CSCI 5408
Data Management, Warehousing & Analytics
Case Study Report

Group 10

Ridham Dabhi, B00853506

Dhruv Tarpara, B00856253

Dhruvesh Patel, B00854218

Dataset Summary

The provided dataset— “Sample Sales Data” [1] gives an insight about the sales done from the year 2003 to the year 2005. It was last updated on November 23, 2016. It contains details of orders and order items. There are 2,823 rows and 25 columns. Each row contains an order line of an order with details like price of the ordered product, their quantity, the date they are ordered and the customer details. Based on the order line item and quantity, the order lines are classified by deal size. The deal size can be Small, Medium, or Large. The dataset contains sales records of vehicles like Cars, Trucks, Trains, Buses, Motorcycles, Planes, and ships. The PriceEach column contains the price of the order product. The Status column shows the status of the order. Its values can be—Shipped, Disputed, Cancelled, In Process, On Hold, or Resolved.
The Qtr_ID, Month_ID, and Year_ID provides the quarter of the year, month and the year the product is ordered, respectively. The ProductLine column contains the type of product. The MSRP and Product Code columns give the Manufacturer Suggested Retail Price and the code to identify the product in the orderline, respectively. The Sales column gives
the total orderline amount. Columns like CustomerName, Phone, AddressLine1, Addressline 2, City, State, PostalCode, Country, and Territory provide the organization’s name, address and contact information. The ContactFirstName and ContactLastName columns give the name of the person from the organization who is placed and is responsible for the order.

Data Warehouse

Data warehousing offers business analysts with systems and tools to effectively organize, interpret, and utilize their data to make business strategy decisions. A data warehouse refers to a database which is kept separately from the operating databases of an organization. Data warehouse systems allow a range of application systems to be integrated. They support information processing by offering a stable platform for the analysis of collected historical data.

Data warehouse is [2]:

- **Subject-oriented**
  A data warehouse is structured around subjects, such as client, product, manufacturer, and revenue.

- **Integrated**
  A data warehouse is generally built by integrating multiple heterogeneous sources, such as relational databases and flat files.

- **Time-variant**
  Data is stored in a data warehouse to provide information from a historical perspective.

- **Non-volatile**
  A data warehouse always consists of a physically separate store of data extracted from the application data contained in the operational environment.

These properties distinguish data warehouses from other data repository systems, such as relational database management systems, and file systems, and transaction processing systems.

Advantages of data warehouse:

- It provides instantaneous access to data from various sources in a single place.
- Historical data can be stored and used to analyse older data and make productive decisions for the future.

Disadvantages of data warehouse:

- Making changes in data types, source schema and ranges of the data in a data warehouse is difficult.
- It is not ideal for unstructured data.

Multidimensional Schema

A data warehouse requires a concise and subject-oriented schema that facilitates effective and efficient analysis. The schemas are designed to address unique requirements of large databases designed for analytical purposes. The most popular data model for data warehousing is the multidimensional model. Such a model can exist in the form of [3]:

- Star schema
  The star schema contains a fact table containing the bulk of data, without redundancy and a set of dimension tables (one for each dimension). The schema graph consists of dimension tables in a radial pattern around the fact table in the center.

- Snowflake schema
  In addition to star schema, the snowflake schema's dimension tables may be kept in normalized form to reduce redundancies. The schema graph consists of central fact table with the dimension tables in a radial pattern around it, which in turn may have other dimension tables around them.

- Fact constellation
  In a fact constellation (also known as galaxy schema), multiple fact tables can exist. Dimension tables can be shared between fact tables.

For the given dataset, we decided to go with the star schema because the star schemas simplify data analysis, i.e., business tools like Cognos and Tableau highly support star schemas for data visualization. It also makes query executions faster and can easily be understood by non-technical businesspeople who might look at the schema.

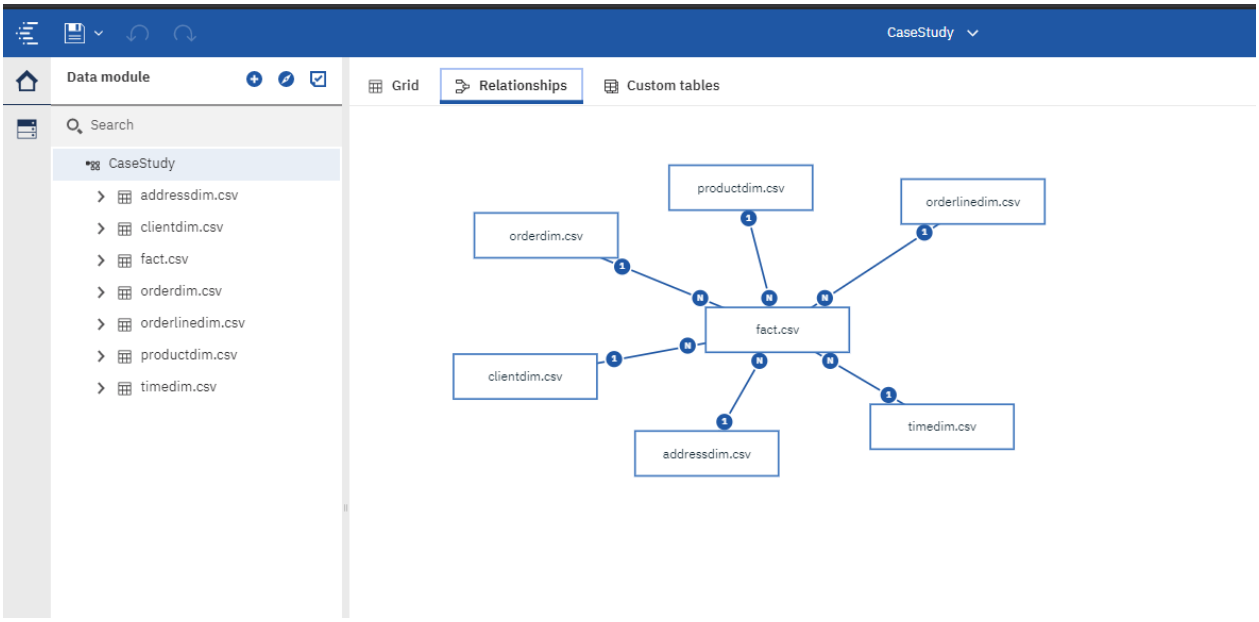

Figure 1. Multidimensional Star Schema

ETL Process for our case study [4]

Extract

We extracted data required for sales visualization from Kaggle [1] in CSV format.

Transform

As a part of cleaning and transforming process, we cleaned the date column as it consisted of redundant data like month and year which already had their separate columns. So, we created a new column storing only the date and not year and month because they had their own columns. Thereafter, we created a fact table using dimension table. To do it we compared rows of dimension table contain same columns in original data table and inserted the primary key values in one different column of the original file. So, we got a CSV file containing the original data plus foreign keys of each dimension table. So finally, we got six dimension tables and one fact table.

Load

As an input we had six dimension tables and one fact table. We had two options for loading data. First, we can import CSV files directly to Cognos BI and create relationships inside Cognos BI itself. And Second, we can load the data using data-source in which we need to store CSV files into MySQL server and create relationships in it. Then using IBM secure gateway, we can connect the local data source into Cognos Analytics and load the data into a data module. We preferred first option, because we were using Cognos Analytics Free version which was providing limited bandwidth and upload capacity, it may take while to upload data in live demo, so we decide to go with it.

Business Intelligence (BI)

Business Intelligence (BI) is a combination of techniques, architectures, technologies, and processes that transforms raw data into useful information that enables business users to make important business-oriented decisions. The eventual goal of BI is to help business users to turn business-related raw data into information to perform critical actions. BI tools are designed to make sense of large amount of raw data that an organization collects over time. It helps to generate interactive reports through which any non-technical person can understand and gain insights of the data.

For the current dataset, we are using Cognos BI Analytics [5]. Cognos is a BI platform that supports analytics, visualization, and interpretation of the data. It also present actionable insights which might help in the decision making.

We found that there are two ways we can upload data—file upload (CSV file) or using server connection (gateway). First, we loaded the CSV files in MySQL server and established relationships among the tables. The, we installed IBM Secure Client Gateway to create data source into Cognos Analytics and the load all data through the gateway. The other way is to import the CSV files directly and create relationships within itself. Due to the limited bandwidth and upload capacity in the Cognos Analytics Free version, importing the cleaned data in the form of CSV files (Fact and dimension tables) is faster.

We separately analysed data and found insights, which we combined into a single Dashboard. We categorised all found insights into nine tabs in Cognos Analytics. We created custom calculations to provide industry level insights.

The following insights were created:

1. Overall Sales of Products based on Location
2. Sales Trend and forecast depending on Time
3. Selling Price Comparison
4. Popular Product based on Location and Time
5. Quantity ordered based on Sales Year
6. Major Clients and their Orders information

Following are images of dashboards that we created and whole dashboard is attached as PDF report as well.

Overview

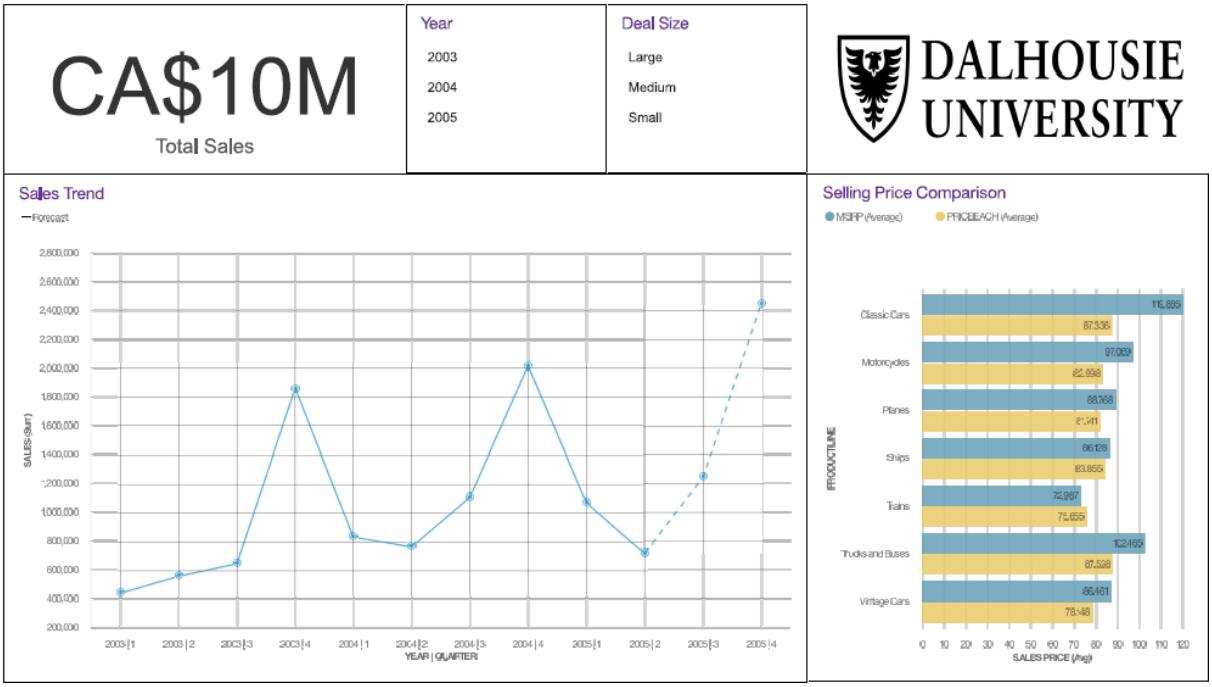

Figure 2. Dashboard Overview Tab

Performance

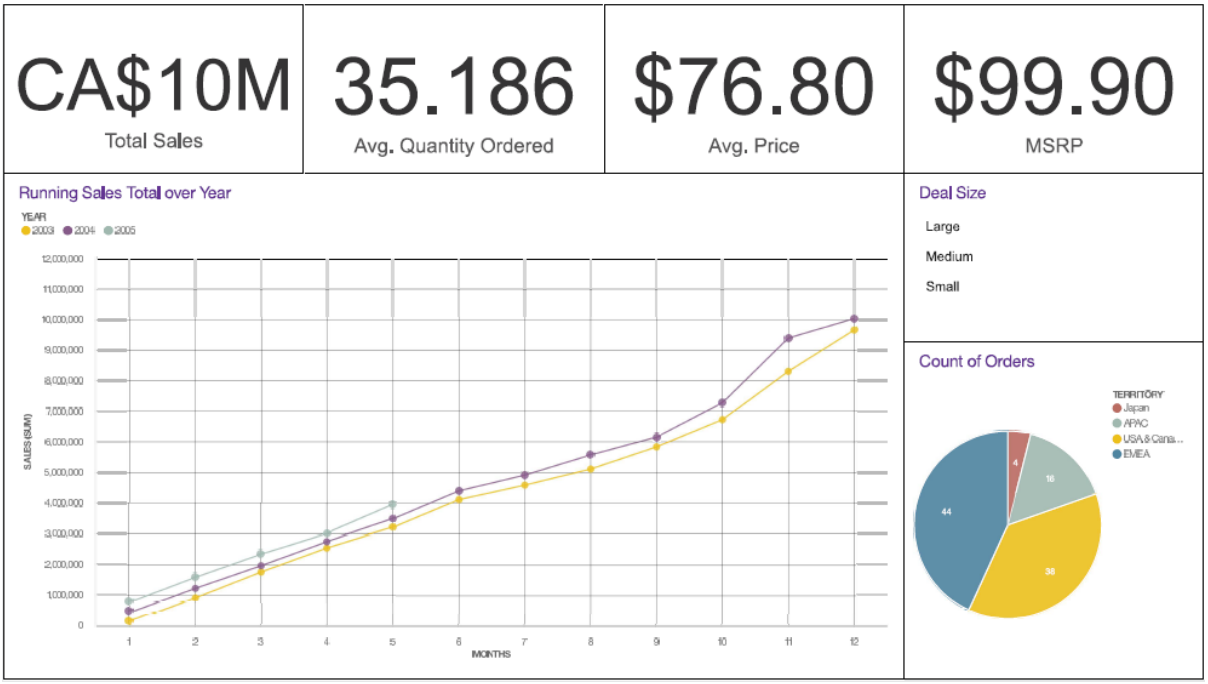

Figure 3. Dashboard Performance Tab

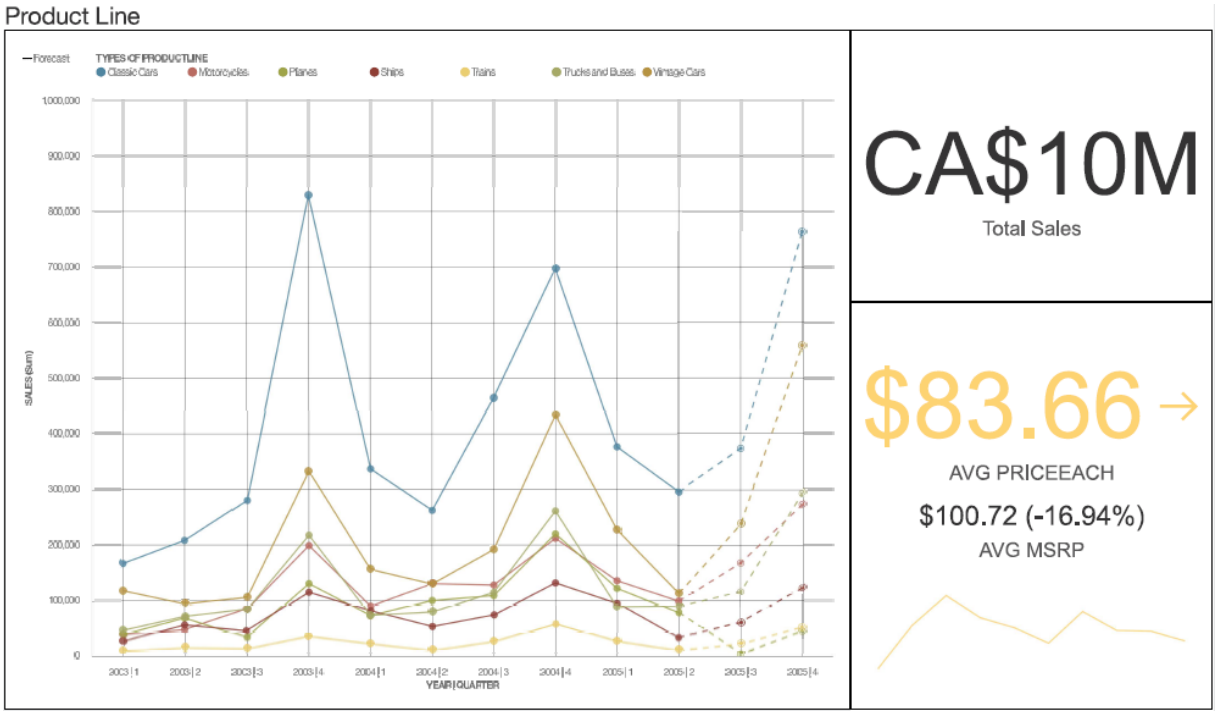

Figure 4. Dashboard Product Line Tab

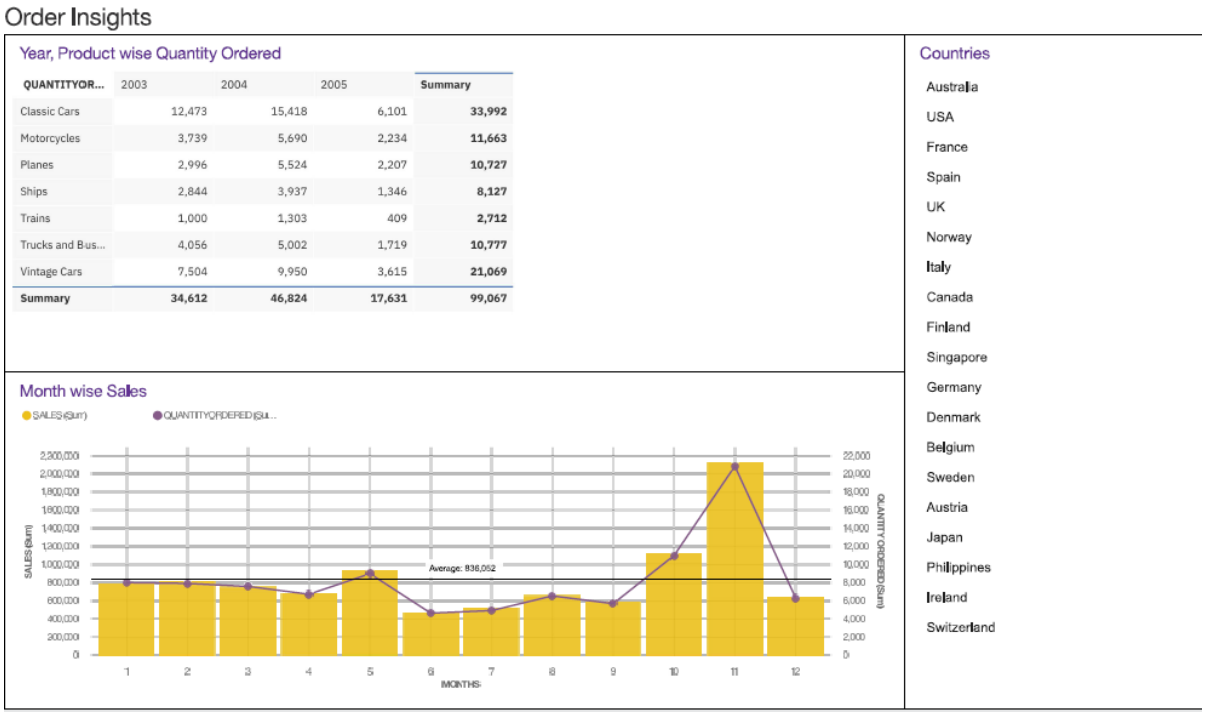

Figure 5. Dashboard Order insight Tab

References

[1] "Sample Sales Data", *Kaggle.com*, 23-Nov-2016. [Online]. Available: https://www.kaggle.com/kyanyoga/sample-sales-data. [Accessed: 03-Mar-2020].

[2] "Data Warehousing - Overview," *tutorialspoint*. [Online]. Available: https://www.tutorialspoint.com/dwh/dwh_overview.htm. [Accessed: 16-Mar-2020].

[3] "Data Warehousing - Schemas," *tutorialspoint*. [Online]. Available: https://www.tutorialspoint.com/dwh/dwh_schemas.htm. [Accessed: 16-Mar-2020].

[4] N. Fatima, "ETL Process and the Steps for its Implementation - Astera Software," Astera, 04-Jul-2019. [Online]. Available: https://www.astera.com/type/blog/etl-process-and-steps/. [Accessed: 24-Mar-2020].

[5] "IBM Cognos Analytics," *IBM Cognos Analytics*, 2018. [Online]. Available: https://www.ibm.com/ca-en/products/cognos-analytics. [Accessed: 24-Mar-2020]