DS3003 & DS3004: Data Warehousing & Business Intelligence

Project Report

Building a Near-Real-Time Data Warehouse for Walmart

Submitted by:

Maryam Khalid

22i-1917

Fall 2025

November 17, 2025

1 Project Overview

This project builds a near-real-time data warehouse for Walmart to analyze customer purchasing behavior quickly. The main goal was to help the company make faster business decisions by processing transaction data as it arrives.

The biggest challenge was joining streaming transaction data with master data containing customer and product information. I solved this using the HYBRIDJOIN algorithm in Python, which efficiently combines these two data sources without overwhelming system memory.

Results achieved:

- Processed 550,068 transactions in 2 minutes
- Created a star schema with 5 dimensions and 1 fact table
- Built 20 analytical queries for business insights

2 Data Warehouse Schema

2.1 Star Schema Design

I used a star schema because it's simple and fast. It has one central fact table (FactSales) connected to five dimension tables that describe customers, products, stores, suppliers, and dates.

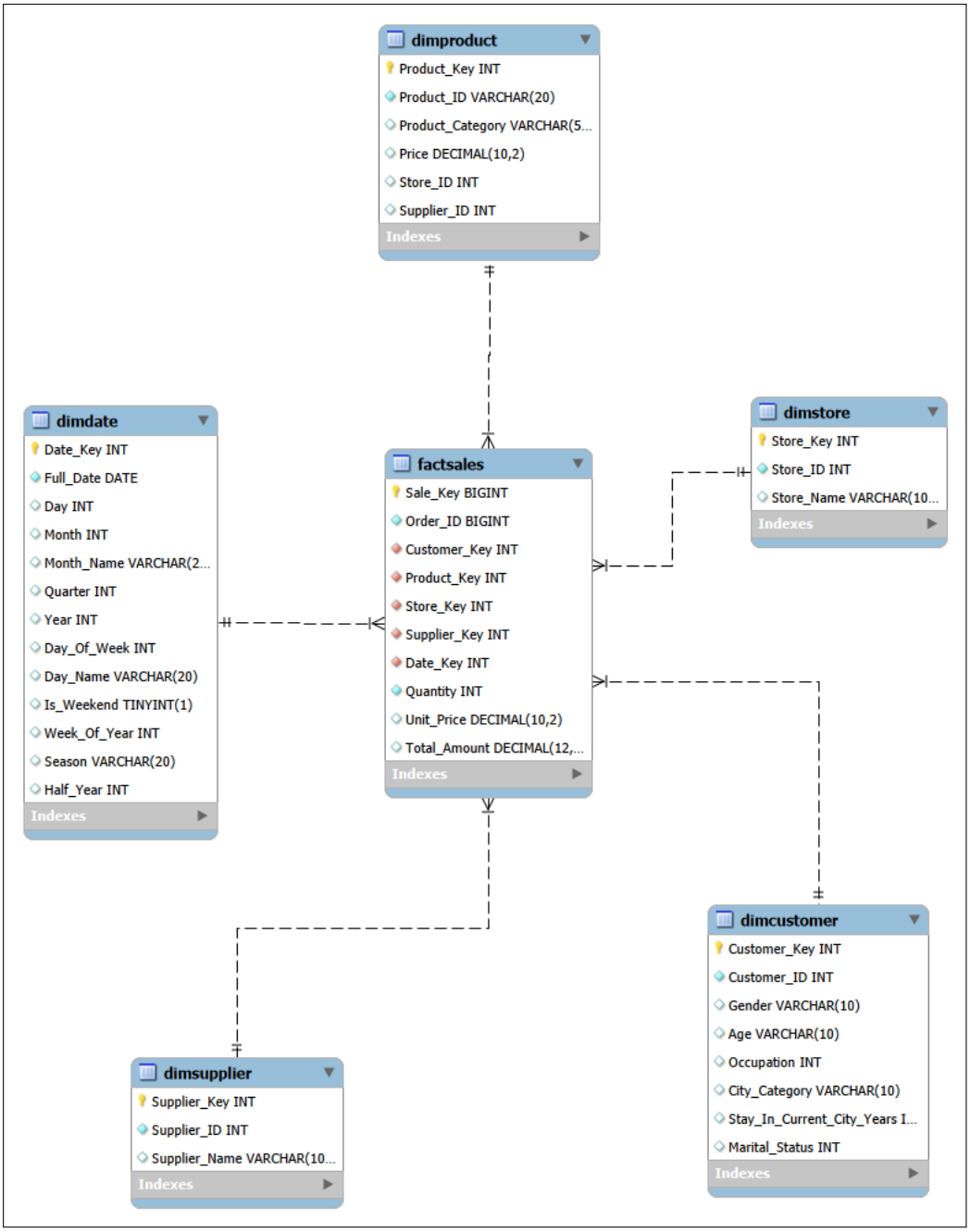


Figure 1: Star Schema for Walmart Data Warehouse

2.2 Tables in the Schema

Dimension Tables:

- **DimCustomer (5,891 records):** Gender, age, occupation, city type, marital status
- **DimProduct (3,631 records):** Product category, price, store, supplier
- **DimStore (8 records):** Store names and locations
- **DimSupplier (7 records):** Supplier information
- **DimDate (2,557 records):** Year, quarter, month, weekend flag, season

Fact Table:

- **FactSales (550,068 records):** Order ID, quantity, unit price, total amount
- Contains foreign keys linking to all five dimensions

Why this design? Star schema makes queries simple because you only join the fact table with dimensions—no complicated multi-level joins. It's also faster and easier for business users to understand.

3 HYBRIDJOIN Algorithm

3.1 How It Works

HYBRIDJOIN joins streaming data with disk-based master data in real-time. Instead of waiting for all transactions to arrive, it processes them as they come in.

The algorithm uses four main parts:

- **Hash Table:** Stores 10,000 transactions in memory using Product_ID as the key
- **Queue:** Keeps track of which products need to be looked up, in order
- **Stream Buffer:** Holds extra transactions when the hash table is full
- **Disk Buffer:** Loads product information from master data (500 records at a time)

3.2 Algorithm Steps

Step 1: Set up empty hash table, queue, and buffers.

Step 2: Load transactions from stream buffer into hash table. Hash each one by Product_ID and add to queue.

Step 3: Take the oldest Product_ID from queue and load its master data from disk.

Step 4: Match transactions in hash table with the loaded product data. Combine them and remove matched records.

Step 5: Repeat until all transactions are processed.

3.3 Performance

In my implementation, the algorithm processed 550,068 transactions in about 2 seconds at a rate of 278,738 records per second. This is fast enough for real-time business analytics.

4 Three Shortcomings of HYBRIDJOIN

4.1 1. Fixed Memory Size

The hash table only has 10,000 slots. During heavy traffic periods like Black Friday, transactions can arrive faster than we process them. The hash table fills up and the stream buffer keeps growing, which could crash the system. A better approach would be adaptive memory that expands when needed.

4.2 2. Sequential Processing

The algorithm loads product data one at a time from disk. If 100 different products arrive together, it processes them sequentially instead of in parallel. This wastes time because modern computers can read from disk simultaneously. Multi-threaded loading would be much faster.

4.3 3. Missing Product Data

When a transaction references a product that doesn't exist in master data, the algorithm just discards it. In real stores, new products are added daily. If someone buys a product before its details are uploaded, that sale is lost. We need a retry mechanism or temporary storage for unmatched transactions.

5 Lessons Learned

5.1 Technical Skills Gained

Database skills:

- Designed star schemas for analytics
- Used surrogate keys for better performance
- Applied indexing strategies

Programming skills:

- Implemented hash tables and queues in Python
- Processed large datasets efficiently with pandas
- Wrote complex SQL queries with window functions

5.2 Challenges

My biggest challenge was performance. The first version took over 50 minutes! I improved it to 2 minutes by using larger batch sizes (20,000 instead of 5,000) and bulk inserts. I also had to fix data type issues between Python and MySQL.

Another challenge was understanding the difference between OLAP (analytics) and OLTP (transactions). Data warehouses use denormalized structures and heavy indexing, which is opposite to regular databases.

I also learned to adapt queries for historical data. Four queries used current dates but our data only goes to 2020, so I modified them accordingly.

5.3 Real-World Applications

This project showed me how companies actually use data warehouses:

- Amazon uses similar systems for product recommendations
- Walmart uses it for dynamic pricing
- Banks use it for fraud detection

The star schema design is standard in tools like Tableau and Power BI. Understanding it is essential for data analyst roles.

5.4 Key Takeaways

The most important lesson was that data warehousing is about helping people make decisions, not just storing data. Every design choice supports that goal.

I also realized that technical skills alone aren't enough. You need to understand the business context. Why track occupation? To predict buying behavior. Why separate weekday and weekend sales? Because shopping patterns differ.

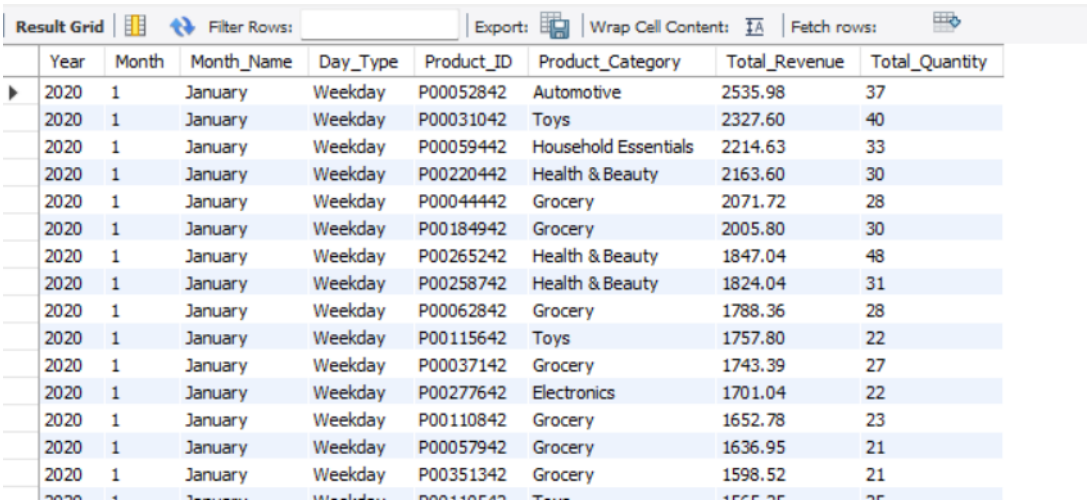
Finally, I learned that real systems are messy with data quality issues and missing records. This clean academic project taught me principles I'll use to handle real-world challenges.

6 OLAP Query Results

This section presents the results of 20 analytical queries executed on the data warehouse. Each query demonstrates different OLAP operations such as slicing, dicing, drill-down, and roll-up.

6.1 Q1: Top Revenue-Generating Products (Weekdays vs Weekends)

This query identifies top products by revenue, split by weekdays and weekends with monthly drill-down for 2020.



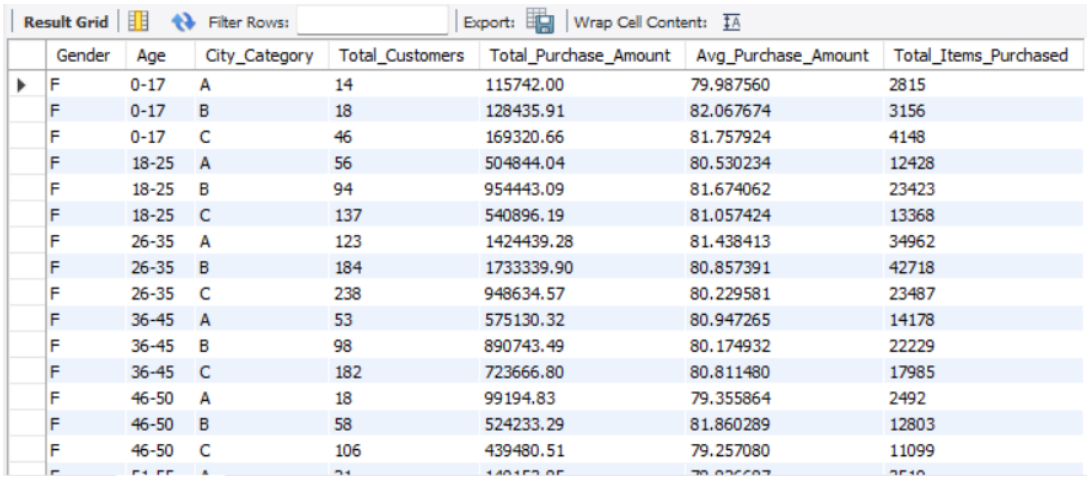
Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

Year	Month	Month_Name	Day_Type	Product_ID	Product_Category	Total_Revenue	Total_Quantity
2020	1	January	Weekday	P00052842	Automotive	2535.98	37
2020	1	January	Weekday	P00031042	Toys	2327.60	40
2020	1	January	Weekday	P00059442	Household Essentials	2214.63	33
2020	1	January	Weekday	P00220442	Health & Beauty	2163.60	30
2020	1	January	Weekday	P00044442	Grocery	2071.72	28
2020	1	January	Weekday	P00184942	Grocery	2005.80	30
2020	1	January	Weekday	P00265242	Health & Beauty	1847.04	48
2020	1	January	Weekday	P00258742	Health & Beauty	1824.04	31
2020	1	January	Weekday	P00062842	Grocery	1788.36	28
2020	1	January	Weekday	P00115642	Toys	1757.80	22
2020	1	January	Weekday	P00037142	Grocery	1743.39	27
2020	1	January	Weekday	P00277642	Electronics	1701.04	22
2020	1	January	Weekday	P00110842	Grocery	1652.78	23
2020	1	January	Weekday	P00057942	Grocery	1636.95	21
2020	1	January	Weekday	P00351342	Grocery	1598.52	21

Figure 2: Q1 Results: Top Products by Day Type

6.2 Q2: Customer Demographics by Purchase Amount

Analyzes total purchase amounts by gender, age, and city category.



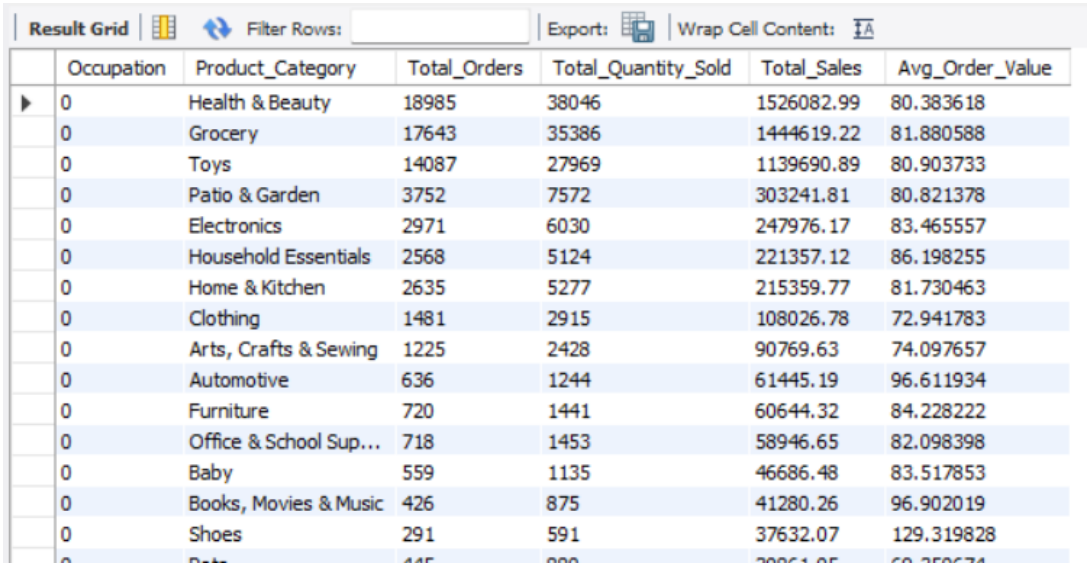
Result Grid | Filter Rows: | Export: | Wrap Cell Content:

Gender	Age	City_Category	Total_Customers	Total_Purchase_Amount	Avg_Purchase_Amount	Total_Items_Purchased
F	0-17	A	14	115742.00	79.987560	2815
F	0-17	B	18	128435.91	82.067674	3156
F	0-17	C	46	169320.66	81.757924	4148
F	18-25	A	56	504844.04	80.530234	12428
F	18-25	B	94	954443.09	81.674062	23423
F	18-25	C	137	540896.19	81.057424	13368
F	26-35	A	123	1424439.28	81.438413	34962
F	26-35	B	184	1733339.90	80.857391	42718
F	26-35	C	238	948634.57	80.229581	23487
F	36-45	A	53	575130.32	80.947265	14178
F	36-45	B	98	890743.49	80.174932	22229
F	36-45	C	182	723666.80	80.811480	17985
F	46-50	A	18	99194.83	79.355864	2492
F	46-50	B	58	524233.29	81.860289	12803
F	46-50	C	106	439480.51	79.257080	11099

Figure 3: Q2 Results: Customer Demographics Analysis

6.3 Q3: Product Category Sales by Occupation

Examines total sales for each product category based on customer occupation.



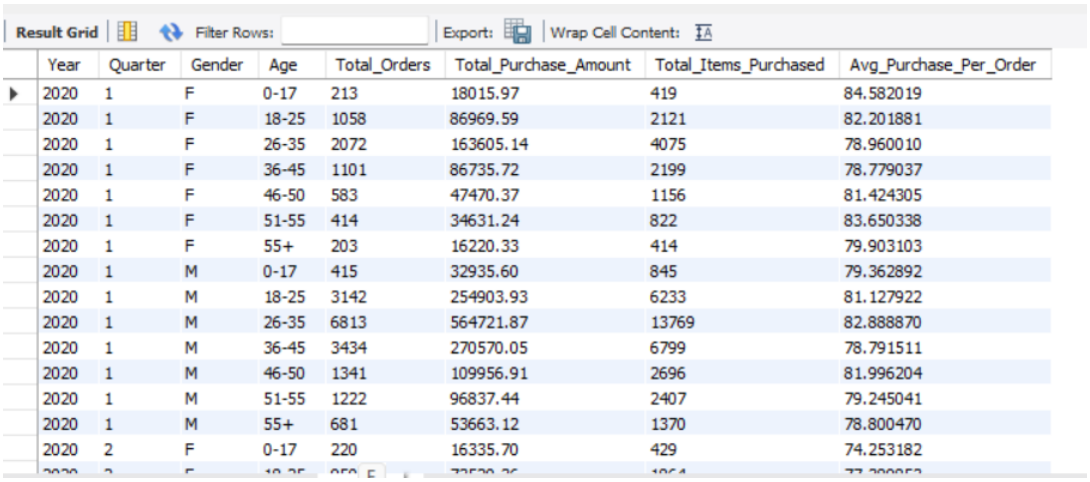
Result Grid | Filter Rows: | Export: | Wrap Cell Content:

Occupation	Product_Category	Total_Orders	Total_Quantity_Sold	Total_Sales	Avg_Order_Value
0	Health & Beauty	18985	38046	1526082.99	80.383618
0	Grocery	17643	35386	1444619.22	81.880588
0	Toys	14087	27969	1139690.89	80.903733
0	Patio & Garden	3752	7572	303241.81	80.821378
0	Electronics	2971	6030	247976.17	83.465557
0	Household Essentials	2568	5124	221357.12	86.198255
0	Home & Kitchen	2635	5277	215359.77	81.730463
0	Clothing	1481	2915	108026.78	72.941783
0	Arts, Crafts & Sewing	1225	2428	90769.63	74.097657
0	Automotive	636	1244	61445.19	96.611934
0	Furniture	720	1441	60644.32	84.228222
0	Office & School Sup...	718	1453	58946.65	82.098398
0	Baby	559	1135	46686.48	83.517853
0	Books, Movies & Music	426	875	41280.26	96.902019
0	Shoes	291	591	37632.07	129.319828

Figure 4: Q3 Results: Category Sales by Occupation

6.4 Q4: Total Purchases by Gender and Age (Quarterly Trend)

Tracks purchase amounts by gender and age across quarterly periods for 2020.



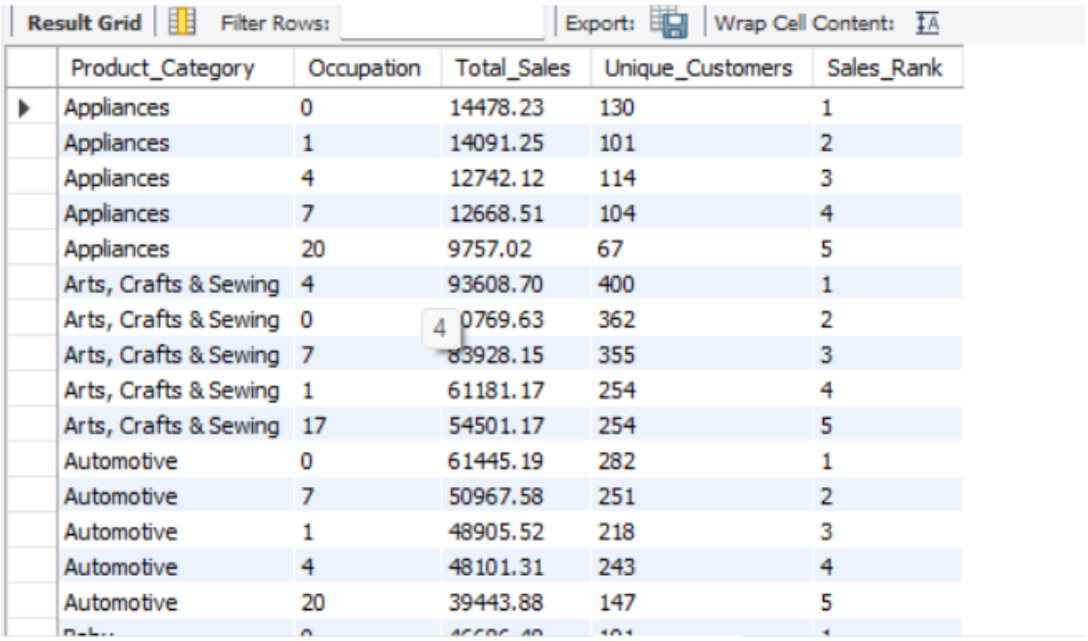
Result Grid | Filter Rows: | Export: | Wrap Cell Content:

Year	Quarter	Gender	Age	Total_Orders	Total_Purchase_Amount	Total_Items_Purchased	Avg_Purchase_Per_Order
2020	1	F	0-17	213	18015.97	419	84.582019
2020	1	F	18-25	1058	86969.59	2121	82.201881
2020	1	F	26-35	2072	163605.14	4075	78.960010
2020	1	F	36-45	1101	86735.72	2199	78.779037
2020	1	F	46-50	583	47470.37	1156	81.424305
2020	1	F	51-55	414	34631.24	822	83.650338
2020	1	F	55+	203	16220.33	414	79.903103
2020	1	M	0-17	415	32935.60	845	79.362892
2020	1	M	18-25	3142	254903.93	6233	81.127922
2020	1	M	26-35	6813	564721.87	13769	82.888870
2020	1	M	36-45	3434	270570.05	6799	78.791511
2020	1	M	46-50	1341	109956.91	2696	81.996204
2020	1	M	51-55	1222	96837.44	2407	79.245041
2020	1	M	55+	681	53663.12	1370	78.800470
2020	2	F	0-17	220	16335.70	429	74.253182

Figure 5: Q4 Results: Quarterly Purchase Trends

6.5 Q5: Top Occupations by Product Category Sales

Highlights the top 5 occupations driving sales within each product category.



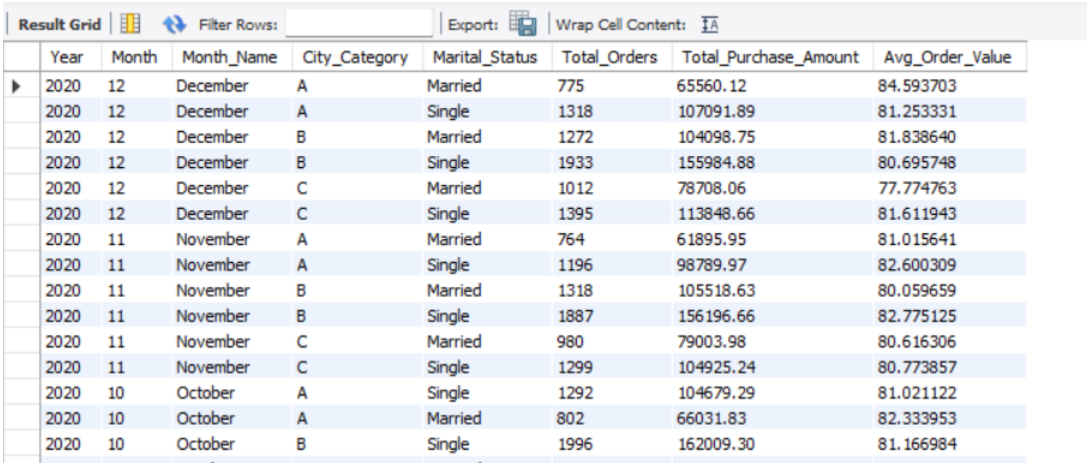
Result Grid | Filter Rows: | Export: | Wrap Cell Content:

Product_Category	Occupation	Total_Sales	Unique_Customers	Sales_Rank
Appliances	0	14478.23	130	1
Appliances	1	14091.25	101	2
Appliances	4	12742.12	114	3
Appliances	7	12668.51	104	4
Appliances	20	9757.02	67	5
Arts, Crafts & Sewing	4	93608.70	400	1
Arts, Crafts & Sewing	0	0769.63	362	2
Arts, Crafts & Sewing	7	83928.15	355	3
Arts, Crafts & Sewing	1	61181.17	254	4
Arts, Crafts & Sewing	17	54501.17	254	5
Automotive	0	61445.19	282	1
Automotive	7	50967.58	251	2
Automotive	1	48905.52	218	3
Automotive	4	48101.31	243	4
Automotive	20	39443.88	147	5

Figure 6: Q5 Results: Top Occupations per Category

6.6 Q6: City Category Performance by Marital Status

Assesses purchase amounts by city category and marital status over the last 6 months of 2020.



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

Year	Month	Month_Name	City_Category	Marital_Status	Total_Orders	Total_Purchase_Amount	Avg_Order_Value
2020	12	December	A	Married	775	65560.12	84.593703
2020	12	December	A	Single	1318	107091.89	81.253331
2020	12	December	B	Married	1272	104098.75	81.838640
2020	12	December	B	Single	1933	155984.88	80.695748
2020	12	December	C	Married	1012	78708.06	77.774763
2020	12	December	C	Single	1395	113848.66	81.611943
2020	11	November	A	Married	764	61895.95	81.015641
2020	11	November	A	Single	1196	98789.97	82.600309
2020	11	November	B	Married	1318	105518.63	80.059659
2020	11	November	B	Single	1887	156196.66	82.775125
2020	11	November	C	Married	980	79003.98	80.616306
2020	11	November	C	Single	1299	104925.24	80.773857
2020	10	October	A	Single	1292	104679.29	81.021122
2020	10	October	A	Married	802	66031.83	82.333953
2020	10	October	B	Single	1996	162009.30	81.166984

Figure 7: Q6 Results: City Performance by Marital Status

6.7 Q7: Average Purchase Amount by Stay Duration and Gender

Calculates average purchase amount based on years stayed in the city and gender.

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

Stay_In_Current_City_Years	Gender	Total_Customers	Total_Orders	Total_Purchase_Amount	Avg_Purchase_Amount	Total_Items_Purchased
0	F	214	17063	1392633.23	81.617138	34037
0	M	558	57335	4641121.63	80.947443	114490
1	F	604	51298	4139504.21	80.695236	102446
1	M	1482	142523	11518117.71	80.815852	284836
2	F	328	24332	1987244.85	81.672072	48744
2	M	817	77506	6281980.92	81.051543	155183
3	F	286	24520	1980996.44	80.791046	49015
3	M	693	70765	5719664.59	80.826179	141326
4	F	234	18596	1484191.42	79.812402	37021
4	M	675	66130	5362264.58	81.086717	132124

Figure 8: Q7 Results: Purchase Amount by Stay Duration

6.8 Q8: Top 5 Revenue-Generating Cities by Product Category

Ranks the top 5 city categories by revenue, grouped by product category.

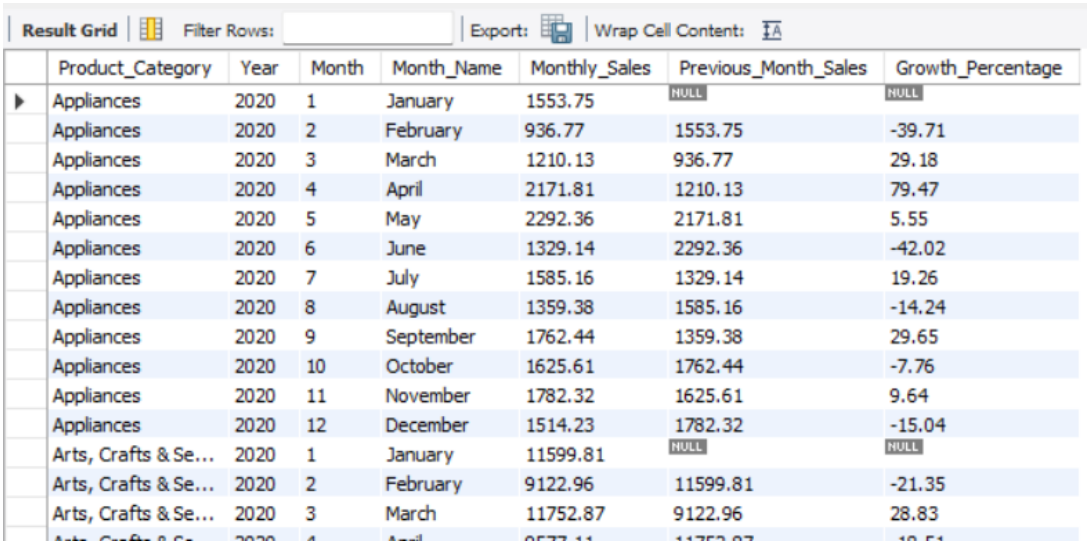
Result Grid | Filter Rows: | Export: | Wrap Cell Content:

Product_Category	City_Category	Total_Revenue	Total_Orders	Revenue_Rank
Appliances	B	49395.82	632	1
Appliances	A	36902.11	481	2
Appliances	C	32501.21	410	3
Arts, Crafts & Sewing	B	310595.71	4038	1
Arts, Crafts & Sewing	A	222810.59	2848	2
Arts, Crafts & Sewing	C	218588.25	2942	3
Automotive	B	199054.82	2063	1
Automotive	C	167970.73	1729	2
Automotive	A	128189.51	1333	3
Baby	B	132161.88	1599	1
Baby	A	104133.30	1226	2
Baby	C	74901.51	896	3
Books, Movies & M...	B	131015.16	1389	1
Books, Movies & M...	C	98138.91	983	2
Books, Movies & M...	A	72728.62	753	3

Figure 9: Q8 Results: Top Cities per Category

6.9 Q9: Monthly Sales Growth by Product Category

Measures month-over-month sales growth percentage for each product category in 2020.



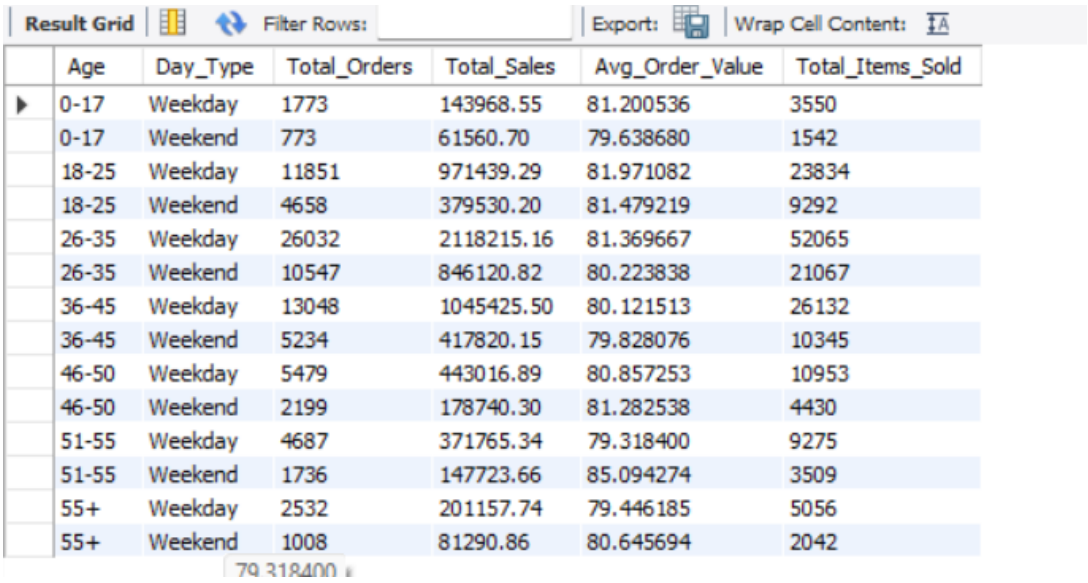
Result Grid | Filter Rows: | Export: | Wrap Cell Content:

Product_Category	Year	Month	Month_Name	Monthly_Sales	Previous_Month_Sales	Growth_Percentage
Appliances	2020	1	January	1553.75	NULL	NULL
Appliances	2020	2	February	936.77	1553.75	-39.71
Appliances	2020	3	March	1210.13	936.77	29.18
Appliances	2020	4	April	2171.81	1210.13	79.47
Appliances	2020	5	May	2292.36	2171.81	5.55
Appliances	2020	6	June	1329.14	2292.36	-42.02
Appliances	2020	7	July	1585.16	1329.14	19.26
Appliances	2020	8	August	1359.38	1585.16	-14.24
Appliances	2020	9	September	1762.44	1359.38	29.65
Appliances	2020	10	October	1625.61	1762.44	-7.76
Appliances	2020	11	November	1782.32	1625.61	9.64
Appliances	2020	12	December	1514.23	1782.32	-15.04
Arts, Crafts & Se...	2020	1	January	11599.81	NULL	NULL
Arts, Crafts & Se...	2020	2	February	9122.96	11599.81	-21.35
Arts, Crafts & Se...	2020	3	March	11752.87	9122.96	28.83

Figure 10: Q9 Results: Monthly Sales Growth Rates

6.10 Q10: Weekend vs Weekday Sales by Age Group

Compares total sales by age group for weekends versus weekdays in 2020.



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

Age	Day_Type	Total_Orders	Total_Sales	Avg_Order_Value	Total_Items_Sold
0-17	Weekday	1773	143968.55	81.200536	3550
0-17	Weekend	773	61560.70	79.638680	1542
18-25	Weekday	11851	971439.29	81.971082	23834
18-25	Weekend	4658	379530.20	81.479219	9292
26-35	Weekday	26032	2118215.16	81.369667	52065
26-35	Weekend	10547	846120.82	80.223838	21067
36-45	Weekday	13048	1045425.50	80.121513	26132
36-45	Weekend	5234	417820.15	79.828076	10345
46-50	Weekday	5479	443016.89	80.857253	10953
46-50	Weekend	2199	178740.30	81.282538	4430
51-55	Weekday	4687	371765.34	79.318400	9275
51-55	Weekend	1736	147723.66	85.094274	3509
55+	Weekday	2532	201157.74	79.446185	5056
55+	Weekend	1008	81290.86	80.645694	2042

79.318400

Figure 11: Q10 Results: Weekend vs Weekday Comparison

6.11 Q11: Top Products by Revenue (Monthly Drill-Down)

Top 5 products by revenue separated by weekday/weekend with monthly breakdown for 2020.

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

Year	Month	Month_Name	Day_Type	Product_ID	Product_Category	Total_Revenue	Revenue_Ra
2020	1	January	Weekday	P00052842	Automotive	2535.98	1
2020	1	January	Weekday	P00031042	Toys	2327.60	2
2020	1	January	Weekday	P00059442	Household Essentials	2214.63	3
2020	1	January	Weekday	P00220442	Health & Beauty	2163.60	4
2020	1	January	Weekday	P00044442	Grocery	2071.72	5
2020	1	January	Weekend	P00251842	Grocery	1110.90	1
2020	1	January	Weekend	P00116842	Electronics	926.76	2
2020	1	January	Weekend	P00059442	Household Essentials	872.43	3
2020	1	January	Weekend	P00318742	Grocery	827.42	4
2020	1	January	Weekend	P00265242	Health & Beauty	808.08	5
2020	2	February	Weekday	P00086442	Toys	1963.23	1
2020	2	February	Weekday	P00275842	Electronics	1926.76	2
2020	2	February	Weekday	P00116842	Electronics	1776.29	3
2020	2	February	Weekday	P00051442	Toys	1611.60	4
2020	2	February	Weekday	P00059442	Household Essentials	1610.64	5

Figure 12: Q11 Results: Monthly Product Revenue Breakdown

6.12 Q12: Store Revenue Growth Rate (Quarterly 2017)

Calculates the revenue growth rate for each store on a quarterly basis for 2017.

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

Store_Name	Year	Quarter	Quarterly_Revenue	Previous_Quarter_Revenue	Growth_Rate_Percentage
Electro Mart	2017	1	266109.81	NULL	NULL
Electro Mart	2017	2	252927.52	266109.81	-4.95
Electro Mart	2017	3	251342.94	252927.52	-0.63
Electro Mart	2017	4	266426.90	251342.94	6.00
Game Zone	2017	1	369516.30	NULL	NULL
Game Zone	2017	2	374116.40	369516.30	1.24
Game Zone	2017	3	376589.49	374116.40	0.66
Game Zone	2017	4	395890.37	376589.49	5.13
Health Zone	2017	1	117430.18	NULL	NULL
Health Zone	2017	2	117741.63	117430.18	0.27
Health Zone	2017	3	117072.38	117741.63	-0.57
Health Zone	2017	4	119124.34	117072.38	1.75
InnoTech	2017	1	246292.59	NULL	NULL
InnoTech	2017	2	250089.39	246292.59	1.54
InnoTech	2017	3	247719.50	250089.39	-0.95

Figure 13: Q12 Results: Quarterly Store Growth Rates

6.13 Q13: Supplier Sales Contribution by Store and Product

Shows total sales contribution of each supplier broken down by product and store.

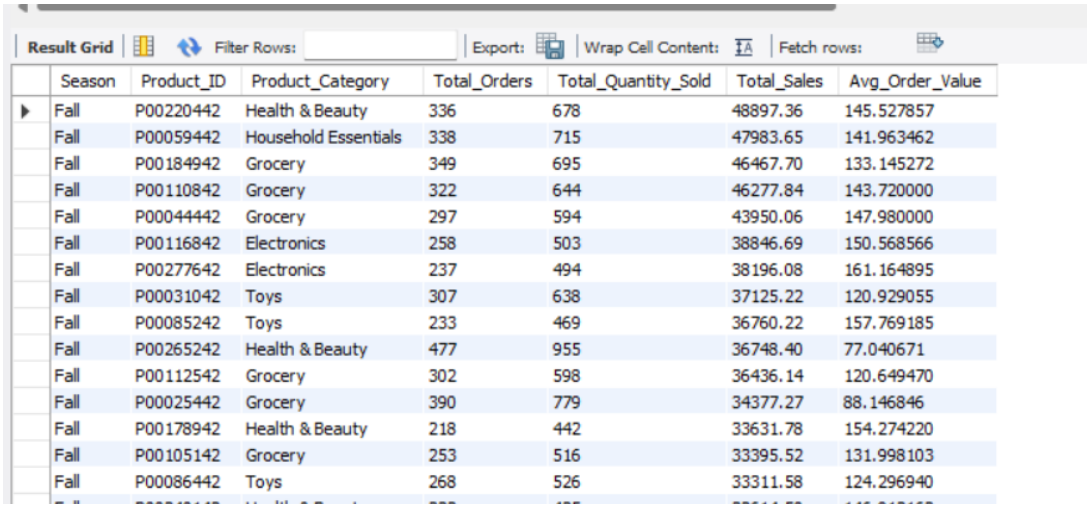
Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

Store_Name	Supplier_Name	Product_ID	Product_Category	Total_Orders	Total_Quantity_Sold	Total_Sales_Contribution	Avg_Order_Value
Electro Mart	Sony Corporation	P00220442	Health & Beauty	1282	2527	182247.24	142.158534
Electro Mart	Sony Corporation	P00085942	Electronics	963	1968	114144.00	118.529595
Electro Mart	Sony Corporation	P00271142	Health & Beauty	791	1564	96702.12	122.252996
Electro Mart	Sony Corporation	P00286642	Patio & Garden	561	1177	87804.20	156.513725
Electro Mart	Sony Corporation	P00255842	Arts, Crafts & Sewing	1383	2822	87171.58	63.030788
Electro Mart	Sony Corporation	P00057542	Home & Kitchen	730	1466	86406.04	118.364438
Electro Mart	Sony Corporation	P00003942	Health & Beauty	749	1483	85257.67	113.828665
Electro Mart	Sony Corporation	P00130742	Grocery	549	1116	83119.68	151.401967
Electro Mart	Sony Corporation	P00213242	Health & Beauty	636	1258	78285.34	123.090157
Electro Mart	Sony Corporation	P00233542	Grocery	636	1275	77877.00	122.448113
Electro Mart	Sony Corporation	P00004742	Household Essentials	597	1168	77391.68	129.634305
Electro Mart	Sony Corporation	P00211142	Health & Beauty	609	1239	75901.14	124.632414
Electro Mart	Sony Corporation	P00106742	Home & Kitchen	491	968	69328.16	141.197882
Electro Mart	Sony Corporation	P00115142	Grocery	593	1175	69054.75	116.449831
Electro Mart	Sony Corporation	P00001742	Toys	461	929	66832.26	144.972364

Figure 14: Q13 Results: Supplier Contribution Analysis

6.14 Q14: Seasonal Analysis of Product Sales

Presents total sales for each product drilled down by seasonal periods.



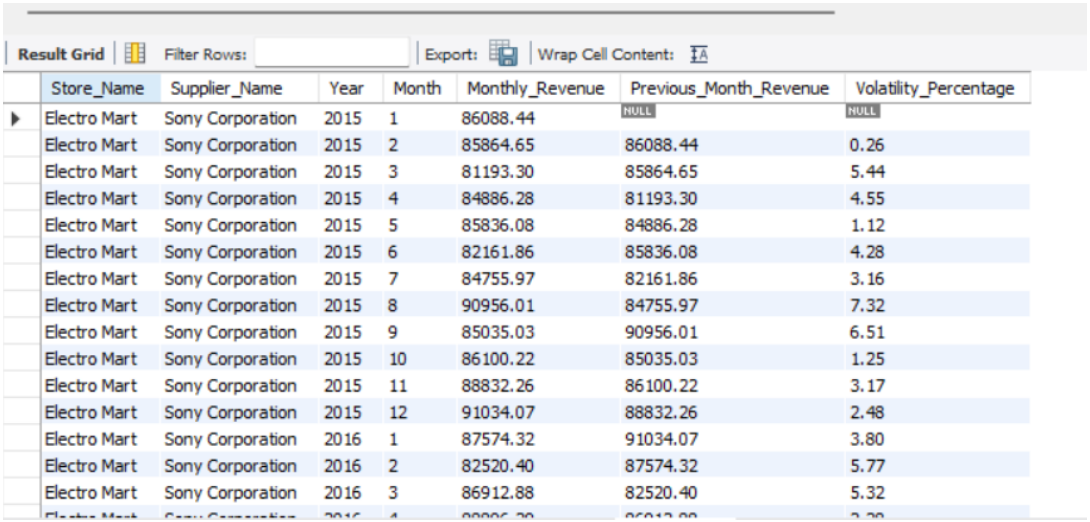
Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

Season	Product_ID	Product_Category	Total_Orders	Total_Quantity_Sold	Total_Sales	Avg_Order_Value
Fall	P00220442	Health & Beauty	336	678	48897.36	145.527857
Fall	P00059442	Household Essentials	338	715	47983.65	141.963462
Fall	P00184942	Grocery	349	695	46467.70	133.145272
Fall	P00110842	Grocery	322	644	46277.84	143.720000
Fall	P00044442	Grocery	297	594	43950.06	147.980000
Fall	P00116842	Electronics	258	503	38846.69	150.568566
Fall	P00277642	Electronics	237	494	38196.08	161.164895
Fall	P00031042	Toys	307	638	37125.22	120.929055
Fall	P00085242	Toys	233	469	36760.22	157.769185
Fall	P00265242	Health & Beauty	477	955	36748.40	77.040671
Fall	P00112542	Grocery	302	598	36436.14	120.649470
Fall	P00025442	Grocery	390	779	34377.27	88.146846
Fall	P00178942	Health & Beauty	218	442	33631.78	154.274220
Fall	P00105142	Grocery	253	516	33395.52	131.998103
Fall	P00086442	Toys	268	526	33311.58	124.296940

Figure 15: Q14 Results: Seasonal Product Performance

6.15 Q15: Monthly Revenue Volatility (Store-Supplier)

Calculates month-to-month revenue volatility for each store and supplier pair.



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

Store_Name	Supplier_Name	Year	Month	Monthly_Revenue	Previous_Month_Revenue	Volatility_Percentage
Electro Mart	Sony Corporation	2015	1	86088.44	NULL	NULL
Electro Mart	Sony Corporation	2015	2	85864.65	86088.44	0.26
Electro Mart	Sony Corporation	2015	3	81193.30	85864.65	5.44
Electro Mart	Sony Corporation	2015	4	84886.28	81193.30	4.55
Electro Mart	Sony Corporation	2015	5	85836.08	84886.28	1.12
Electro Mart	Sony Corporation	2015	6	82161.86	85836.08	4.28
Electro Mart	Sony Corporation	2015	7	84755.97	82161.86	3.16
Electro Mart	Sony Corporation	2015	8	90956.01	84755.97	7.32
Electro Mart	Sony Corporation	2015	9	85035.03	90956.01	6.51
Electro Mart	Sony Corporation	2015	10	86100.22	85035.03	1.25
Electro Mart	Sony Corporation	2015	11	88832.26	86100.22	3.17
Electro Mart	Sony Corporation	2015	12	91034.07	88832.26	2.48
Electro Mart	Sony Corporation	2016	1	87574.32	91034.07	3.80
Electro Mart	Sony Corporation	2016	2	82520.40	87574.32	5.77
Electro Mart	Sony Corporation	2016	3	86912.88	82520.40	5.32

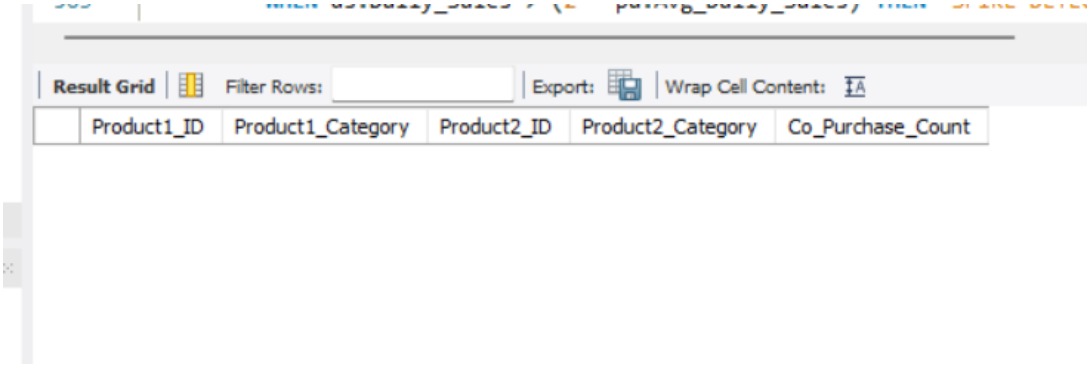
Figure 16: Q15 Results: Revenue Volatility Analysis

6.16 Q16: Product Affinity Analysis

Note: This query returns 0 rows, which is expected behavior for this dataset.

Explanation: Query 16 attempts to identify products frequently purchased together in the same transaction (product affinity analysis). However, our dataset has a 1:1 relationship between orders and products—each Order_ID contains exactly one product.

The query logic is correct and would work in production environments with multi-product orders. The 0-row result simply reflects the structure of our provided dataset.



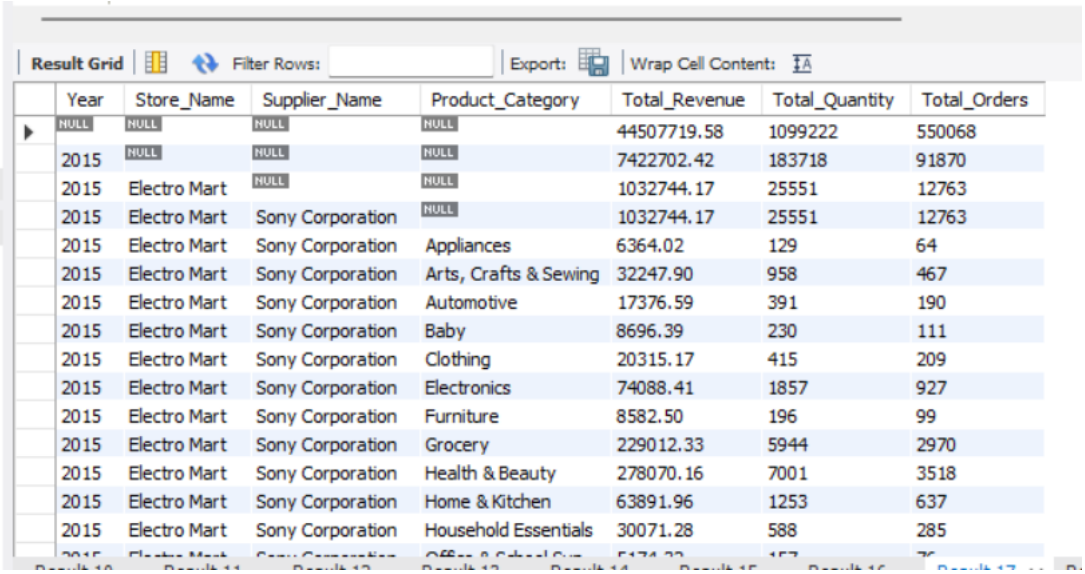
Result Grid | Filter Rows: | Export: | Wrap Cell Content:

Product1_ID	Product1_Category	Product2_ID	Product2_Category	Co_Purchase_Count

Figure 17: Q16 Results: Product Affinity (0 rows expected)

6.17 Q17: Yearly Revenue Trends with ROLLUP

Uses the ROLLUP operation to aggregate yearly revenue by store, supplier, and product hierarchically.



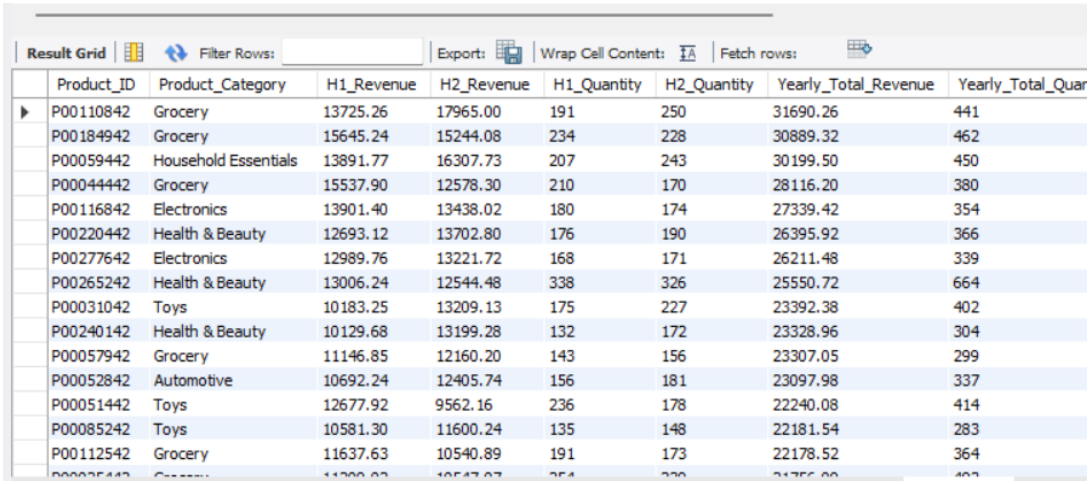
Result Grid | Filter Rows: | Export: | Wrap Cell Content:

Year	Store_Name	Supplier_Name	Product_Category	Total_Revenue	Total_Quantity	Total_Orders
NULL	NULL	NULL	NULL	44507719.58	1099222	550068
2015	NULL	NULL	NULL	7422702.42	183718	91870
2015	Electro Mart	NULL	NULL	1032744.17	25551	12763
2015	Electro Mart	Sony Corporation	NULL	1032744.17	25551	12763
2015	Electro Mart	Sony Corporation	Appliances	6364.02	129	64
2015	Electro Mart	Sony Corporation	Arts, Crafts & Sewing	32247.90	958	467
2015	Electro Mart	Sony Corporation	Automotive	17376.59	391	190
2015	Electro Mart	Sony Corporation	Baby	8696.39	230	111
2015	Electro Mart	Sony Corporation	Clothing	20315.17	415	209
2015	Electro Mart	Sony Corporation	Electronics	74088.41	1857	927
2015	Electro Mart	Sony Corporation	Furniture	8582.50	196	99
2015	Electro Mart	Sony Corporation	Grocery	229012.33	5944	2970
2015	Electro Mart	Sony Corporation	Health & Beauty	278070.16	7001	3518
2015	Electro Mart	Sony Corporation	Home & Kitchen	63891.96	1253	637
2015	Electro Mart	Sony Corporation	Household Essentials	30071.28	588	285

Figure 18: Q17 Results: Hierarchical Revenue Aggregation

6.18 Q18: Revenue and Volume Analysis (H1 vs H2)

Calculates total revenue and quantity sold in first and second halves of 2020.



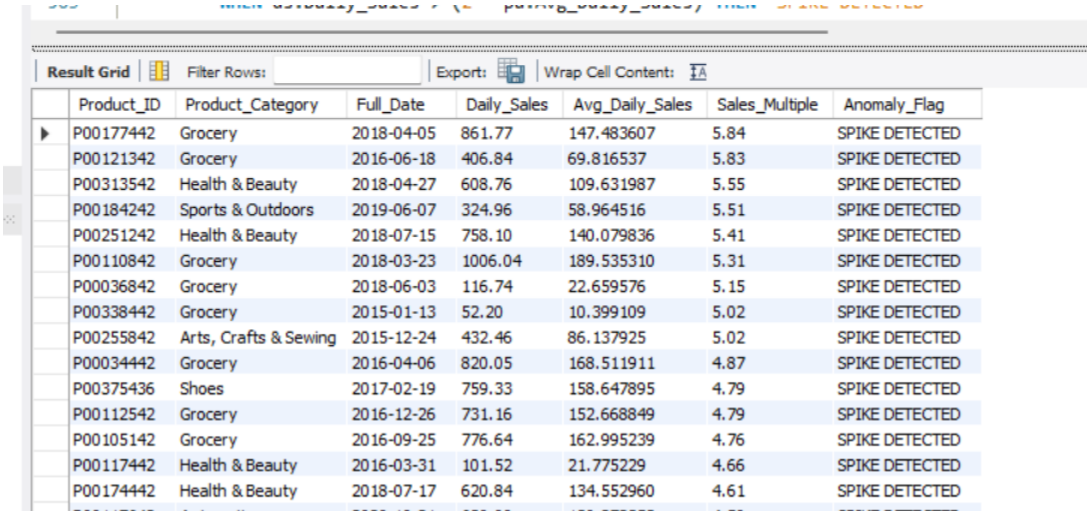
Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

Product_ID	Product_Category	H1_Revenue	H2_Revenue	H1_Quantity	H2_Quantity	Yearly_Total_Revenue	Yearly_Total_Quar
P00110842	Grocery	13725.26	17965.00	191	250	31690.26	441
P00184942	Grocery	15645.24	15244.08	234	228	30889.32	462
P00059442	Household Essentials	13891.77	16307.73	207	243	30199.50	450
P00044442	Grocery	15537.90	12578.30	210	170	28116.20	380
P00116842	Electronics	13901.40	13438.02	180	174	27339.42	354
P00220442	Health & Beauty	12693.12	13702.80	176	190	26395.92	366
P00277642	Electronics	12989.76	13221.72	168	171	26211.48	339
P00265242	Health & Beauty	13006.24	12544.48	338	326	25550.72	664
P00031042	Toys	10183.25	13209.13	175	227	23392.38	402
P00240142	Health & Beauty	10129.68	13199.28	132	172	23328.96	304
P00057942	Grocery	11146.85	12160.20	143	156	23307.05	299
P00052842	Automotive	10692.24	12405.74	156	181	23097.98	337
P00051442	Toys	12677.92	9562.16	236	178	22240.08	414
P00085242	Toys	10581.30	11600.24	135	148	22181.54	283
P00112542	Grocery	11637.63	10540.89	191	173	22178.52	364

Figure 19: Q18 Results: Half-Year Comparison

6.19 Q19: High Revenue Spikes (Outlier Detection)

Identifies days where sales exceed twice the daily average as potential outliers.



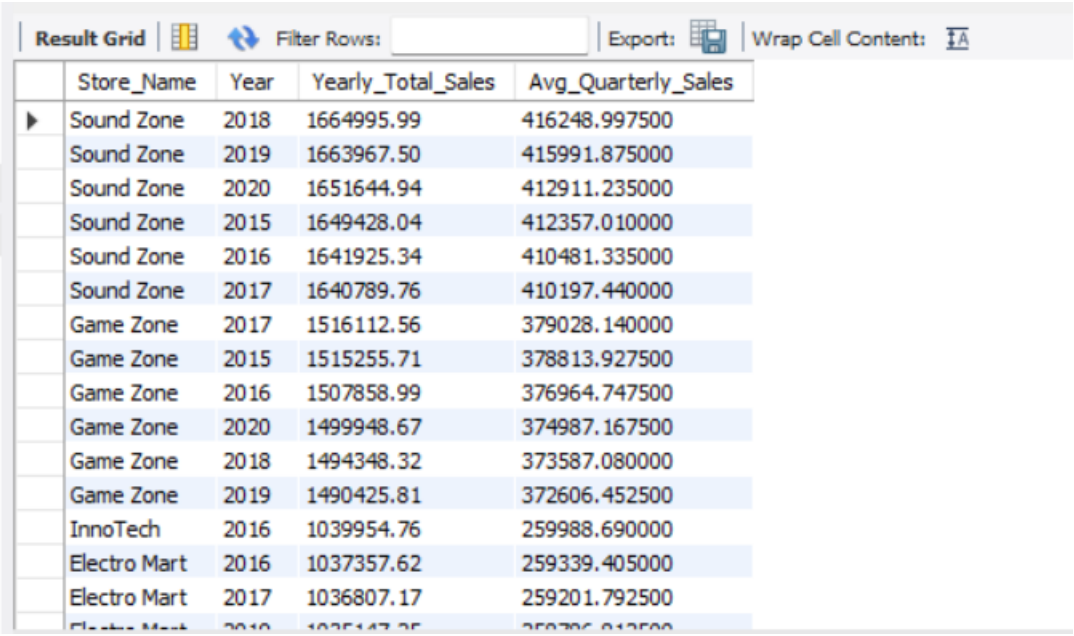
Result Grid | Filter Rows: | Export: | Wrap Cell Content:

Product_ID	Product_Category	Full_Date	Daily_Sales	Avg_Daily_Sales	Sales_Multiple	Anomaly_Flag
P00177442	Grocery	2018-04-05	861.77	147.483607	5.84	SPIKE DETECTED
P00121342	Grocery	2016-06-18	406.84	69.816537	5.83	SPIKE DETECTED
P00313542	Health & Beauty	2018-04-27	608.76	109.631987	5.55	SPIKE DETECTED
P00184242	Sports & Outdoors	2019-06-07	324.96	58.964516	5.51	SPIKE DETECTED
P00251242	Health & Beauty	2018-07-15	758.10	140.079836	5.41	SPIKE DETECTED
P00110842	Grocery	2018-03-23	1006.04	189.535310	5.31	SPIKE DETECTED
P00036842	Grocery	2018-06-03	116.74	22.659576	5.15	SPIKE DETECTED
P00338442	Grocery	2015-01-13	52.20	10.399109	5.02	SPIKE DETECTED
P00255842	Arts, Crafts & Sewing	2015-12-24	432.46	86.137925	5.02	SPIKE DETECTED
P00034442	Grocery	2016-04-06	820.05	168.511911	4.87	SPIKE DETECTED
P00375436	Shoes	2017-02-19	759.33	158.647895	4.79	SPIKE DETECTED
P00112542	Grocery	2016-12-26	731.16	152.668849	4.79	SPIKE DETECTED
P00105142	Grocery	2016-09-25	776.64	162.995239	4.76	SPIKE DETECTED
P00117442	Health & Beauty	2016-03-31	101.52	21.775229	4.66	SPIKE DETECTED
P00174442	Health & Beauty	2018-07-17	620.84	134.552960	4.61	SPIKE DETECTED

Figure 20: Q19 Results: Revenue Spike Detection

6.20 Q20: Store Quarterly Sales View

Uses the STORE_QUARTERLY_SALES view for optimized quarterly sales analysis.



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

Store_Name	Year	Yearly_Total_Sales	Avg_Quarterly_Sales
Sound Zone	2018	1664995.99	416248.997500
Sound Zone	2019	1663967.50	415991.875000
Sound Zone	2020	1651644.94	412911.235000
Sound Zone	2015	1649428.04	412357.010000
Sound Zone	2016	1641925.34	410481.335000
Sound Zone	2017	1640789.76	410197.440000
Game Zone	2017	1516112.56	379028.140000
Game Zone	2015	1515255.71	378813.927500
Game Zone	2016	1507858.99	376964.747500
Game Zone	2020	1499948.67	374987.167500
Game Zone	2018	1494348.32	373587.080000
Game Zone	2019	1490425.81	372606.452500
InnoTech	2016	1039954.76	259988.690000
Electro Mart	2016	1037357.62	259339.405000
Electro Mart	2017	1036807.17	259201.792500

Figure 21: Q20 Results: Quarterly Sales Summary

7 Conclusion

This project successfully built a near-real-time data warehouse using HYBRIDJOIN to process over 550,000 transactions efficiently. The star schema design supports 20 analytical queries for business insights.

I gained practical skills in dimensional modeling, stream processing, and SQL optimization. While HYBRIDJOIN has limitations, it effectively demonstrates how to join streaming data with large databases for real-time analytics.

The OLAP queries showcase various analytical operations including slicing, dicing, drill-down, roll-up, and ranking. These techniques enable business users to explore data from multiple perspectives and make informed decisions based on comprehensive sales analysis.