differences between traditional joins and **MESHJOIN** in terms of memory management and computational efficiency.

Additionally, this project improved my skills in:

- **SQL** for querying and data manipulation.
- **Java programming** for database connections and ETL implementation.
- Working with large datasets and optimizing performance through algorithms like MESHJOIN.

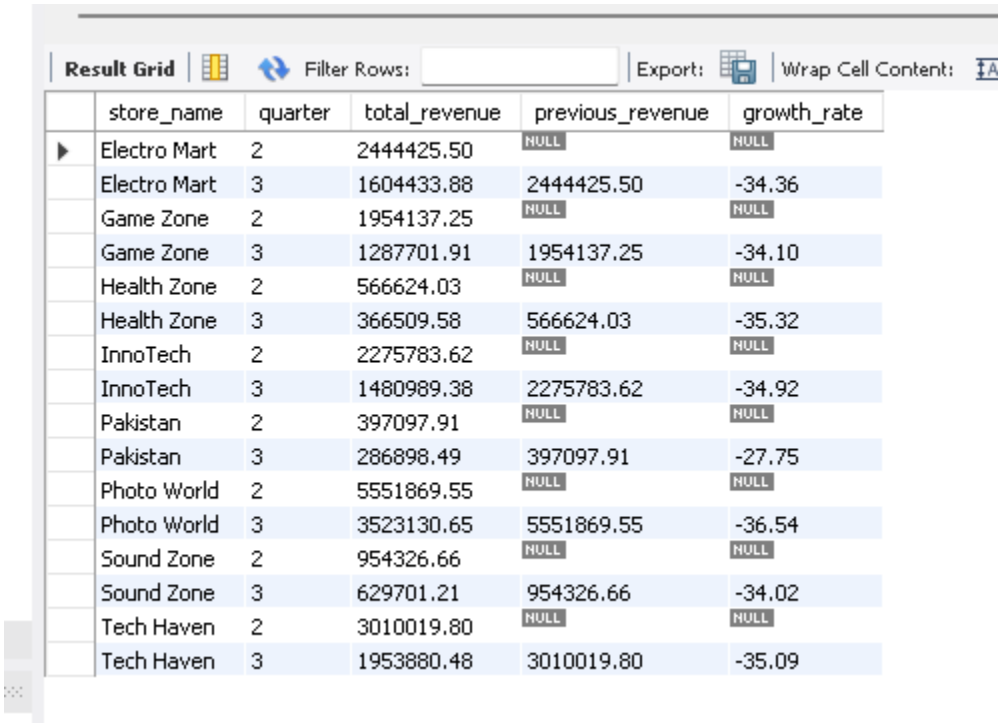
6. Olap Queries:

Q1


Result Grid | Filter Rows: | Export: | Wrap Cell Content:

product_name	month	day_type	total_revenue
Canon EOS-1D X Mark III DSLR Camera	2019-04	Weekday	1104998.30
Canon EOS-1D X Mark III DSLR Camera	2019-08	Weekday	643499.01
Canon EOS R5 Mirrorless Camera	2019-04	Weekday	559998.40
Nikon D850 DSLR Camera	2019-04	Weekday	416998.61
LG C1 OLED 4K TV	2019-04	Weekday	403198.56

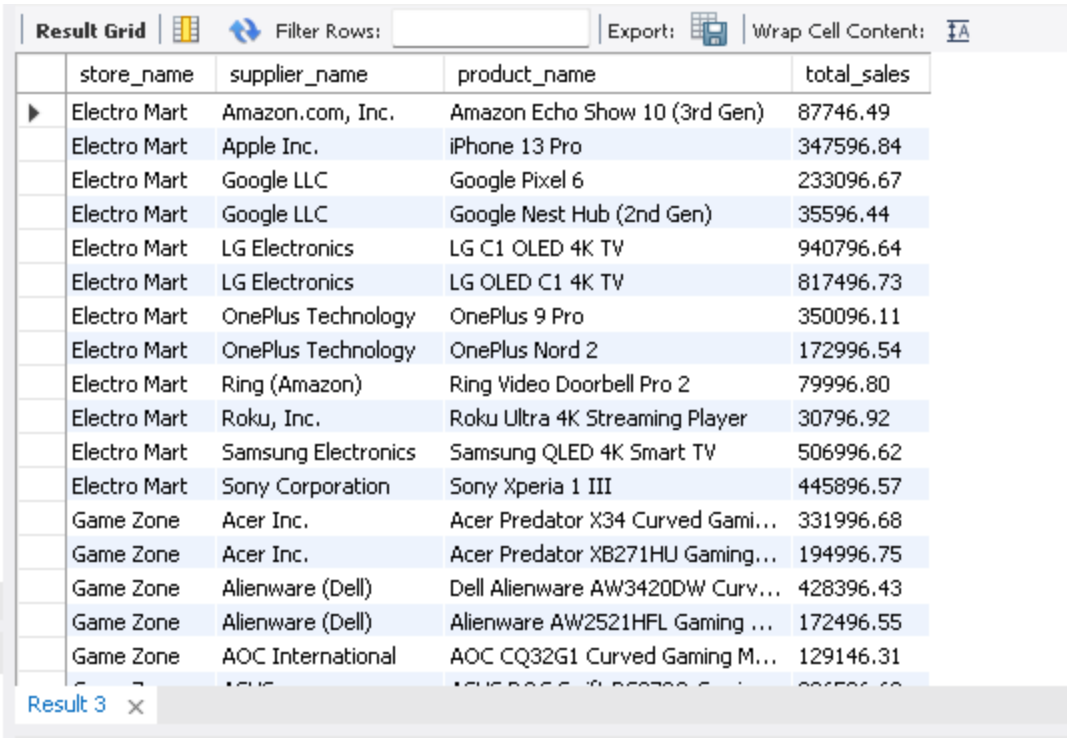
Q2



Result Grid | Filter Rows: | Export: | Wrap Cell Content:

store_name	quarter	total_revenue	previous_revenue	growth_rate
Electro Mart	2	2444425.50	NULL	NULL
Electro Mart	3	1604433.88	2444425.50	-34.36
Game Zone	2	1954137.25	NULL	NULL
Game Zone	3	1287701.91	1954137.25	-34.10
Health Zone	2	566624.03	NULL	NULL
Health Zone	3	366509.58	566624.03	-35.32
InnoTech	2	2275783.62	NULL	NULL
InnoTech	3	1480989.38	2275783.62	-34.92
Pakistan	2	397097.91	NULL	NULL
Pakistan	3	286898.49	397097.91	-27.75
Photo World	2	5551869.55	NULL	NULL
Photo World	3	3523130.65	5551869.55	-36.54
Sound Zone	2	954326.66	NULL	NULL
Sound Zone	3	629701.21	954326.66	-34.02
Tech Haven	2	3010019.80	NULL	NULL
Tech Haven	3	1953880.48	3010019.80	-35.09

Q3

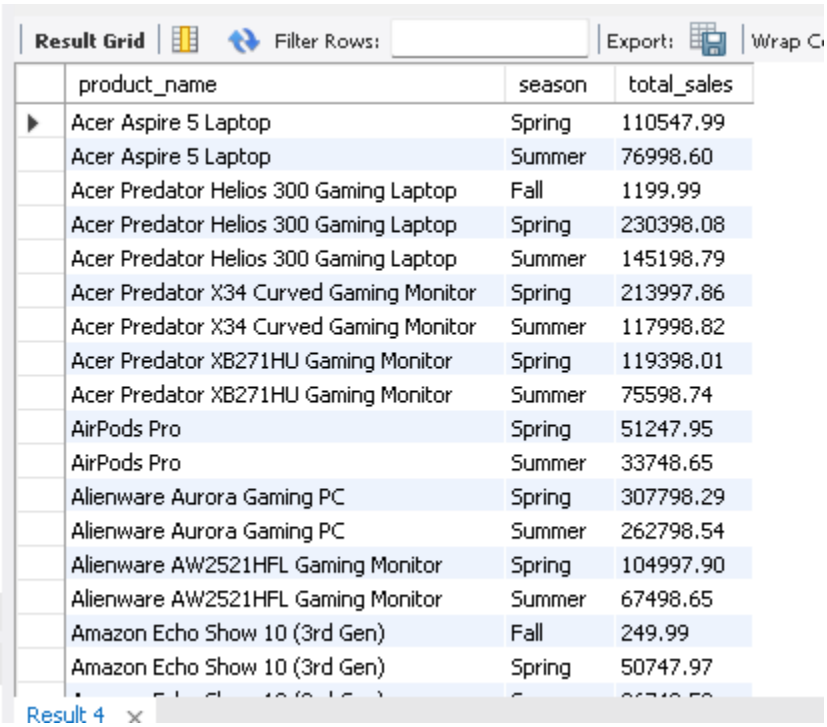


Result Grid | Filter Rows: | Export: | Wrap Cell Content:

store_name	supplier_name	product_name	total_sales
Electro Mart	Amazon.com, Inc.	Amazon Echo Show 10 (3rd Gen)	87746.49
Electro Mart	Apple Inc.	iPhone 13 Pro	347596.84
Electro Mart	Google LLC	Google Pixel 6	233096.67
Electro Mart	Google LLC	Google Nest Hub (2nd Gen)	35596.44
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 Technology	OnePlus 9 Pro	350096.11
Electro Mart	OnePlus Technology	OnePlus Nord 2	172996.54
Electro Mart	Ring (Amazon)	Ring Video Doorbell Pro 2	79996.80
Electro Mart	Roku, Inc.	Roku Ultra 4K Streaming Player	30796.92
Electro Mart	Samsung Electronics	Samsung QLED 4K Smart TV	506996.62
Electro Mart	Sony Corporation	Sony Xperia 1 III	445896.57
Game Zone	Acer Inc.	Acer Predator X34 Curved Gami...	331996.68
Game Zone	Acer Inc.	Acer Predator XB271HU Gaming...	194996.75
Game Zone	Alienware (Dell)	Dell Alienware AW3420DW Curv...	428396.43
Game Zone	Alienware (Dell)	Alienware AW2521HFL Gaming ...	172496.55
Game Zone	AOC International	AOC CQ32G1 Curved Gaming M...	129146.31

Result 3

Q4

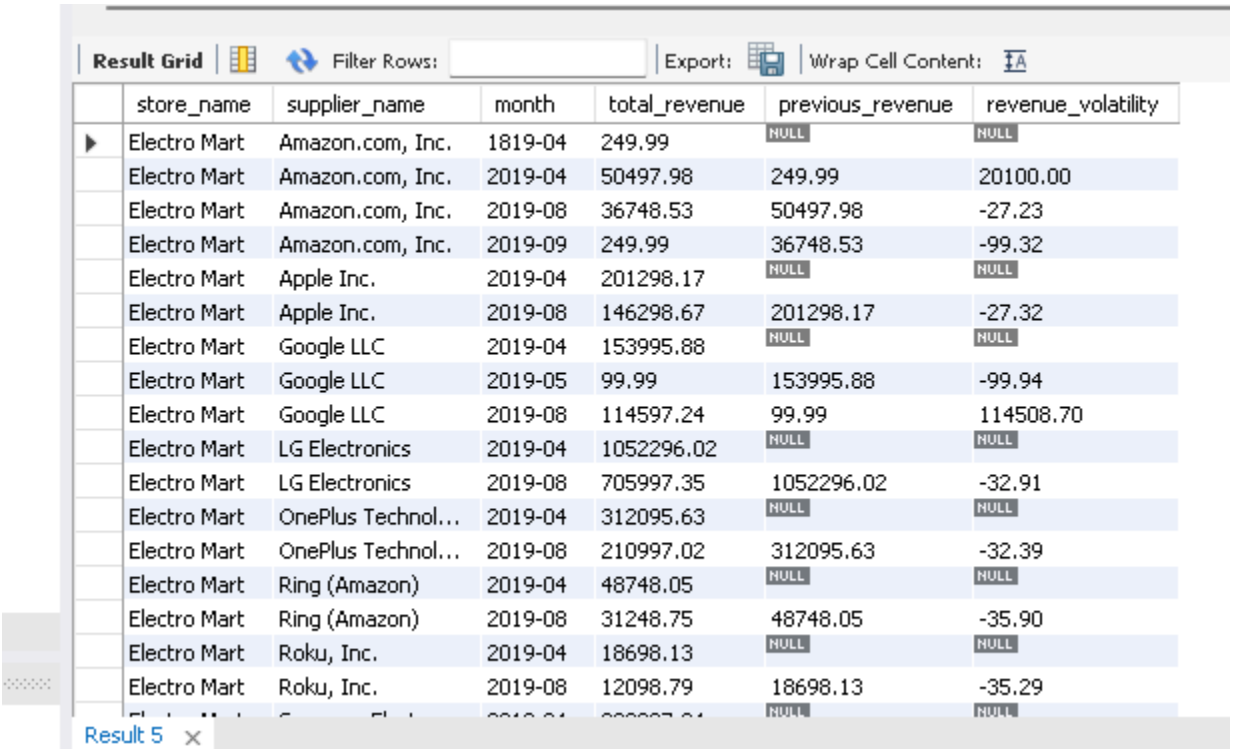


Result Grid | Filter Rows: | Export: | Wrap C

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	Fall	1199.99
Acer Predator Helios 300 Gaming Laptop	Spring	230398.08
Acer Predator Helios 300 Gaming Laptop	Summer	145198.79
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
Amazon Echo Show 10 (3rd Gen)	Fall	249.99
Amazon Echo Show 10 (3rd Gen)	Spring	50747.97

Result 4

Q5

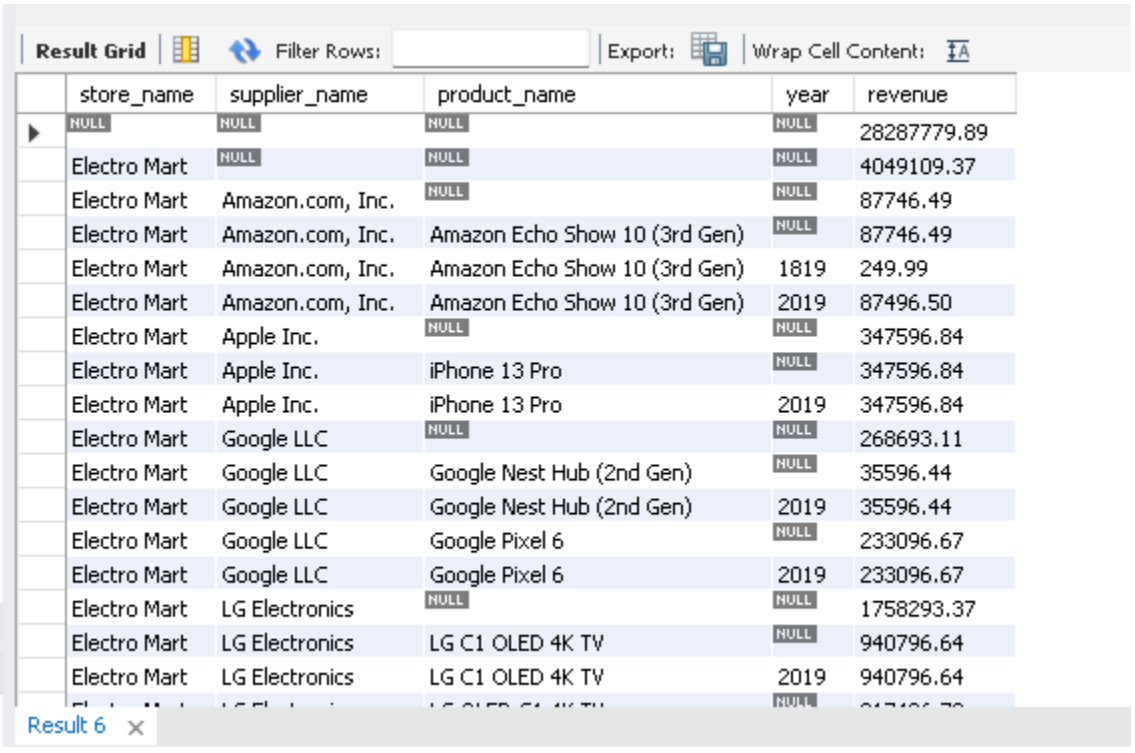


Result Grid | Filter Rows: | Export: | Wrap Cell Content:

store_name	supplier_name	month	total_revenue	previous_revenue	revenue_volatility
Electro Mart	Amazon.com, Inc.	1819-04	249.99	NULL	NULL
Electro Mart	Amazon.com, Inc.	2019-04	50497.98	249.99	20100.00
Electro Mart	Amazon.com, Inc.	2019-08	36748.53	50497.98	-27.23
Electro Mart	Amazon.com, Inc.	2019-09	249.99	36748.53	-99.32
Electro Mart	Apple Inc.	2019-04	201298.17	NULL	NULL
Electro Mart	Apple Inc.	2019-08	146298.67	201298.17	-27.32
Electro Mart	Google LLC	2019-04	153995.88	NULL	NULL
Electro Mart	Google LLC	2019-05	99.99	153995.88	-99.94
Electro Mart	Google LLC	2019-08	114597.24	99.99	114508.70
Electro Mart	LG Electronics	2019-04	1052296.02	NULL	NULL
Electro Mart	LG Electronics	2019-08	705997.35	1052296.02	-32.91
Electro Mart	OnePlus Technol...	2019-04	312095.63	NULL	NULL
Electro Mart	OnePlus Technol...	2019-08	210997.02	312095.63	-32.39
Electro Mart	Ring (Amazon)	2019-04	48748.05	NULL	NULL
Electro Mart	Ring (Amazon)	2019-08	31248.75	48748.05	-35.90
Electro Mart	Roku, Inc.	2019-04	18698.13	NULL	NULL
Electro Mart	Roku, Inc.	2019-08	12098.79	18698.13	-35.29

Result 5

Q7

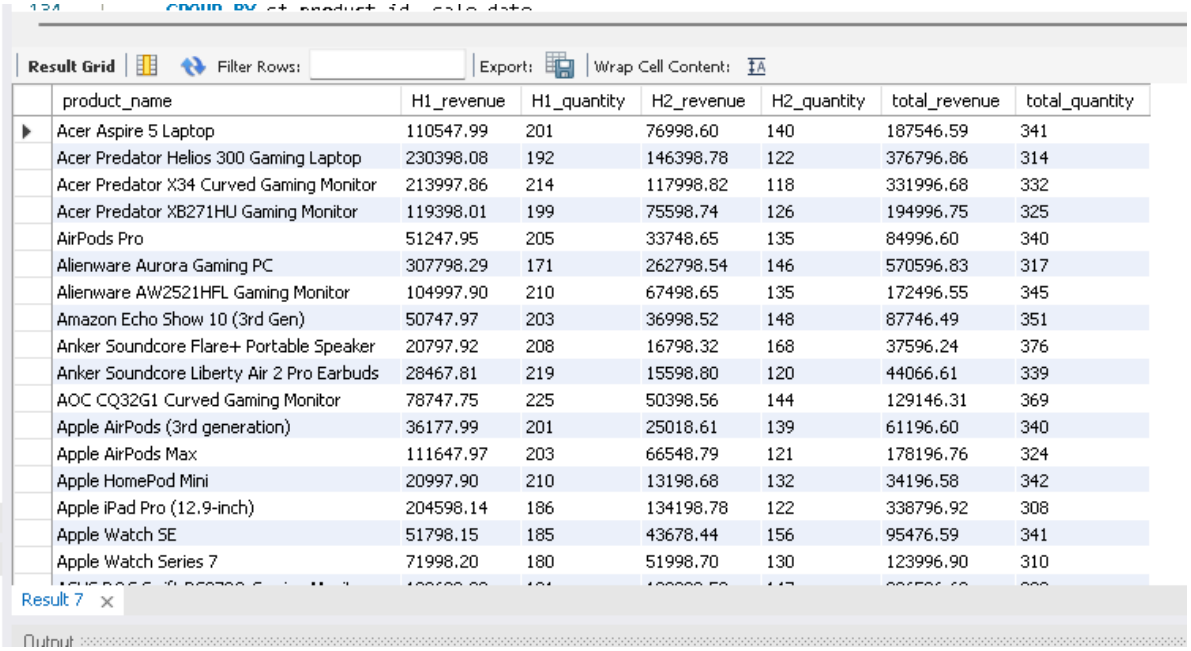


Result Grid | Filter Rows: | Export: | Wrap Cell Content:

store_name	supplier_name	product_name	year	revenue
NULL	NULL	NULL	NULL	28287779.89
Electro Mart	NULL	NULL	NULL	4049109.37
Electro Mart	Amazon.com, Inc.	NULL	NULL	87746.49
Electro Mart	Amazon.com, Inc.	Amazon Echo Show 10 (3rd Gen)	NULL	87746.49
Electro Mart	Amazon.com, Inc.	Amazon Echo Show 10 (3rd Gen)	1819	249.99
Electro Mart	Amazon.com, Inc.	Amazon Echo Show 10 (3rd Gen)	2019	87496.50
Electro Mart	Apple Inc.	NULL	NULL	347596.84
Electro Mart	Apple Inc.	iPhone 13 Pro	NULL	347596.84
Electro Mart	Apple Inc.	iPhone 13 Pro	2019	347596.84
Electro Mart	Google LLC	NULL	NULL	268693.11
Electro Mart	Google LLC	Google Nest Hub (2nd Gen)	NULL	35596.44
Electro Mart	Google LLC	Google Nest Hub (2nd Gen)	2019	35596.44
Electro Mart	Google LLC	Google Pixel 6	NULL	233096.67
Electro Mart	Google LLC	Google Pixel 6	2019	233096.67
Electro Mart	LG Electronics	NULL	NULL	1758293.37
Electro Mart	LG Electronics	LG C1 OLED 4K TV	NULL	940796.64
Electro Mart	LG Electronics	LG C1 OLED 4K TV	2019	940796.64

Result 6

Q8



Result Grid | Filter Rows: | Export: | Wrap Cell Content:

product_name	H1_revenue	H1_quantity	H2_revenue	H2_quantity	total_revenue	total_quantity
Acer Aspire 5 Laptop	110547.99	201	76998.60	140	187546.59	341
Acer Predator Helios 300 Gaming Laptop	230398.08	192	146398.78	122	376796.86	314
Acer Predator X34 Curved Gaming Monitor	213997.86	214	117998.82	118	331996.68	332
Acer Predator XB271HU Gaming Monitor	119398.01	199	75598.74	126	194996.75	325
AirPods Pro	51247.95	205	33748.65	135	84996.60	340
Alienware Aurora Gaming PC	307798.29	171	262798.54	146	570596.83	317
Alienware AW2521HFL Gaming Monitor	104997.90	210	67498.65	135	172496.55	345
Amazon Echo Show 10 (3rd Gen)	50747.97	203	36998.52	148	87746.49	351
Anker Soundcore Flare+ Portable Speaker	20797.92	208	16798.32	168	37596.24	376
Anker Soundcore Liberty Air 2 Pro Earbuds	28467.81	219	15598.80	120	44066.61	339
AOC CQ32G1 Curved Gaming Monitor	78747.75	225	50398.56	144	129146.31	369
Apple AirPods (3rd generation)	36177.99	201	25018.61	139	61196.60	340
Apple AirPods Max	111647.97	203	66548.79	121	178196.76	324
Apple HomePod Mini	20997.90	210	13198.68	132	34196.58	342
Apple iPad Pro (12.9-inch)	204598.14	186	134198.78	122	338796.92	308
Apple Watch SE	51798.15	185	43678.44	156	95476.59	341
Apple Watch Series 7	71998.20	180	51998.70	130	123996.90	310

Result 7

Output

Q9

159

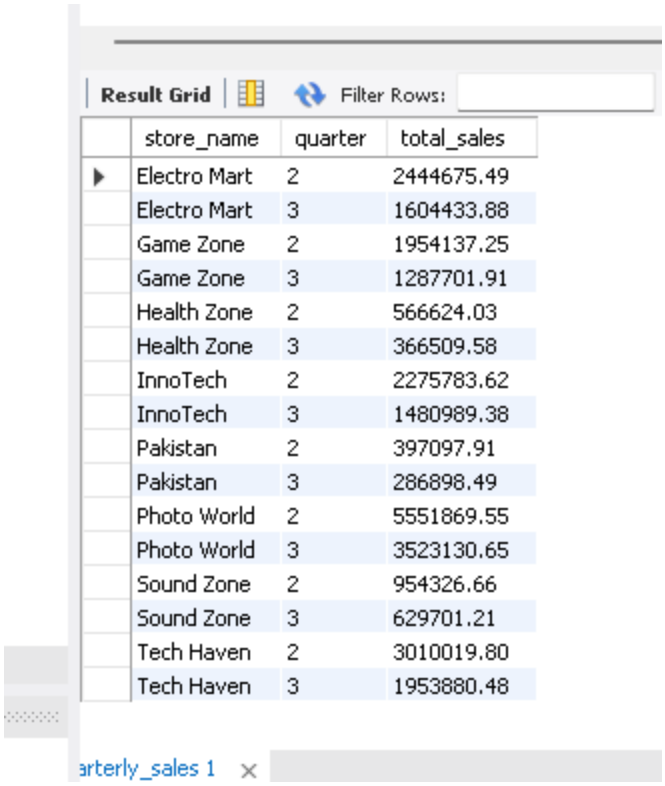
Question 10

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

product_id	product_name	sale_date	daily_sales	avg_sales	anomaly
1	iPhone 13 Pro	2019-04-18	12099.89	5698.308852	Spike
1	iPhone 13 Pro	2019-04-29	14299.87	5698.308852	Spike
1	iPhone 13 Pro	2019-08-27	13199.88	5698.308852	Spike
2	Dell XPS 13 Laptop	2019-04-20	18199.86	6563.883934	Spike
2	Dell XPS 13 Laptop	2019-04-23	15599.88	6563.883934	Spike
2	Dell XPS 13 Laptop	2019-04-26	20799.84	6563.883934	Spike
2	Dell XPS 13 Laptop	2019-04-29	14299.89	6563.883934	Spike
3	Samsung QLED 4K Smart TV	2019-04-13	19499.87	8311.420000	Spike
3	Samsung QLED 4K Smart TV	2019-04-16	17999.88	8311.420000	Spike
4	Sony WH-1000XM4 Headphones	2019-04-11	4549.87	2208.953279	Spike
5	iPad Air	2019-04-29	7799.87	3459.942333	Spike
6	Xbox Series X	2019-04-13	8499.83	2703.335763	Spike
6	Xbox Series X	2019-04-25	6499.87	2703.335763	Spike
7	AirPods Pro	2019-04-04	3499.86	1393.386885	Spike
8	Alienware Aurora Gaming PC	2019-08-30	25199.86	9203.174677	Spike
9	Canon EOS R5 Mirrorless Camera	2019-04-13	48999.86	21844.765172	Spike
10	Fitbit Charge 5	2019-04-17	2159.88	920.604590	Spike

Result 8

Q10



7. Conclusion:

This project helped me understand how data warehouses are designed, implemented, and used for business analytics. The **Star Schema** architecture enables fast querying of large datasets, and the **MESHJOIN algorithm** provides a way to efficiently handle huge amounts of transactional data during the ETL process. By implementing these concepts, I've gained a deeper understanding of data warehousing and optimization techniques..