STUDENTS NAMES/NUMBERS:

IHENACHO DANIEL CHIEBUKA – 20067317

MRUNAL SUNIL FULZELE - 20029230

BHARATH SHAKTHIVEL BALAIH RAVEENDRAN – 20057027

PROGRAMME: **MSC IN DATA ANALYTICS**

LECTURER NAME: **DR. SHAZIA A AFZAL**

MODULE/SUBJECT TITLE: **B9DA111/DATA STORAGE SOLUTIONS FOR DATA ANALYTICS**

By submitting this assignment, we are confirming that:

- **This assignment is all our own work;**
- **Any sources used have been referenced;**
- **We have followed the Generative AI instructions/scale set out in the assignment brief;**
- **We have read the college rules regarding academic integrity in the QAH Part B Section 3, and the Generative AI Guidelines, and understand that penalties will be applied accordingly if work is found not to be my/our own.**
- **We understand that all work uploaded is submitted via Ouriginal, whereby a text-matching report will show**

Any similarities with other texts.

CONTENTS

1. INTRODUCTION

A data warehouse is centralised and specialised repository that consolidates and aggregates data from multiple internal, external or a combination of both sources, including **OLTP** (Online Transaction Processing) systems within an organisation, which is used for **OLAP** (Online Analytical Processing) operations and further integration for **BI** (Business Intelligence).

For our proof of concept, we have selected the Global Electronics Retailer data for the creation of the project, which is available on Maven Analytics. The dataset is a comprehensive collection of transactional data for a fictitious global electronics retailer from years 2016 – 2021. It contains information such as transactions, products and customers.

1.2. REASONS FOR SELECTING THE SUBJECT AREA AND DATA

This dataset satisfies one of the criteria for the project by having dates as a data warehouse but be dated. It also gives realistic information on sales and stores information with products, categories and subcategories by a customer within a state, country and continent.

Using this dataset will permit the practical implementation of dimensional modelling such as the creation of a star schema which is the desire format for a data warehouse or data mart. With this, a suitable star schema can be designed, data migrations can be executed, generate SSIS and SSRS report and implementation of graph database, and SQL & CQL corresponding queries are created.

1.3. VISION AND GOALS

The vision of this project is to create a suitable data warehouse that is well modelled for reporting and analysis of data both efficiently and effectively for the Sales Manager to make key decisions within the business.

The goals are as follows:

- Develop a data warehouse

- Migrate data into the data warehouse using SSIS
- Generate insightful SSRS reports and visualisations using Tableau
- Implement a Graph database
- Compare the differences between SQL and CQL queries.

1.4. KEY STAKEHOLDERS

The Sales Manager is the key stakeholder the data warehouse as it is intended for managers to make key business decisions in the **Sales Department.** He wants timely and credible information about sales over the years of the business transactions from 2016 – 2021 to understand the trends and highlights areas which needs to be addressed or further improved to maximise sales for the next few years by performing descriptive, prescriptive and diagnostic analytics.

1.5. BUSINESS REQUIREMENTS

Business requirements are as follows;

- Understand which gender has the average dollars spent and total quantities ordered
- Which continent had the most quantities of products sold
- Which year has the lowest and highest quantities sold
- Which Brand had the highest quantities sold and over the years
- Which country had the highest average profits made from over the years
- Generate all transactions details list for manager's inspection.
- Create an overall sales dashboard
- To understand Sells Distribution by Brand across continents
- To identify Top Products by Sales
- Know the highest percentage of Customers & Total Sales across Continent
- Identify the highest Customer Type by Country
- Which Month has the highest Sales by Year
- Analyse Sales by Country, Sales and City

2. SCHEMA

The schema employed will be a star schema with the following reasons;

- It is a regular practice which has descriptive attributes that remain constant or change infrequently.
- It offers a simplicity in development and maintenance, while offering an easy and straightforward layout for business users' interpretation.
- Query execution is faster than the use of a snowflake schema because of fewer joins and denormalized dimension tables.
- It makes the ETL operations more beneficial as there are much less confusions.
- Integrity is maintained as each dimension table has a primary key which is made a foreign key in the fact; thereby ensuring a consistent data warehouse.

SCHEMA DEVELOPMENT

- The selected business objective is to evaluate the sales performance for the business. This included sales by country, continent, gender, store and state.
- Granularity of the sales is based on the date dimension formulated from the sales order date. This aided in the development of a time-wise analysis and trend analysis of sales and quantities sold by year, quarters and months

Correct identification of dimensions is needed for the creation of the data warehouse.

Figure 1: Data Warehouse Schema

The following can be explained from the above schema figure;

- Stores Dimension holds information about the store
- The Customer dimension contains all information about the customer
- The Date dimension holds all information about the order dates
- The Product dimension holds all information about the product and its respective categories and subcategories

- The Sales fact table is the aspect of interest as quantities sold, and dollars made will be analysed. It contains all primary keys from dimension tables as foreign keys for easy development of queries.

Decision makers can make informed decisions and gain insights into the sales performance through the balance between simplicity & analytical power offered by the **STAR SCHEMA.**

3. ETL

Using Microsoft SQL Server Management Studio and Microsoft Visual Studio, the data warehouse was created as shown in the figures below;

Figure 2: Customer Dimension Migration

The data to be migrated contains the details of customers in OLE DB source database (except that the OLE DB data source is a (LocalDb) MSSQLLocalDB database called AdvanceNetworking). SQL Server Integration Services will copy the customer data to the destination table. The extract source data is retrieved in the [Customers] table.

In the SSIS package, the transformation between the destination and the source is Data Conversion transformation. This means that one or more columns in the source table Gender, Name, City, State_Code, State, Zip_Code, Country, Continent must be changed in data types or what is considered suitable data types in the target system.

Most of the source columns will be directly mapped to the same destination columns Gender to Gender, Name to Name etc. It is worth noting that CustomerKey is extracted, however does not seem to be selected conversion, thus possibly being processed differently. One of the columns is named Birthday, and it only appears in the destination mapping, which means that its data may be calculated during the conversion process or will be provided at a later stage. Error handling configuration in data flow is one of the steps involved. Basically, the demographics and location data of customers are pulled out, possibly transformed and loaded.

Figure 3: Date Dimension Migration

The purpose of this SSIS data flow is to process sales order date to fill a table with date dimension **Calendar_Dim** in a data warehouse. The following are the workflow:

Source Extraction: The data comes in the **Order_Date** table of OLE DB **sale_data**.

Sorting: The Order_Date in the Sort transformation occurs chronologically in ascending order.

Creation of Derived Columns: Derived Column transformation is a transformation that enhances data in the date column by creating the data which generates:

FullDate: Day of week

DayOfWeek: Whether the weekend/weekday (Weekend/weekday)

DayType: number of the month

Month: Number of months

Quater: Year

Conversion of Types of Data: The critical columns are also converted to ANSI-encoded string (DT_STR) of predetermined lengthsso that they are compatible.

Destination Loading- .**Calendar_Dim** loads data with fast-load optimization in **AdvanceNetworkingWarehouse** database.

Purpose: Inserts a reusable dimension of dates of time-based analysis and normalizing date fields **(weekdays, months, years)** out of raw sales orders. The pipeline makes the warehouse plug-in chronologically sorted and type safe.

Figure 4: Product Dimension Migration

This SSIS package is used to do a direct migration of product data between an operational database and a dimensional warehouse table:

Source Extraction:

Reads **Products** table in database **AdvanceNetworking** at source OLE DB Source.

Columns to be extracted: **ProductKey, Product_Name, Brand, Colour, Unit_Cost_USD, Unit_Price_USD, SubcategoryKey and CategoryKey**.

Destination Loading:

Inserts into **Products_Dim** in **AdvanceNetworkingWarehouse** (OLE DB Destination).

Optimizes loads to fast with locking on the tables.

The check of constraint disabled to maximize speed during load.

Column Mappings:

Most attributes are direct 1:1 (like ProductKey\ ProductKey, Product_Name\ Product_Name, and some others).

Key transformations:

SubcategoryKey (source) and (destination): (Subcategory)

Missing Mapping: Source Subcategory description column is un-mapped which makes it an indication that this dimension will only require keys.

Purpose: The dimension table of products is generated to use in analytical queries, staying with the main attributes and hierarchies of categories. The data transformations are also absent, which means it is a simple structural migration. Performance optimization is made on bulk insert using fast-load settings within the warehouse setting.

Figure 5: Store Dimension Migration

This SSIS package will help in migrating the store related information in an operational database to a dimensional warehouse table. The procedure starts by extracting data in table Stores in AdvanceNetworking database. Some of the important columns that are chosen are StoreKey, Country, State, Square_Meters, and Open_Date. Arguably, the Name column is retrieved, yet it is not processed further which implies that it is not necessary in the target environment.

After extraction, an extremely important deduplication process is carried out. The Sort transformation sorts the records based on StoreKey in ascending order and the option of the removal of the duplicate sort values is turned on. This guarantees that the only unique StoreKey records make it out to the downstream environments, and removes redundant records, and ensures data integrity of the store dimension.

The fixed data is later uploaded into the Stores_Dim table in the AdvanceNetworkingWarehouse database. Its loading strategy uses the fast-load optimization, however, with a high configuration penalty: Table locks, constraint checks, and null retention are all turned off. Although this increases the speed of insertion, it presupposes validity of source data and can be risky in terms of integrity once inconsistencies are present. Columns relating to each column depend on the mappings to columns. Structural consistency is maintained between head and neck.

Generally, the aim is to develop a deduplicated location-consistent store dimension to analytical reporting. This pipeline will sustain a trade-off between dimensional reliability and efficiency in the warehouse environment because it focuses on uniqueness using StoreKey deduplication and tolerates loading trade-offs.

Figure 6: Sale Fact Migrations

The package used in this SSIS will perform the loading of a sales fact table with a combination of transactions with sales data based on dimension key elements that can be achieved with a multi-step approach. At the inlet stage, a custom SQL source query is written in the pipeline to structure raw sales records in the form of a Common Table Expression. This query can compute the line-item total, and it makes use of ROW_NUMBER to first determine the first ProductKey to be used downstream. After the extraction, there are four look up transformations that resolve the operational keys to the data warehouse surrogate keys: Customers_Lookup is the source customer identifier, **Products_Dim_Lookup** is the primary product per **order**, **Store_dim_Lookup** is store identifier, and the **Date_dim_Lookup** is the derivation of a date key based on order date.

There are two data conversion stages involving type compatibility between source data and target schema and standardization of formats of keys and measures. Lastly, the optimized data is loaded into **Sales_Fact** table performing fast-load optimizers where table locking is turned on. The destination mappings consist of degenerate dimensions, solved surrogate keys, and measures. This integration allows referential integrity between the dimensions of warehouse with efficiently structuring the transactional data to allow the analysis.

4. VISUALIZATIONS AND REPORTS

4.1. VISUALIZATIONS

The business requirements for the visualisations are as follows;

- Create an overall sales dashboard
- To understand Sales Distribution by Brand across continents
- To identify Top Products by Sales
- Know the highest percentage of Customers & Total Sales across Continent
- Identify the highest Customer Type by Country
- Which Month has the highest Sales by Year

- Analyse Sales by Country, Sales and City

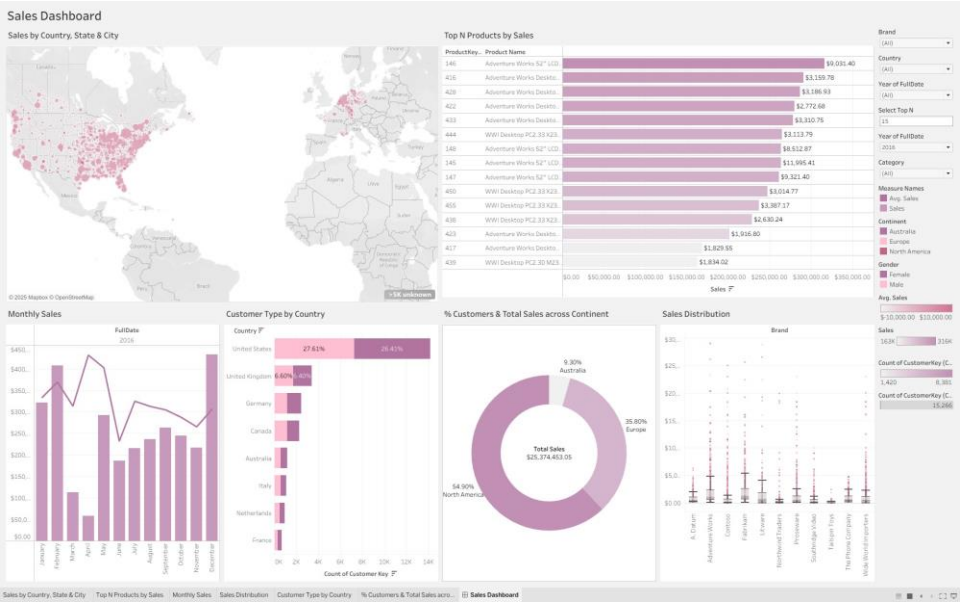


Figure 7: Sales Dashboard

- The primary purpose of creating this dashboard is to analyse the Sales using data from Adventure Works Data Warehouse.
- Filters such as Year, Brand, Country, Product Category, Top N were implemented to analyse the sales accordingly.
- Global action filter is created to make the visualizations more interactable.
- Legends included in the dashboard are Continent, Gender, Average of Sales and Count of Customers.
- The Dashboard uses Purple-Pink theme to be consistent with all the visuals and to make it visually appealing to the target audience.
- The Dashboard contains 6 visuals which will be briefed one by one.

Figure 8: Sales Distribution by Brand

- Adventure Works, Contoso and Fabrikam are some of the top brands which has widest range and greatest concentration of high sales. This indicates the strong overall performance.

- Litware and Northwind Traders have steady sales distributions, and their performance is stable. These brands have very few extreme high sales.
- Southridge Video, Tailspin Toys and The Phone Company are the low performing brands. They don't have extreme sale values as well.
- Sales are spread across all continents such as North America, Europe and Australia and we can see that there is no dominating market for all brands.
- Brands which some outlier points which are above $20,000, shows high value sales occasionally. This causes skewness to sales analysis based on average.

Figure 9: Top N Products by Sales

- The Adventure Works 52" LCD HDTV X590 Black is the highest selling product by a significant margin. Total Sales of that product is approximately $9031.
- Most of the adventure works products are the top performers which consists of products like Desktop PCs and LCD TVs.
- The Adventure Works products tend to dominate the market. This shows the brand loyalty and popularity in the market compared to their competitors like World Wide Importers.
- This visualization has Top N products filter, meaning that we can customise based on the user needs. Let's say if the user wants only to show the top 5 products, it can be achieved by entering the value in the filter.

Figure 10: % of Customers

- North America is the leading and largest revenue generating continent, contributing around 55% of Total Sales.
- Europe is the second largest revenue generating continent, accounting for 36% of Total Sales.
- Australia has the lowest remaining share of 9% of Total Sales.

- The Sale value is around $25.3 Million, with North America and Europe being the top contributors.
- This sales distribution highlights the high dependency on North America and Europe and the potential growth opportunities in low performing regions like Australia.

Figure 11: Customer Type by Country

- United States has the highest customer base with (appx) 54% (27.61% female, 26.41% male), when compared to other countries.
- The 2nd and 3rd country with the highest customer base is United Kingdom (13%) and Germany (9.27%).
- Canada (8.67%) and Australia (4.49%) has moderate customer base which shows that those areas must be focused to boost sales.
- Gender distribution is balanced across all countries. There are only minor differences between male and female percentages.
- Netherlands and France have the least shares which indicates potential markets for targeted expansion to boost the sales.

Figure 12: Monthly Sales

- There was peak Sales during early 2016 and late 2019. It was exceeding $1.1M which means there is strong seasonal trend.
- Sales remained consistent between 2017–2018 but lower compared to peak periods.
- There was a significant growth in late 2019 with multiple months exceeding $1M which shows greater market expansion.
- There is a significant decline in early 2020. The Sales was dropped after the 2019 peak. This might be due to external factors (say global events like COVID 19).
- Average sales remained stable overall, despite fluctuations in monthly sales trends.

Figure 13: Sales by Country, State, City

- Most of sales are concentrated in the eastern and central United States, especially around major cities like New York, Atlanta, and Chicago.
- There is also significant sales activity in Western Europe, particularly in Germany, France, the Netherlands, and the UK.
- There is very less or no sales activity across most of Asia, Africa, and South America which indicates marketing must be focused on these markets.
- The Sales are most clustered in and around urban areas. This shows that there must be focus on metropolitan regions for product or service distribution.

4.2. REPORTS

The business requirements for the reports are as follows;

- Understand which gender has the average dollars spent and total quantities ordered
- Which continent had the most quantities of products sold
- Which year has the lowest and highest quantities sold
- Which Brand had the highest quantities sold and over the years
- Which country had the highest average profits made from over the years
- Generate all transactions details list for manager's inspection.

Figure 14: Gender Summary Information (Design-view)

Gender Summary Information Over Years

Gender	Total Quantity Ordered	Average Amount
Female	41042	$973.22
Male	41827	$954.82

Figure 15: Gender Summary Information (Preview-view)

The above report shows that males have a higher amount of quantities ordered but females have a higher average amount of dollars spent.

Figure 16: Total Quantities of Products Sold in Each Year by Continent and Country (Design-view)

Figure 17: Total Quantities of Products Sold in Each Year by Continent and Country (Preview-view)

North America had the most products sold over the years and 2019 had the highest product quantities sold, while 2021 had the lowest number of products sold.

Figure 18: Total Quantity Sold by Brands and Products over the Years, Quarters and Months (Design-view)

Figure 19: Total Quantity Sold by Brands and Products over the Years, Quarters and Months (Preview-view)

Wide World Importers had the highest product quantities sold over the years while **A. Datum** had the lowest number of product quantities sold.

Figure 20: Average Stores Performance across Each Year, Quarter and Month (Design-view)

Figure 21: Average Stores Performance across Each Year, Quarter and Month (Preview-view)

- In 2016, Australia had the lowest dollar spent, while Germany had the highest.
- In 2017, Canada had the lowest dollars spent, while Australia had the highest dollars spent.
- In 2018, Australia had the lowest dollars spent, while Netherlands had the highest
- In 2019, The United Kingdom had the lowest dollars spent, while France had the highest dollars spent.
- In 2020, Italy had the lowest dollars spent, while Australia had the highest dollars spent
- In 2021, Italy had the lowest dollars spent, while Germany had the highest dollars spent.

Figure 22: Amount Details Information (Design-view)

Figure 23: Amount Details Information (Preview-view)

The Amount Details provides a spreadsheet for the Sales Manager to inspect each order number thereby marking out what order had the highest total amount sold. It has about 599 pages. Based on the screenshot given. Order number 1520023 has the highest dollars spent on 2019-02-28

5. GRAPH DATABASES

Several queries are created in Neo4j and can find the output here. It is also compared with SQL queries to see if the results are consistent accross the two platforms. See Appendix A for codes which includes Queries, Nodes and Relationships

	Order_Number	CustomerKey	Total_Amount	Country
1	"366001"	1269051	854.0	"Online"
2	"366002"	266019	326.0	"Online"
3	"366010"	370077	1495.0	"Online"
4	"366014"	738549	279.99	"Online"
5	"367005"	758280	1738.0	"Online"
6	"367028"	607356	188.0	"Online"

Results Messages

	Order_Number	CustomerKey	Total_Amount	Country
1	366000	265598	68.00	Online
2	366001	1269051	854.00	Online
3	366002	266019	326.00	Online
4	366004	1107461	9553.20	Online
5	366005	844003	1876.00	Online
6	366007	2035771	815.00	Online
7	366008	759705	3135.00	Online
8	366009	254540	109.99	Online
9	366010	370077	1495.00	Online
10	366011	1984985	1003.80	Online
11	366012	1977527	25.00	Online
12	366013	1187306	1299.95	Online
13	366014	738549	279.99	Online
14	366016	1982762	139.93	Online
15	367000	1438050	1000.00	Online
16	367003	980164	939.94	Online
17	367004	1853690	249.95	Online

Query 1: Get orders with store country

```
neo4j$ MATCH (o:Order) RETURN ROUND(MAX(o.amount), 2) AS MaxAmount, ROUND(MIN(o.amount), 2) AS MinAmount, ROUND(AVG(o.amount), 2) AS A…
```

MaxAmount	MinAmount	AvgAmount
28999.9	0.95	963.86

Started streaming 1 records after 9 ms and completed after 42 ms.

Results | Messages

	maximum_total_Amount	minimum_total_Amount	average_total_Amount
1	28999.90	0.95	963.890000

Query 2: Aggregated order amounts

```
neo4j$ MATCH (p:Product) WHERE p.category IN ['5', '2'] RETURN p.product_id, p.name, p.category;
```

	p.product_id	p.name	p.category
1	116	"Adventure Works 20" CRT TV E15 Silver"	"2"
2	117	"Adventure Works 20" CRT TV E15 Black"	"2"
3	118	"Adventure Works 20" CRT TV E15 White"	"2"
4	119	"Adventure Works 13" Color TV E25 Silver"	"2"
5	120	"Adventure Works 13" Color TV E25 Black"	"2"
6	121	"Adventure Works 13" Color TV E25 White"	"2"

Started streaming 587 records after 8 ms and completed after 16 ms.

	ProductKey	Product_Name	Brand	Color	Unit_Cost_USD	Unit_Price_USD	Subcategory	Category
1	116	Adventure Works 20" CRT TV E15 Silver	Adventure Works	Silver	86.67	169.99	02:01:00.000000000	2
2	117	Adventure Works 20" CRT TV E15 Black	Adventure Works	Black	86.67	169.99	02:01:00.000000000	2
3	118	Adventure Works 20" CRT TV E15 White	Adventure Works	White	86.67	169.99	02:01:00.000000000	2
4	119	Adventure Works 13" Color TV E25 Silver	Adventure Works	Silver	61.17	119.99	02:01:00.000000000	2
5	120	Adventure Works 13" Color TV E25 Black	Adventure Works	Black	61.17	119.99	02:01:00.000000000	2
6	121	Adventure Works 13" Color TV E25 White	Adventure Works	White	61.17	119.99	02:01:00.000000000	2
7	122	Adventure Works 19" Portable LCD HDTV M110 Silver	Adventure Works	Silver	128.76	279.99	02:01:00.000000000	2
8	123	Adventure Works 19" Portable LCD HDTV M110 Black	Adventure Works	Black	128.76	279.99	02:01:00.000000000	2
9	124	Adventure Works 19" Portable LCD HDTV M110 White	Adventure Works	White	128.76	279.99	02:01:00.000000000	2
10	125	Adventure Works 19" Color Digital TV E35 Silver	Adventure Works	Silver	73.11	143.4	02:01:00.000000000	2
11	126	Adventure Works 19" Color Digital TV E35 Black	Adventure Works	Black	73.11	143.4	02:01:00.000000000	2
12	127	Adventure Works 19" Color Digital TV E35 White	Adventure Works	White	73.11	143.4	02:01:00.000000000	2
13	128	Adventure Works 19" Color Digital TV E35 Brown	Adventure Works	Brown	73.11	143.4	02:01:00.000000000	2
14	129	Adventure Works 20" Analog CRT TV E45 Silver	Adventure Works	Silver	101.97	200	02:01:00.000000000	2
15	130	Adventure Works 20" Analog CRT TV E45 Black	Adventure Works	Black	101.97	200	02:01:00.000000000	2
16	131	Adventure Works 20" Analog CRT TV E45 White	Adventure Works	White	101.97	200	02:01:00.000000000	2

Query 3: Products with category 5 or 2

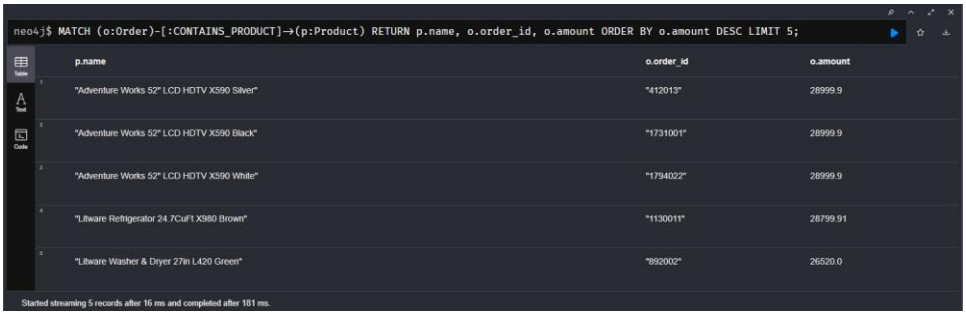
```
neo4j$ MATCH (o:Order)-[:CONTAINS_PRODUCT]→(p:Product) WHERE o.amount > 500 AND p.category IN ['5','2'] RETURN o.order_id, p.product_…
```

	o.order_id	p.product_id	p.name	p.category	o.amount
1	"498005"	116	"Adventure Works 20" CRT TV E15 Silver"	"2"	679.96
2	"596012"	116	"Adventure Works 20" CRT TV E15 Silver"	"2"	1019.94
3	"598012"	116	"Adventure Works 20" CRT TV E15 Silver"	"2"	509.97
4	"669000"	116	"Adventure Works 20" CRT TV E15 Silver"	"2"	849.95
5	"689003"	116	"Adventure Works 20" CRT TV E15 Silver"	"2"	1189.93
6	"927002"	116	"Adventure Works 20" CRT TV E15 Silver"	"2"	1019.94

Started streaming 2874 records after 16 ms and completed after 41 ms, displaying first 1000 rows.

	Order_Number	ProductKey	Product_Name	Category	Total_Amount
1	366004	163	Adventure Works 52" LCD HDTV X790W White	2	9553.20
2	366011	128	Adventure Works 19" Color Digital TV E35 Brown	2	1003.80
3	367003	153	Adventure Works 26" 720p LCD HDTV M140 Silver	2	939.94
4	367005	319	SV Car Video LCD9.2W X9281 Silver	2	1738.00
5	367012	145	Adventure Works 52" LCD HDTV X590 Silver	2	8699.97
6	367016	206	Litware Home Theater System 5.1 Channel M515 Black	2	1707.00
7	367021	1445	The Phone Company Touch Screen Phones 26-1.4" M250	5	1072.00
8	367024	302	SV Car Video LCD9.2W X9280 Black	2	7992.00
9	369012	1423	The Phone Company Touch Screen Phones - CRT M11 Bl	5	567.00
10	370007	126	Adventure Works 19" Color Digital TV E35 Black	2	1003.80
11	371012	147	Adventure Works 52" LCD HDTV X590 White	2	2899.99
12	372000	148	Adventure Works 52" LCD HDTV X590 Brown	2	8699.97
13	377006	320	SV Car Video LCD7W M7082 Silver	2	699.00
14	378005	125	Adventure Works 19" Color Digital TV E35 Silver	2	573.60
15	378009	336	SV Car Video LCD7W M7082 Brown	2	2097.00
16	378010	145	Adventure Works 52" LCD HDTV X590 Silver	2	8699.97
17	379004	321	SV Car Video LCD7 M7001 Silver	2	658.00

Query 4: High value orders for category 5 or 2 products

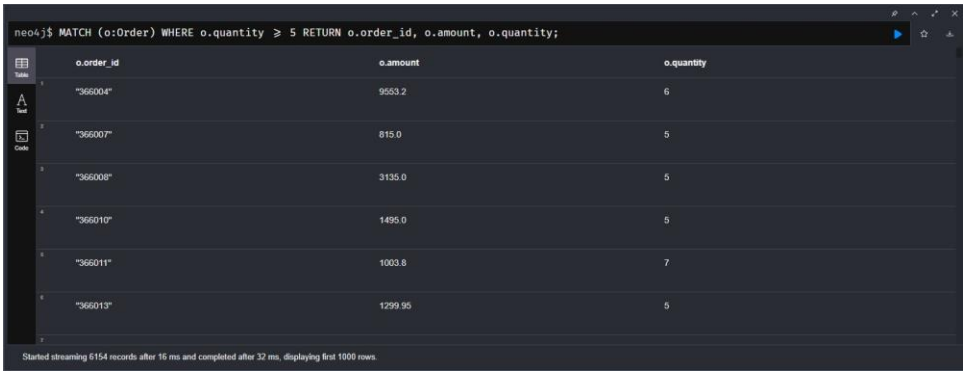


	Product_Name	Order_Number	Total_Amount
1	Adventure Works 52" LCD HDTV X590 Silver	412013	28999.90
2	Adventure Works 52" LCD HDTV X590 Black	1731001	28999.90
3	Adventure Works 52" LCD HDTV X590 White	1794022	28999.90
4	Litware Refrigerator 24.7CuFt X980 Brown	1130011	28799.91
5	Litware Washer & Dryer 27in L420 Green	892002	26520.00

Query 5: Top 5 most expensive orders (with product info)

	Name	Order_Number	Total_Amount
1	Kristin Oster	412013	28999.90
2	Angelo Nolan	1794022	28999.90
3	Wayne Banks	1731001	28999.90
4	Esin Holman	1130011	28799.91
5	Jacqueline Casias	892002	26520.00

Query 6: Top 5 customers by order amount



	Order_Number	CustomerKey	ProductKey	StoreKey	DateKey	Total_Amount	Quantity
1	366004	1107461	163	38	251537	9553.20	6
2	366007	2035771	666	43	251537	815.00	5
3	366008	759705	1060	29	251537	3135.00	5
4	366010	370077	618	0	251537	1495.00	5
5	366011	1984985	128	66	251537	1003.80	7
6	366013	1187306	1654	41	251537	1299.95	5
7	366016	1982762	1253	63	251537	139.93	7
8	367007	1730985	65	57	251564	1267.00	7
9	367009	1512558	53	57	251564	1480.00	5
10	367013	815458	41	31	251564	1392.00	6
11	367015	911025	2296	42	251564	1879.50	7
12	367019	793573	371	29	251564	4193.00	7
13	367020	799366	448	29	251564	1619.40	6
14	367024	1713253	302	50	251564	7992.00	8
15	367029	1260533	1223	66	251564	3220.00	7
16	369004	1389462	1961	65	251636	1399.95	5

Query 7: Orders with quantity >= 5

5.1. COMAPRISON TO RELATIONAL DATABASES

Table 1: Comparison of SQL TO CQL

Aspect	Relational (SQL)	Graph (CQL / Neo4j)
Data Modelling	Tables with foreign keys and joins	Nodes and relationships, more intuitive for connections
Query Complexity	Grows rapidly with joins	Natural and often simpler syntax for relationships
Performance	Slower on deep joins or multi-table traversals	Faster for highly connected data (e.g., paths, networks)
Use Case Fit	Structured, transactional data	Relationships-focused scenarios (e.g., social, sales)
Ease of Understanding	Requires understanding normalized schema	Model reflects real-world entities and relations
Scalability	Optimized for tabular scalability	Scales well with connected data, less so for flat data

6. CONCLUSIONS

The project was successful executed and produced the data warehouse solution for the Global Electronics Retailer dataset, by integrating a star schema design, ETL processes, reporting, and graph database capabilities to support the necessary insights needed for managerial decision making for stakeholders.

In a broad scope the following was explored;

- **Schema & ETL:** A dimensional modelling approach via the use of a star schema was employed to enable efficient OLAP operations & reporting. The ETL process/pipeline was implemented using SSIS and SQL Server

to ensure accurate extraction, transformation and loading of the data into the data warehouse. Key processes were data type conversions, surrogate key lookups and deduplication, which resulted to a clean and consistent data warehouse system.

- **Insights & Reporting:** With the use of Tableau & SSRS, the system delivered actionable information which aligned with the business requirements. North America was identified as the top performing region, while Adventure Works dominated in terms of product sold. Females showed higher average dollars spent despite males placing more orders. Seasonal sales patterns were identified, with 2019 marking the peak year. Regions such as Australia and some European countries are underperforming but could be avenues for growth opportunities once further analytics are carried out.
- **Graph vs Relational:** Graph database implementation in Neo4j highlighted the advantages of relational to a graph database. Compared to query execution in SQL, CQL offered better performance for complex relationship.

The project met its objectives by providing a scalable, efficient and insightful data warehouse infrastructure which empowers the business stakeholders, such as the **Sales Manager**, to make data-driven decisions that enhance operational effectiveness and identify opportunities.

7. BIBLIOGRAPHY

Bismart (2023) Data warehousing: ETL, OLAP and OLTP. Available at: https://blog.bismart.com/en/data-warehousing-olap-oltp (Accessed: 24 July 2025).

Maven Analytics (2024) Global Electronics Retailer dataset. Available at: https://mavenanalytics.io/data-playground?order=date_added%2Cdesc&search=Global%20Electronics%20Retailer (Accessed: 16 June 2025)

APPENDIX A – NEO 4J CODE

Query 1: Get orders with store country

MATCH (sto:Store) OPTIONAL MATCH (c:Customer)-[:PLACED]->(o:Order)-[:ORDERED_FROM]->(sto) RETURN o.order_id AS Order_Number, c.customer_id AS CustomerKey, o.amount AS Total_Amount, sto.country AS Country;

Query 2: Aggregated order amounts

MATCH (o:Order) RETURN ROUND(MAX(o.amount), 2) AS MaxAmount, ROUND(MIN(o.amount), 2) AS MinAmount, ROUND(AVG(o.amount), 2) AS AvgAmount;

Query 3: Products with category 5 or 2

MATCH (p:Product) WHERE p.category IN ['5', '2'] RETURN p.product_id, p.name, p.category;

Query 4: High value orders for category 5 or 2 products

MATCH (o:Order)-[:CONTAINS_PRODUCT]->(p:Product) WHERE o.amount > 500 AND p.category IN ['5','2'] RETURN o.order_id, p.product_id, p.name, p.category, o.amount;

Query 5: Top 5 most expensive orders (with product info)

MATCH (o:Order)-[:CONTAINS_PRODUCT]->(p:Product) RETURN p.name, o.order_id, o.amount ORDER BY o.amount DESC LIMIT 5;

Query 6: Top 5 customers by order amount

MATCH (c:Customer)-[:PLACED]->(o:Order) RETURN c.name, o.order_id, o.amount ORDER BY o.amount DESC LIMIT 5;

Query 7: Orders with quantity >= 5

MATCH (o:Order) WHERE o.quantity >= 5 RETURN o.order_id, o.amount, o.quantity;

Nodes & Relationships

Customer Node

LOAD CSV WITH HEADERS FROM 'file:///customers.csv' AS row CREATE (:Customer { customer_id: toInteger(row.CustomerKey), name: row.Name, gender: row.Gender, country: row.Country, state: row.State, city: row.City, continent: row.Continent });

Product Node

LOAD CSV WITH HEADERS FROM 'file:///products.csv' AS row CREATE (:Product { product_id: toInteger(row.ProductKey), name: row.Product_Name, brand: row.Brand, category: row.Category, price: toFloat(row.Unit_Price_USD) });

Store Node

LOAD CSV WITH HEADERS FROM 'file:///stores.csv' AS row CREATE (:Store { store_id: toInteger(row.StoreKey), country: row.Country, state: row.State, square_meters: toInteger(row.Square_Meters), open_date: row.Open_Date });

Date Node

LOAD CSV WITH HEADERS FROM 'file:///dates.csv' AS row CREATE (:Date { date_id: toInteger(row.date_key), full_date: row.FullDate, month: row.Month, quarter: row.Quarter_, year: row.Year });

Order Node and Relationships

1. LOAD CSV WITH HEADERS FROM 'file:///orders.csv' AS row MERGE (c:Customer {customer_id: toInteger(row.CustomerKey)}) MERGE (p:Product {product_id: toInteger(row.ProductKey)}) MERGE (s:Store {store_id: toInteger(row.StoreKey)}) MERGE (d:Date {date_id: toInteger(row.DateKey)}) CREATE (o:Order { order_id: row.Order_Number, amount: toFloat(row.Total_Amount), quantity: toInteger(row.Quantity) }) MERGE (c)-[:PLACED]->(o)

MERGE (o)-[:CONTAINS_PRODUCT]->(p) MERGE (o)-[:ORDERED_FROM]->(s) MERGE (o)-[:ON_DATE]->(d);

2. LOAD CSV WITH HEADERS FROM 'file:///products.csv' AS row MERGE (cat:category {category: row.Category}) WITH cat,row MATCH (p:product {product_id: toInteger(row.ProductKey)}) MERGE (p)-[:BELONGS_TO]->(cat)

3. LOAD CSV WITH HEADERS FROM 'file:///customers.csv' AS row MERGE (city:City {name: row.City}) WITH city,row MATCH (c:Customer {customer_id: toInteger(row.CustomerKey)}) MERGE (c)-[:LIVES_IN]->(city)

4. LOAD CSV WITH HEADERS FROM 'file:///customers.csv' AS row MERGE (country:Country {name: row.Country}) WITH country,row MATCH (s:Store {store_id: toInteger(row.StoreKey)}) MERGE (s)-[:LOCATED_IN]->(country)

APPENDIX B – SQL CODE

```
select sf.Order_Number,sf.CustomerKey,sf.Total_Amount,sto.Country

from Sales_Fact sf right join

Stores_Dim sto on sf.StoreKey = sf.StoreKey


select ROUND( Max(Total_Amount),2)
maximum_total_Amount,ROUND( MIN(Total_Amount),2)
minimum_total_Amount,ROUND( AVG(Total_Amount),2 )average_total_
Amount

from Sales_Fact
```

```
select  *  from Products_Dim where Category in (5,2)


select sf.Order_Number,pd.ProductKey ,
pd.Product_Name,pd.Category,sf.Total_Amount  from Sales_Fact sf

inner join Products_Dim pd on sf.ProductKey = pd.ProductKey where
sf.Total_Amount > 500

and pd.Category in (5,2)


select Top 5  Product_Name,Order_Number , Total_Amount from
Sales_Fact sf

right join Products_Dim pd on pd.ProductKey = sf.ProductKey order by
Total_Amount desc


select Top 5  cd.Name,Order_Number , Total_Amount from  Sales_Fact sf

right join Customers_Dim cd on cd.CustomerKey = sf.CustomerKey order
by Total_Amount desc


select *  from Sales_Fact where Quantity >=5
```