Table of Contents

Introduction:

SMP Retail is the largest retail chain, having hundreds of stores in different states in the UK. Every manager in the store will look at the staff members, inventory stock, promotions, and store operations. They only consider basic sales data from the store. That means the customer experience is not consistent in the stores. Sometimes, there will be a full stock of items; sometimes, there will be out-of-stock. As a result, the organisation doesn't understand the customer orders, sales of the product and how the store performs in this scenario.

As hundreds of stores are generating large numbers of transactions daily from the POS systems, The SMP Store can't understand the overall wealth of this data for seeing the sales patterns, the behaviour of the customers and product sales, which helps to improve the store performance.

To improve this, SMP wants to implement a data warehouse that can analyse the data from various sources like POS transactions, customer data, product data, and store data. By analysing this data, the organisation can decide the customer's behaviour, product sales, and store performance, which leads to improving sales and managing the inventory.

The data warehouse implementation will have several tasks; in the first task, we will design an ERD Diagram explaining the key concepts for the following benefits and challenges of implementing a data warehouse and analyse different techniques relevant to this scenario.

In task 2, we will implement the data warehouse using Snowflake, which has facts and dimensions related to the store. In task 3, the data cleaning and visualising will be done using BI tools after creating the data warehouse, which helps to analyse the data easily.

Finally, we identify ethical issues related to data privacy, security, and bias in the context of this data warehouse and propose strategies to mitigate the risks.

Task 1: DESIGN

Critical concepts of databases in the retail scenario:

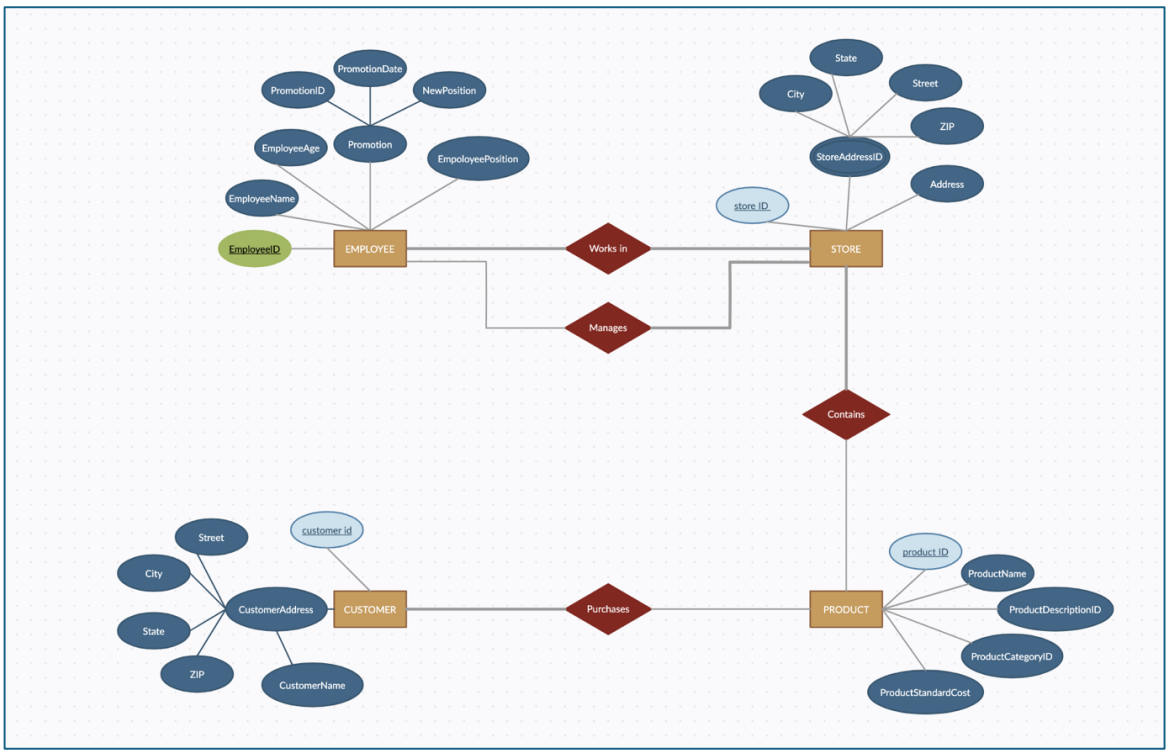

Fig 1: ERD Diagram

The above diagram is the ERD diagram, which shows the relation between entities and now a few critical concepts of this database can be explained.

For managing the massive amounts of data, databases are required. A few ideas that are important to SMP Retail are:

Relational Databases: The sources in this data warehouse, like POS transactions, customers, products, and store data, will be arranged in the tables inside an RDMS so that data can be stored effectively and retrieved.

Schema: Schemas like Star and Snowflake are essential for managing the data in SMP. The SQL server queries that describe the schema have multiple facts and dimension tables like Facts tables- Sales and Inventory, and Dimensions tables- Customer, Product, Store and Time dimensions. So, by designing this schema, it will be easy to organise the data in a Snowflake structure, which helps for analysis.

Indexes: Having indexes in the tables makes retrieving the data from the queries faster. In this schema, CustomerID in the Customer Dimension table, the same as ProductID from Product Dimension, can be indexed as frequently queried columns.

Normalization: This database uses this method to reduce data redundancy and make the data consistent (MKS075, 2023). As normalization is used in this database, the tables will be increased as customers will be stored in a separate table named CustomerDetails table. also,

product Descriptions will be stored in the ProductDescription table, which can prevent duplications in the data.

The Benefits and challenges of setting up a data warehouse:

Benefits:

Centralized data: By centralizing the whole data from all sources like Sales, Customer behavior, and inventory management, we can get a single source for this data warehouse that can be analyzed for the SMP for the business.

Historical Analysis: By storing the past data, SMP can analyze the trend for sales and can make long-term decisions for the improvement of the store's performance.

Increased operational effectiveness: SMP can locate the areas where they can cut expenses, control the inventory, and improve supply chain coordination.

Enhanced profitability: By setting up a data warehouse SMP can have a detailed understanding of the prices of the products and other things and optimize promotion.

Challenges:

Data Quality: SMP can face challenges while implementing the data warehouse, like data cleaning and transforming, as accuracy and consistency are important in the data.

Data Integration: Integrating data from different sources like POS and the customer database will be difficult, so for this, the ETL (Extract, Transform and Load) process will be required.

Scalability: As the SMP store locations increase, the data also increases, so the data warehouse should manage a large amount of data as stores expand.

Data Security: As the data warehouse has the data of every individual and organization, data privacy and security play an important role in storing the data.

Data Capture, Cleaning, Transformation, and Integration Techniques:

1. Data Capture:

ETL Procedures:

a. ETL (Extract, Transform and Load) procedures are important for gathering data from different sources and transforming the data from the source to the data warehouse. (raman_257, 2023)

b. During the extraction, the data can be retrieved from sources using different mechanisms, such as data replication and file transfers.

c. Moreover, during this extraction phase, the data will be cleaned and filtered for future use.

d. This ETL procedure can take the POS transaction data from the databases and perform the data quality checks for the transformation in the future.

Database Replication:

a. This method creates a copy of the database, like POS, in a separate environment designed for the data warehouse.

b. There are a few more approaches to implement this method like transactional replication and changing data capture.

c. Before loading into the data warehouse, data like POS transactions will be replicated from the primary transactional database to a staging area and processed further.

2. Data Cleaning:

Data profiling:

a. As the data may contain some null values or missing values, improper data type and names of the columns and rows, the entire source data should be analyzed for implementing this method. (talend.com, 2024)

b. There will be some tools that give the metrics and reports on the data quality that can be used, so those tools are required for performing this profiling.

c. In this data warehouse, if the customer data is profiled, the results may be duplicates, missing values in address or birth dates or incorrect information. So these things will be marked for further cleaning.

Data Scrubbing:

a. As the profiling shows the problems, the data should be cleaned with the help of scrubbing procedures.

b. A few changes are required, like removing null, substituting values, and changing the data format.

3. Data Transformation:

Data mapping:

a. This technique must map the data fields from source to destination in this data warehouse schema.

b. From sources and destination, the mapping rules are set for the converting, cleansing and structuring data into a usable format (Tibco, 2024). For example, product description fields from source to warehouse will be mapped, and a specialised ProductDescription table ensures that product data is shown consistently.

Data aggregation:

a. Data aggregation is summarising the data for creating higher level aggregates, which involves calculating sums, averages, counts, and other functions using particular derived dimensions.

b. For instance, daily transactions can be summarised monthly or yearly with store or product categories for analysis.

4. Data Integration:

Master data management:

a. This will be done by creating a single and trustworthy master data entity like customers, products, and stores around multiple source systems. (Qlik, n.d.)

b. To guarantee data integrity, this technique includes finding and solving the conflicts, duplication, and inconsistencies in required entities, like removing duplicate customer data from multiple source systems with the help of established match criteria, which makes individual customer data in the data warehouse.

Data deduplication:

a. This technique identifies the duplicates from single/multiple sources and eliminates them. (Sagacity, n.d.)

b. This will be done using Matching Algorithms and established deduplication rules to identify and eliminate redundant data. For instance, deduplicating products based on their names and descriptions to store a single and consistent product master.

TASK 2: IMPLEMENTATION:

Building the Data Warehouse: SNOWFLAKE SCHEMA

Database name : gubba_bhanu_shankar

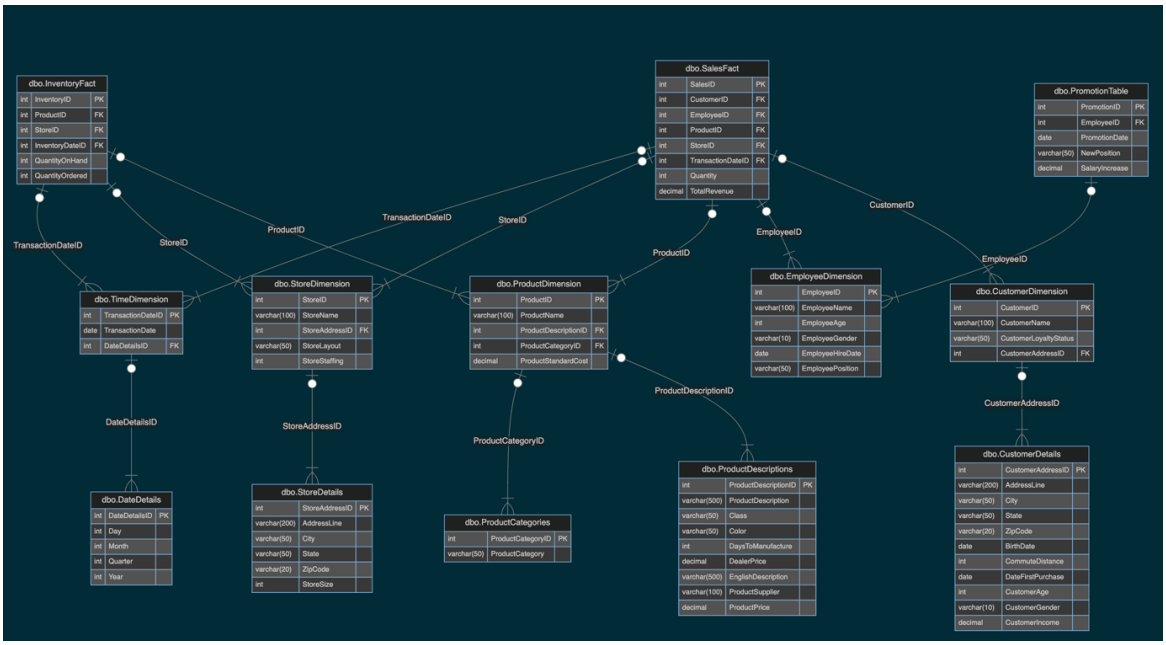

Fig 2: Snowflake Schema

The above figure shows a snowflake schema diagram, which is the same as a star schema; however, in the snowflake, the dimension tables will be further normalized into sub-dimensions and form a snowflake pattern.

As the data warehouse has been built using the SQL server queries, we need first to build the dimension tables, and then fact tables, as the dimension tables will have the primary keys.

According to the scenario, there should be dimensions of Customer, Product, Store and POS transactions. So accordingly, the dimension tables have been created; let's see details about these dimensions:

There are five dimension tables with five sub-dimensions.

1. CustomerDimension: This table holds the data about the customers visiting the store. This customer data will help SMP Retail to make the analysis of customer behavior, customer segments and also purchasing patterns based on loyalty status. As the CustomerDetails is the sub-dimension of the CustomerDimension, it has the attributes related to the customer address like Address line, City, State, Zip Code, and other customer details that can be seen in the above diagram.

2. ProductDimension: This table holds the data about the products in the store. The product data helps the SMP understand the product performance and identify popular and unpopular products to analyse product trends based on the suppliers. The ProductDescriptions table has detailed information about the product description like Class, Color, DaysToManufacture and other details. Whereas ProductCategories tables have only two values, they are ProductCategoryID and ProductCategory. Those two tables act as the sub-dimensions for the ProductDimension table.

3. StoreDimension: This table holds the data about the Store, which helps to compare the performance of the stores in different locations. Moreover, with the help of this data SMP can identify factors affecting sales like size and staffing. Since StoreDetails has data about the store, like AddressLine, City, State and Zip Code, acts as a sub-dimension for the StoreDimension table. From this, they can make informed decisions regarding store operations.

4. TimeDimension: This table holds the data about time-related information like transactions. This makes to help sales trends over time, identifying seasonal sales and comparing store performance across the different periods.

5. EmployeeDimension: This table holds the Employee data for the analysis of employee performance, sorting out top employees by their performance and understanding the impact of promotions on sales. The PromotionTable stores the promotion details like new positions and salary increases that can be analysed in joining with EmployeeDimension and SalesFact tables.

As the dimensions and sub-dimensions are ready, we must create facts accordingly. So, in this case, we create two essential fact tables for the SMP Retail store. They are SalesFact and InventoryFact:

1. SalesFact: This table holds the data about the sales which is the central fact table in the schema joining all the dimension tables. With the help of this particular table, SMP can gain insights into sales performance, customer behavior, product trends and store operations.

2. InventoryFact: This table holds the data about the store's stock, which helps monitor the inventory and optimise and identify overstocks in the different stores across the state.

These facts and dimensions are the essential tables for creating the data warehouse for SMP Retail. The Snowflake divides the dimensional tables into hierarchical tiers, enabling effective Sales and Inventory analysis in SMP.

TASK 3

Emerging trends and Ethical concerns:

Emerging trends:

A. Integration with Business Intelligence (BI) tools:

Various BI tools like Microsoft Power BI and Tableau can be linked to the Data warehouse for cleaning the data and creating interactive modern dashboards for visualising the data. Data cleaning and Data visualisation can make the data more reliable. For now, I am using the Power BI tool for cleaning the data and then creating the Dashboard with the same data. The dataset is an experimental dataset taken from Kaggle.

Below is the link of uncleaned dataset:

https://docs.google.com/spreadsheets/d/1BcQbnSL4F4uj4ybFrqgHQxpbkREC8mkO/edit?usp=sharing&ouid=105791675541851059979&rtpof=true&sd=true

The dataset is about “HR Data”, which includes the data about employees like employee count, gender, age, job role, marital status, and educational background. This dataset shows the relationship between these variables and the attrition rate, as this dataset also includes data about the employee attributes and variables that could affect employee attrition.

Now we can see the data cleaning and visualising process step by step process:

1. First, open the Power BI application on the device and click “Get data” in the upper navigation bar. It gives options for file formats like Excel, CSV, SQL server, etc.

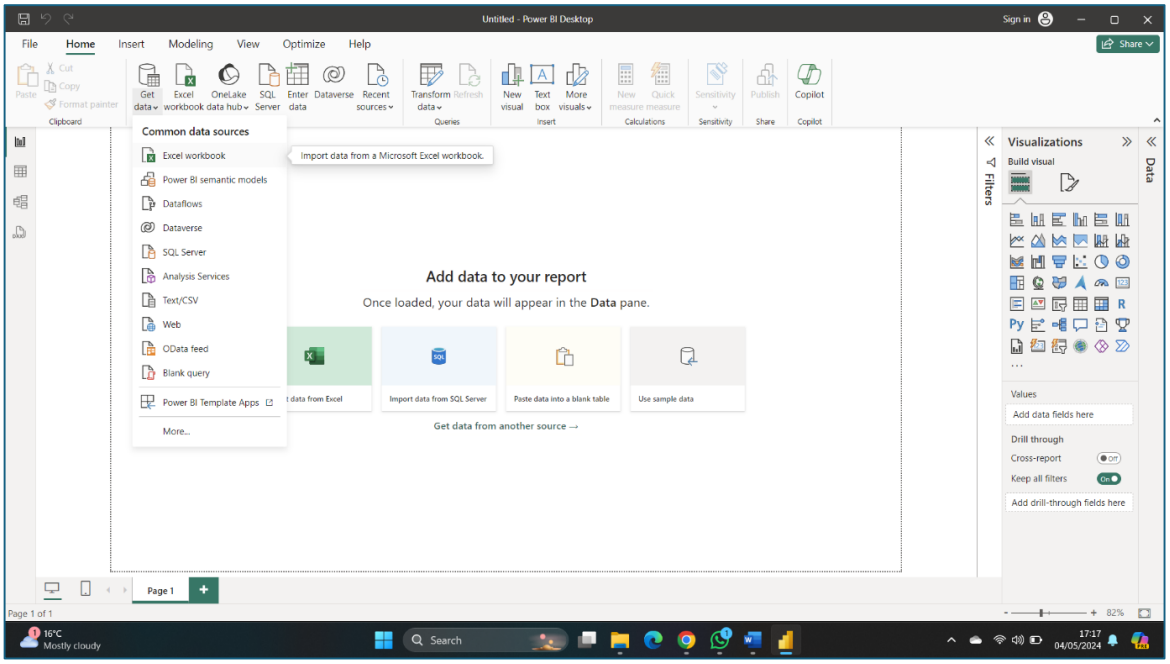

Figure 3

2. As the experimental dataset is in Excel format, we chose that, and then it shows the sheets in that file. Required sheets can be selected and click on the “Load” button.

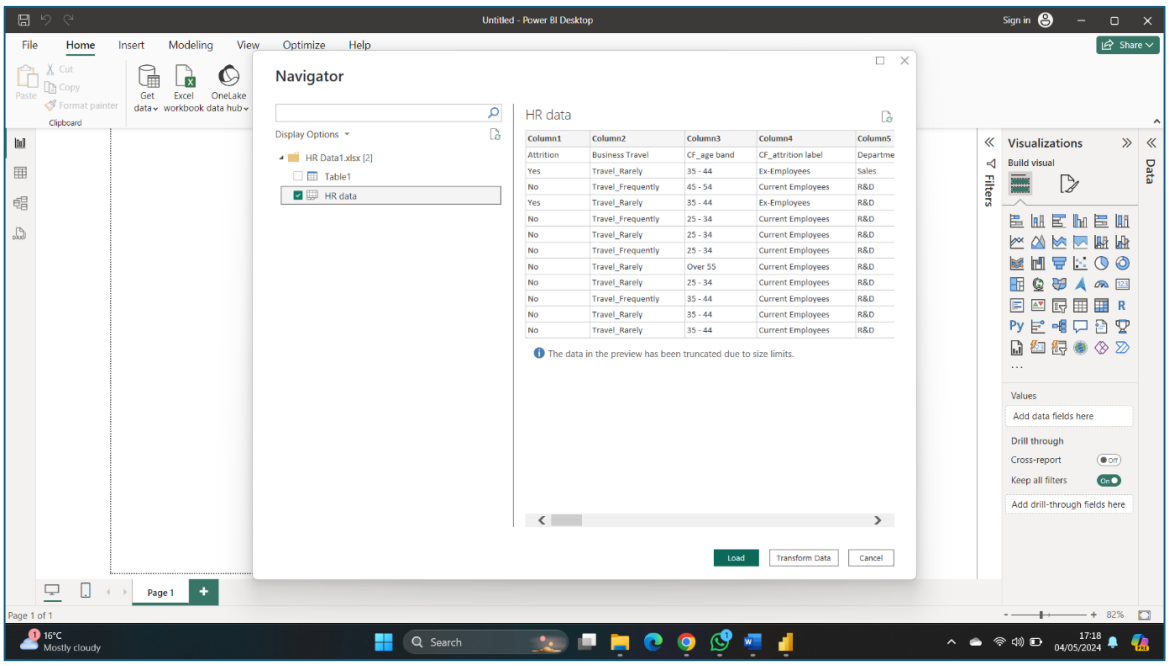

Figure 4

3. After loading the data, it redirects to the Power BI homepage, where there are a few options like the Report page, data page (Table view) and Data flow, which can be seen in the left

sidebar of the home page. As the report page is used to visualise the data, the table view is used to add any new measures, which can be seen below.

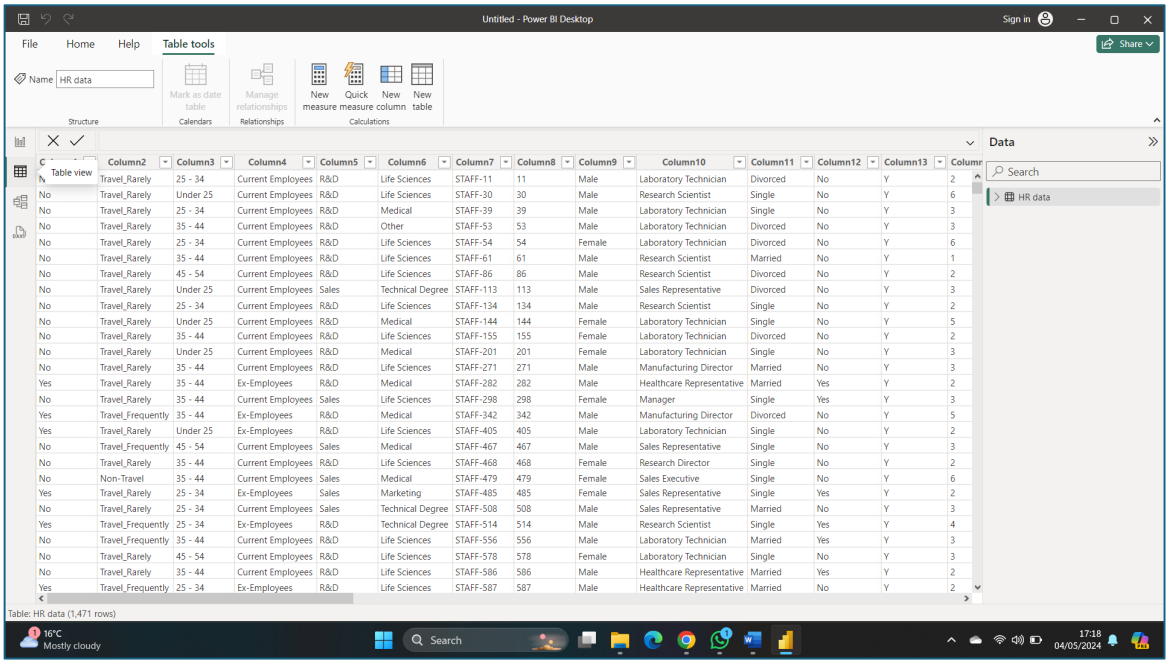

Figure 5

4. We can start cleaning the data by clicking the “Transform data” option on the report page.

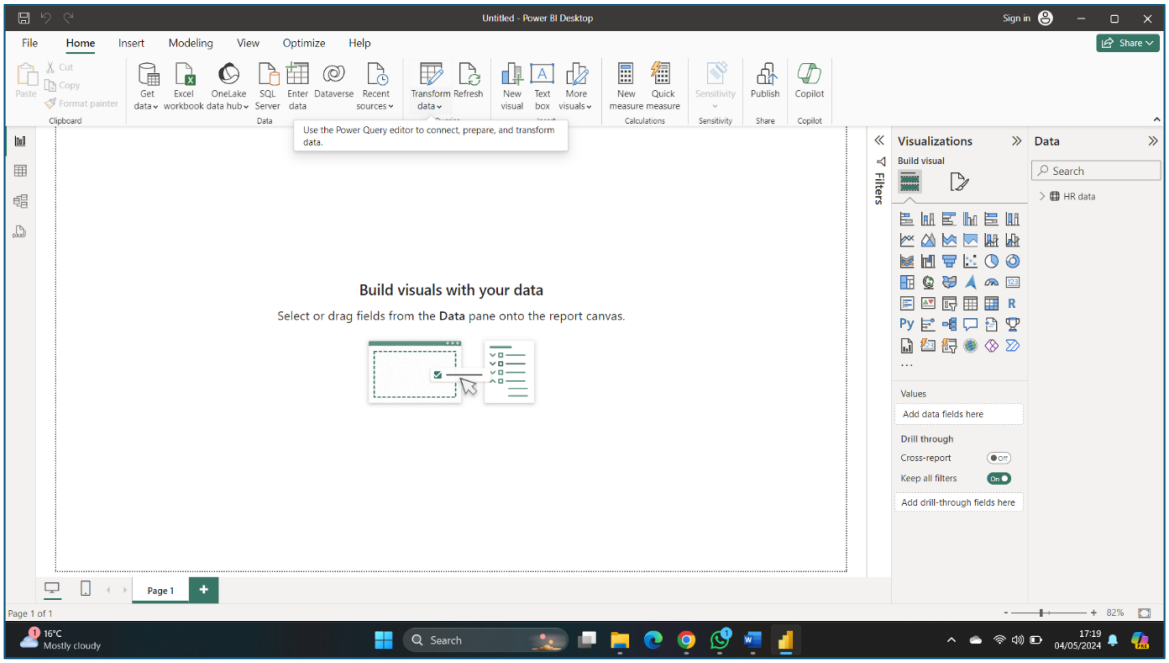

Figure 6

5. After clicking that, it redirects to the Power Query editor, where we can clean the data accordingly, and the cleaning steps can be seen on the right side of the page under the heading "Applied Steps", which can be seen in the below image.

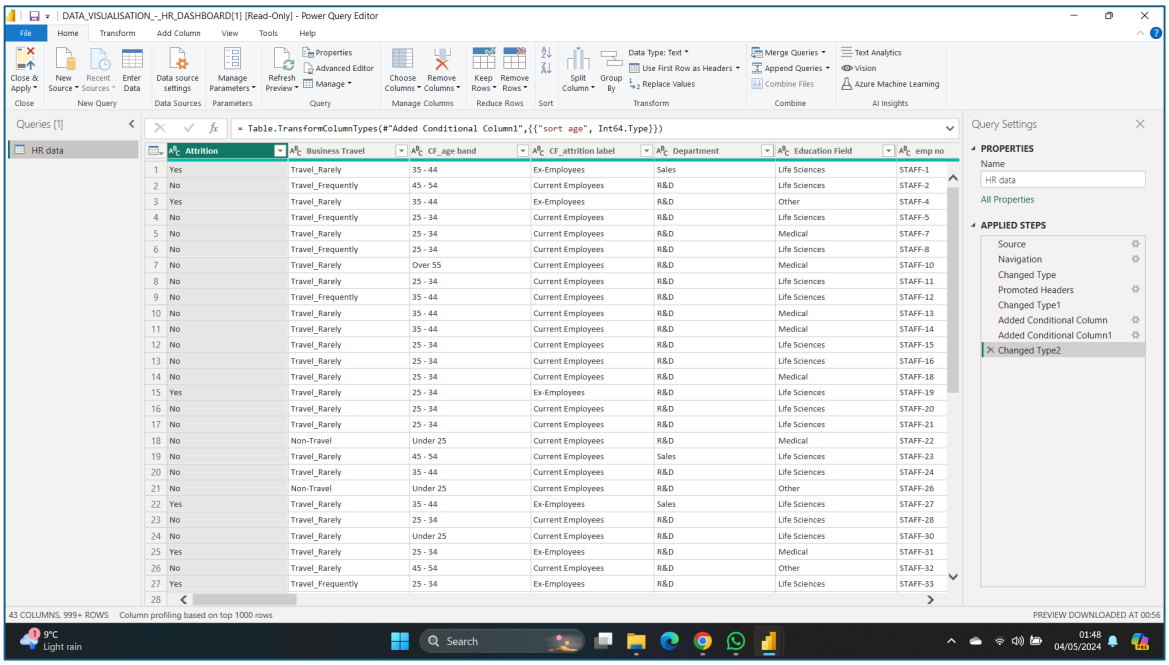

Figure 7

6. The above data has been cleaned in different forms like changing data type, which can be helpful for metrics in the visualisation; if the data type is wrong, there will be some errors during the visualisation, promoting headers, adding conditional columns for the cards for accurate results and other cleaning steps can be followed based on the requirements.

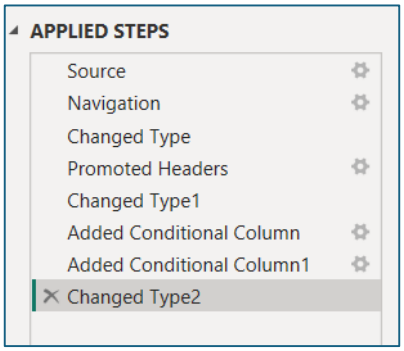

Figure 8

Below is the link to cleaned dataset:

https://docs.google.com/spreadsheets/d/17wldMdZs2Gbs1APaGBxZomOSHthAbXAT/edit?usp=sharing&ouid=105791675541851059979&rtpof=true&sd=true

7. After the cleaning is done, for the visualization, we can choose the option “Close and Apply” to go back to the Power BI home page.

8. Visualisation can be done for the cleaned data by using the tools in the application, like the below dashboard, Cards, charts, and metrics have been used to show and analyse the data in detail.

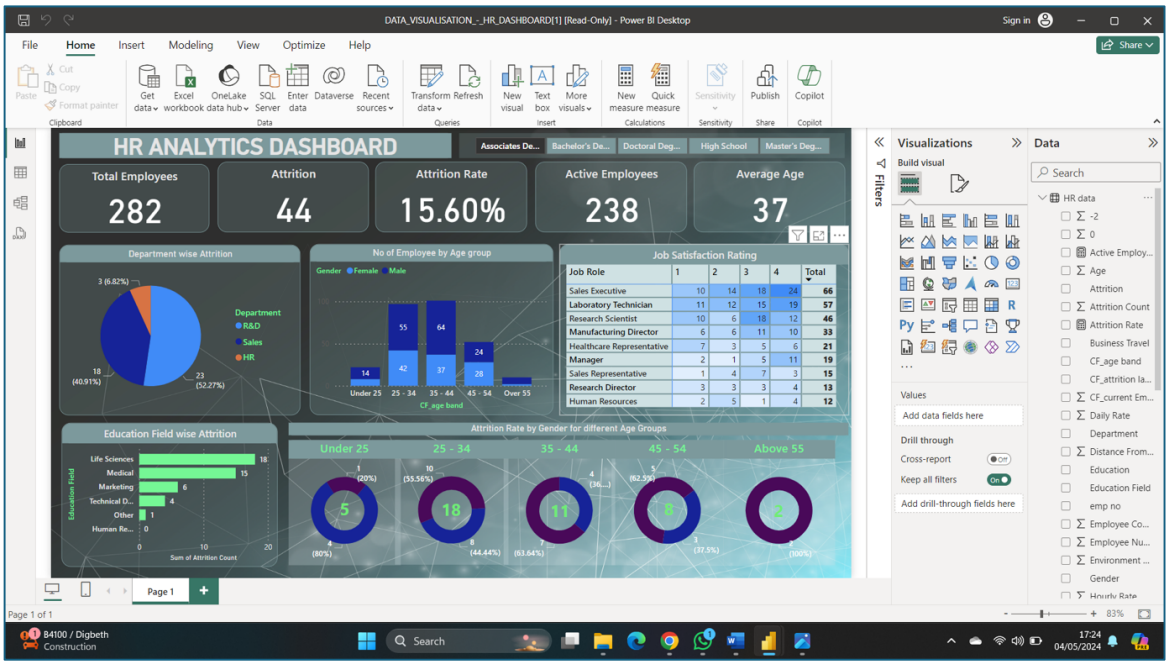

Figure 9

Moreover, different types of charts have been added to the dashboard to analyse the data in different ways, like a pie chart for “Department wise Attrition”, bar chart for “Number of Employees by Age Group”, heat map for job Satisfaction ratings and similar donut charts for “Attrition Rate by Gender for different age groups. “Finally, adding a heading and slicer to visualise the data for a particular education field selected in the dashboard.

Using this application, the following dashboard has been created, which helps the data be visualised in different forms by picking certain fields, years, and ages to view graphically.

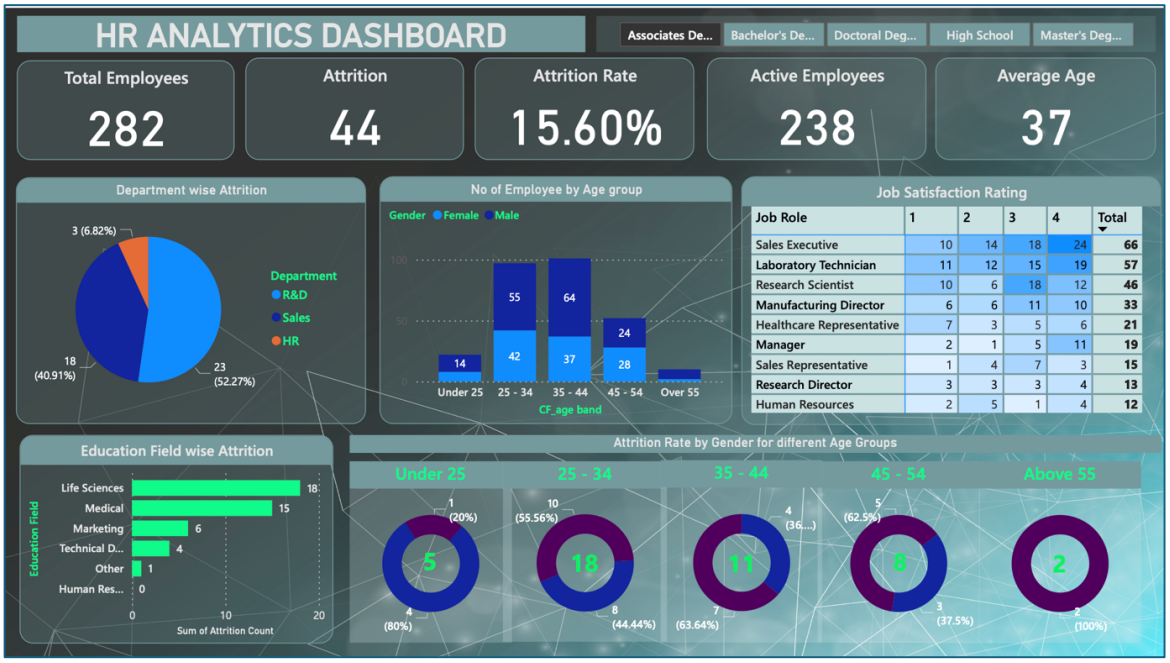

Figure 10

As the data can be cleaned and visualised by loading an Excel file, the same thing can be done by connecting the SQL server Database directly to the application. Let us see how we can connect it:

1. By clicking the “Get data” option in the application, we can see the option “SQL Server database.”

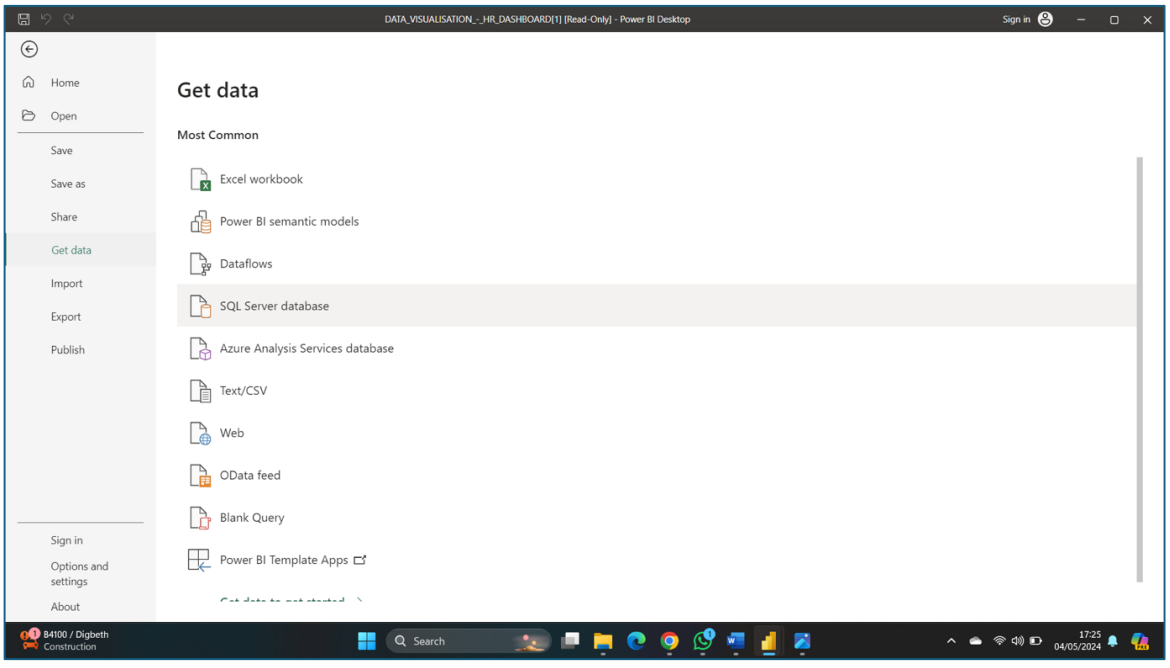

Figure 11

2. Choosing the option opens a window asking for database information like Server, Database name, which is optional and data connectivity mode, which we can see below.

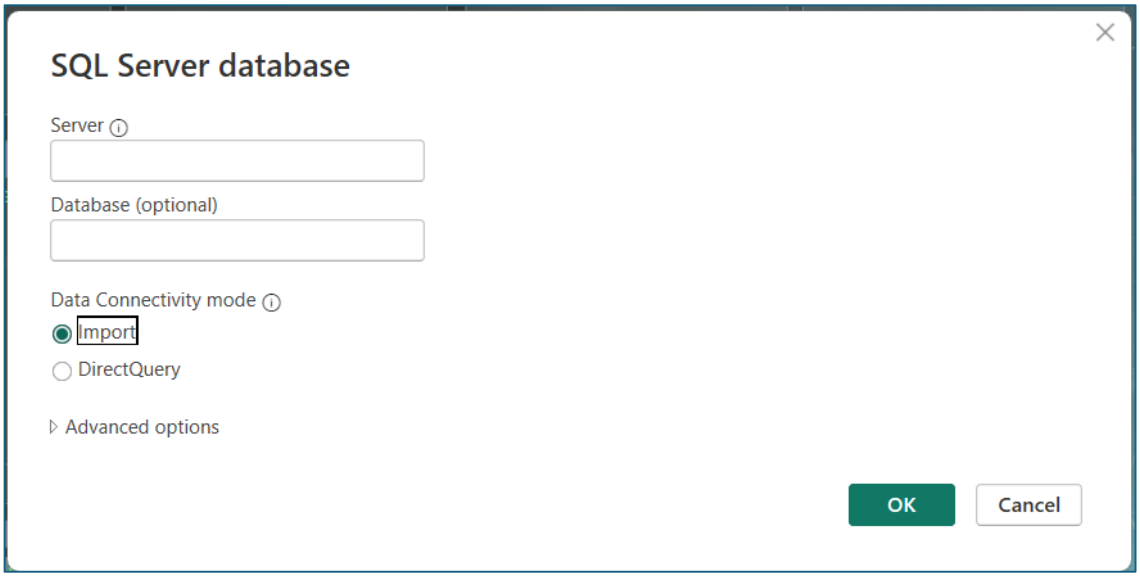

Figure 12

3. After giving the information, it again shows another window where it asks for the credentials for connecting to the database like Username and Password.

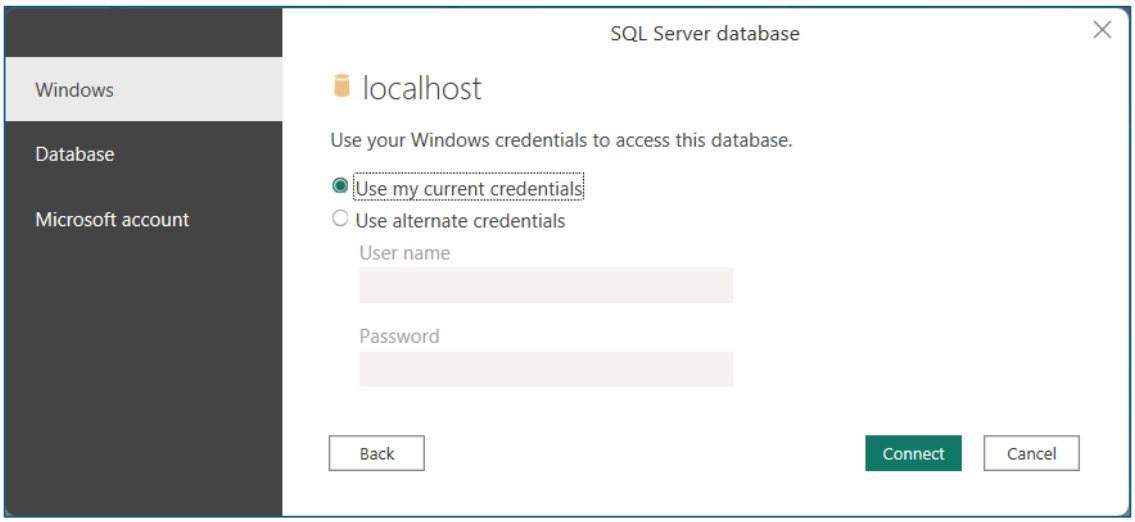

Figure 13

4. By clicking the “Connect” button, the database will be connected to the application and is ready for cleaning and visualisation.

Applications of Machine learning or Big Data technologies:

Machine Learning is becoming popular in today's data warehouse, which can capture a large amount of data from various sources and store it in a single platform for easy retrieval (Weldon, n.d.).

Multiple benefits can be taken if ML works with Data, like the ability to scale up and handle a large amount of data, including different data types like structured, unstructured, and semi-structured. (Perrow, n.d.)

As machine learning tools can be quickly adapted to the current technology and trends. The companies can build the data according to the new trend in the algorithm, and the ML environment should make the adjustments required.

These algorithms can be taught on the past data for future demands and optimising the inventory levels.

Ethical concerns:

A. Data Privacy and Security:

SMP will collect and store the customer's data, and it should take precautions to avoid any misuse and protect the customer data like their personal information and transactional history, which has the details of their cards.

So, by developing strong access control and data encryption, SMP can prevent data loss. (JUREK, 2023)

B. Bias:

Identifying bias in a data warehouse needs a systematic approach that covers the entire data lifecycle from data collection to data consumption. (Tamba, n.d.)Bias can affect the validity and reliability of the data.

Strategies for Mitigating Ethical Risks:

1. For the data collection with the customers, firstly, privacy rules and consent are created.

2. Mitigating bias in the data warehouse requires technical measures to minimize the sources of bias. (Venkaata, n.d.)

3. Implementing data governance practices that have transparency and compliance with the project.

4. Awareness programs and providing training for the staff on data privacy and security best practices.

By addressing these ethical concerns, SMP Retail can control the data while maintaining ethics.

Bibliography

MKS075. (2023, Feb 21). *Difference between Normalization and Denormalization*. Retrieved from GeeksforGeeks: https://www.geeksforgeeks.org/difference-between-normalization-and-denormalization/

Tibco. (2024). *What is data transformation?* . Retrieved from Tibco : https://www.tibco.com/glossary/what-is-data-transformation#:~:text=Data%20transformation%20is%20the%20process,that%20of%20the%20destination%20system.

Qlik. (n.d.). *Data Integration* . Retrieved from Qlik: https://www.qlik.com/us/data-integration

Sagacity. (n.d.). *What is Data Deduplication? How to Improve Data Uniqueness* . Retrieved from Sagacity: https://www.sagacitysolutions.co.uk/about/news-and-blog/data-deduplication/#:~:text=Deduplication%20can%20reduce%20redundant%20data,as%20long%20for%20data%20retrieval.

Weldon, D. (n.d.). *Machine learning and data warehousing: What it is, why it matters*. Retrieved from TechBeacon: https://techbeacon.com/enterprise-it/machine-learning-data-warehousing-what-it-why-it-matters

Perrow, M. (n.d.). *Essential data types: What you need to know*. Retrieved from TechBeacon: https://content.microfocus.com/guide-to-modern-data-warehouse-tb/essential-data-types?lx=hLKzJ4&utm_source=techbeacon&utm_medium=referral&utm_campaign=%207014J000000dVnvQAE&_ga=2.82265492.306219457.1715543451-653813248.1715543449

JUREK, W. (2023, November 01). *Data Warehouse Security: A Comprehensive Guide*. Retrieved from cobry: https://www.cobry.co.uk/how-to-secure-your-data-warehouse#:~:text=Data%20encryption%20is%20a%20fundamental,it%20in%20its%20data%20store.

Tamba, A. (n.d.). *How can you identify and mitigate bias in data warehousing and business intelligence?* . Retrieved from linkedin: https://www.linkedin.com/advice/0/how-can-you-identify-mitigate-bias-data-warehousing-

cho1c#:~:text=Bias%20in%20data%20warehousing%20and%20BI%20is%20any%20deviation%20from,%2C%20misleading%2C%20or%20unfair%20outcomes.

Venkaata, S. R. (n.d.). *How can you identify and mitigate bias in data warehousing and business intelligence?* Retrieved from linkedin: https://www.linkedin.com/advice/0/how-can-you-identify-mitigate-bias-data-warehousing-cho1c#:~:text=Bias%20in%20data%20warehousing%20and%20BI%20is%20any%20deviation%20from,%2C%20misleading%2C%20or%20unfair%20outcomes.

raman_257. (2023, Feb 02). *ETL Process in Data Warehouse* . Retrieved from geeksforgeeks: https://www.geeksforgeeks.org/etl-process-in-data-warehouse/

talend.com. (2024). *What is Data Profiling?* . Retrieved from talend: https://www.talend.com/uk/resources/what-is-data-profiling/#:~:text=Data%20profiling%20is%20the%20process,then%20leverage%20to%20their%20advantage.

The below link has files of the assignment – Cleaned dataset, Uncleaned dataset, Power Bi file, and SQL Server file.

https://drive.google.com/drive/folders/115790InrvGK8muZm2YRrwaQtr-RE3cNH?usp=sharing

Conclusion:

SMP Retail has hundreds of stores across the state, and each manager makes their own decisions based on the sales data from each store, which leads to inconsistent customer experiences and inventory.

The central data warehouse has been created to solve this issue, combining data from different sources like POS transactions, customer details, and product details. It also stores information in one place so that the SMP will get an idea of what's happening in all the stores. Building the data warehouse includes implementing a snowflake schema, which organises the entire data into fact tables, dimensions, and sub-dimensions, which helps to analyse the useful data. Moreover, with this cleaned data, the visualisation has been done for greater insights. With this, SMP can integrate modern tools for BI, even machine learning or big data technologies in the future. This helps in discovering patterns and predicting demand. However, SMP needs to

be careful with data privacy and security. Creating policies and procedures following them will maintain the trust of the customers. Overall, a data warehouse is an essential step for the SMP in which data can be used as an asset by focusing more on customer's needs. Continuously showing interest in data-driven technologies and following ethical practices, SMP can reach the top level in the retail industry.