CSCI 5408 Data Management, Warehousing and Analytics

Sales Case Study

April 2, 2020

Group 5:

Smit Panchal, B00828070

Meeta Chanchlani, B00835734

Vivek Sakariya, B00848519

Table of Contents

INTRODUCTION

Data analytics and visualization is important for a business system in many ways and can provide meaningful insights to the business executives. A database can include different fields and the explanation of it can be obtained from the metadata provided on the Kaggle website. There are many ways in which a dataset can be interpreted and the selection of necessary attributes becomes crucial for the further process.

Data Warehousing

Data warehousing is a procedure of acquiring and allocating the data from various sources to provide meaningful insights in the successful working of a business. It is typically used to connect and analyze business data from heterogenous sources. The data warehouse is most intricate part of a Business Intelligence system for further analytics and visualization. It is a mix of technologies and components which assists the relative use of the data. It is an electronic storage of a large amount of information by a business which is designed for query and analysis instead of transaction processing. It is flow where raw data is transformed into information and making it available to users in a timely manner to make a difference [1].

It mainly consists of three steps namely:

1. Data cleaning
2. Data integration
3. Data consolidation

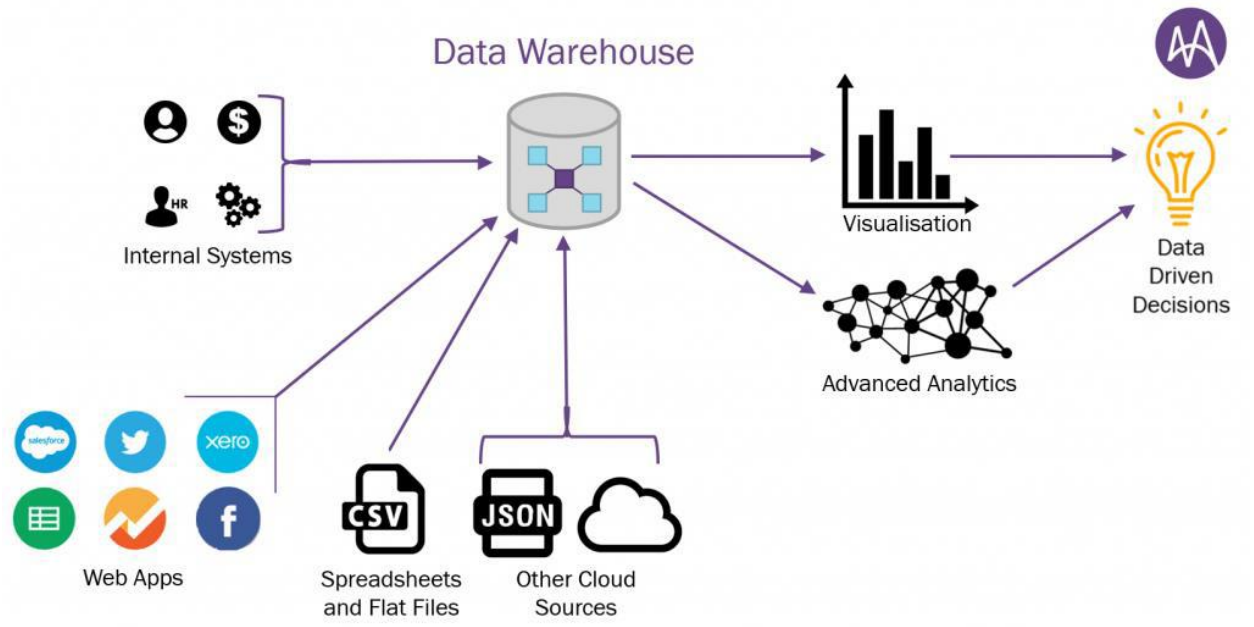


Figure 1: Data warehousing [2]

In figure 1, there is clarification of how a system can be articulated into different ways and how different sources are used to design a data analytics system. There are sources like internal systems, cloud systems, web application and many more which are fed in the schema design process, creating a well-versed informative data warehouse. This warehouse is then used for visualization and gaining meanings in a business setting.

ETL Process

The abbreviation for ETL is Extract, Transform and Load. The following diagram explains the way how the process work in a detailed way.

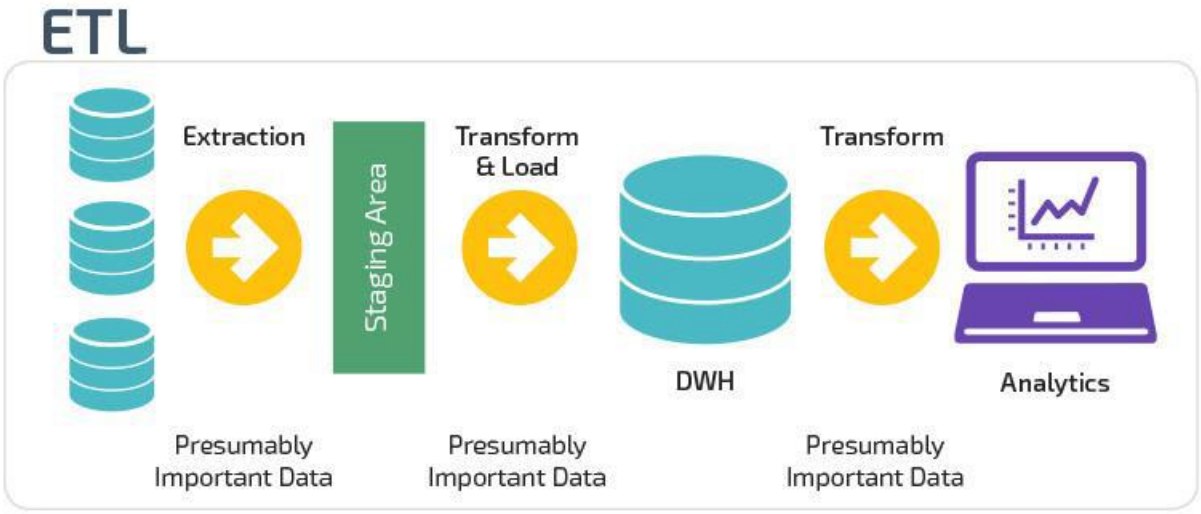


Figure 2: ETL Process [3]

Extract – Collecting data from various sources

Transform – Convert the data into understandable form

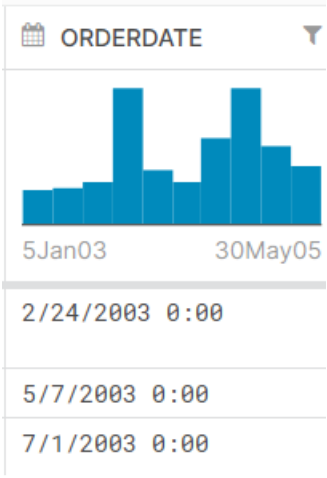
Load – Store the transformed data on the target system

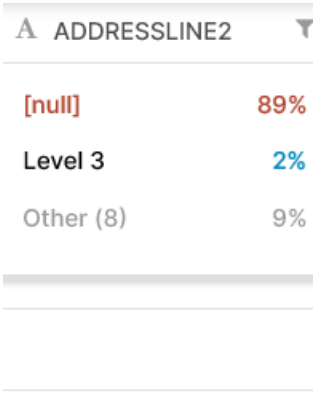
Dataset Exploration

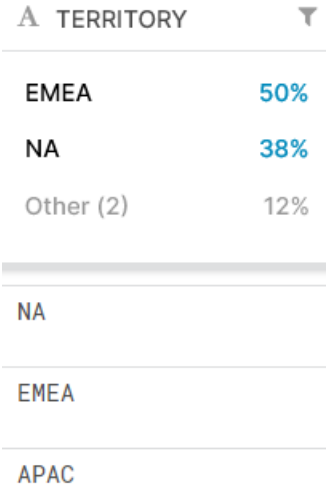
The data is collected from https://www.kaggle.com/kyanyoga/sample-sales-data [2] which is a Kaggle platform having various kinds of data regarding to different subjects. There are a total of 25 fields in the sales data. There are various fields which can define the relationships and further be used for visualization. The dataset is about an Automobile Sales and is inspired for retail analytics, originally written by Maria Carina. There are 2823 records of different orders. The data is collected between the year 2003 and 2005. The number of unique products are 109 distributed among 19 countries and 92 unique customers. There are 307 unique orders among 4 territories.

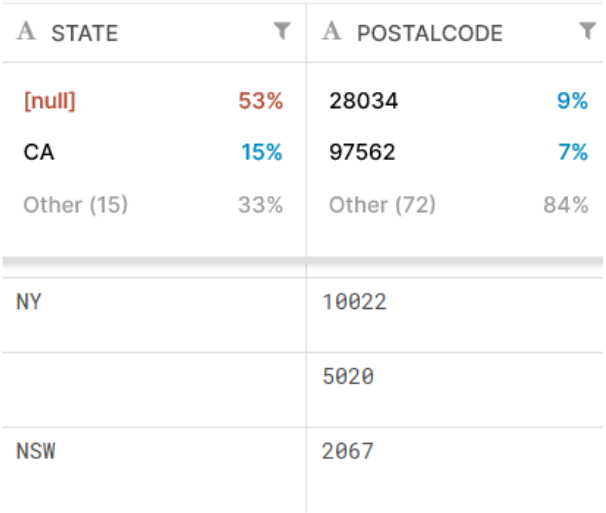
Data Cleaning

The below figure defines certain fields which are to be cleaned having some errors or might need some cleaning in the values.









The orderdate field contains the time which is not necessary and the addressline2 has null values in it. The territory, state and postalcode also has been cleaned which had null and empty values in it. The cleaning is done by a python script which used pandas library in it.

Schema Design

The initial schema design was created as a star schema keeping the basic information of fact into the consideration. It is clearly defined in the figure about the various dimensions of raw data that was originally collected.

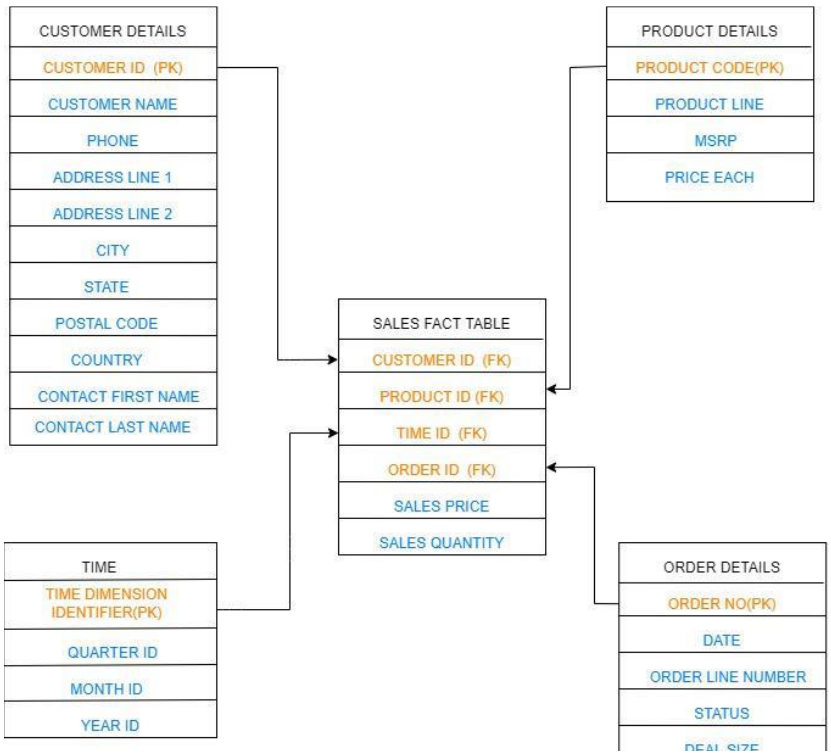


Figure 3: Initial Design (Star Schema)

The final schema which is chosen as a structural base for the development of a data warehouse is depicted in the figure below. There more dimensions added related to the fact table. Therefore, it makes a system more simplified to collect simpler data modules from the warehouse.

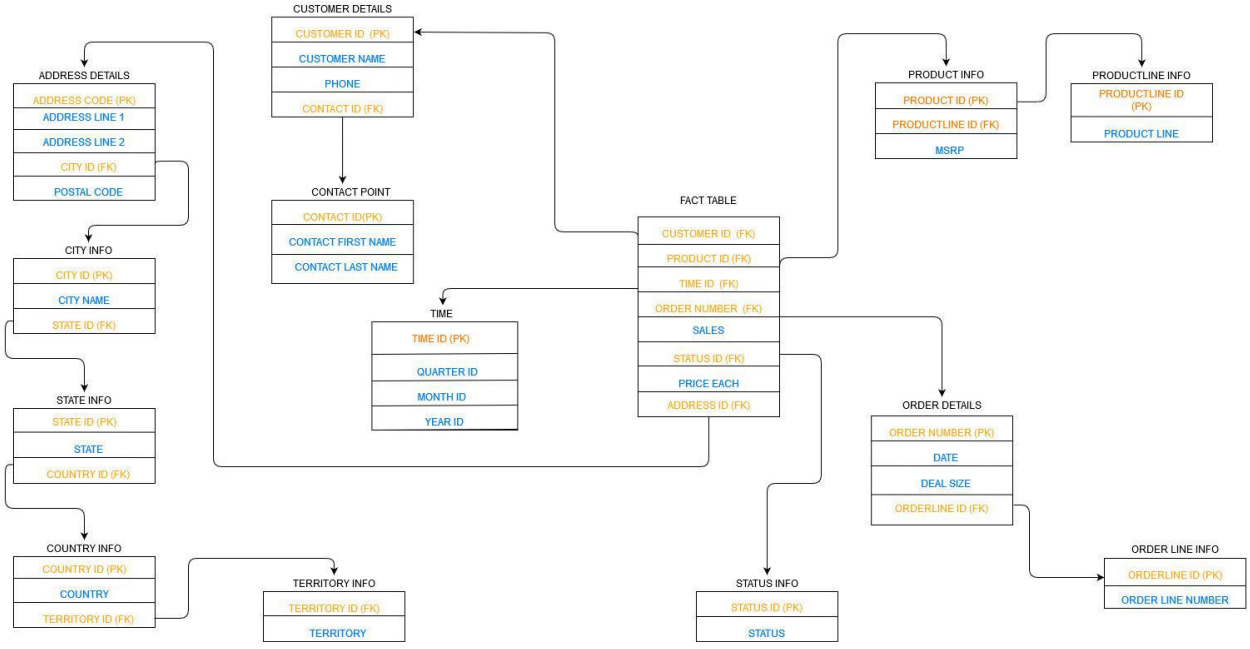


Figure 4: Final Design (Snowflake Schema)

Python Script

```
import pandas as pd
readcsv = pd.read_csv('sales_data_sample.csv', encoding="latin1")
readcsv['PRICEEACH'] = round(readcsv['PRICEEACH'], 2)
readcsv['SALES'] = round(readcsv['SALES'], 2)


readcsv['PHONE'] = readcsv['PHONE'].str.replace('.', '')
readcsv['PHONE'] = readcsv['PHONE'].str.replace('+', '')
readcsv['PHONE'] = readcsv['PHONE'].str.replace('-', '')
readcsv['PHONE'] = readcsv['PHONE'].str.replace('(', '')
readcsv['PHONE'] = readcsv['PHONE'].str.replace(')', '')
readcsv['PHONE'] = readcsv['PHONE'].str.replace(' ', '')


readcsv['ORDERDATE'] = readcsv['ORDERDATE'].str.split(" ",1,expand = True)
df = readcsv.fillna({
'ADDRESSLINE2' : 'NA',
'STATE' : 'NA',
'POSTALCODE' : 'NA',
'TERRITORY' : 'NA'

})

mywrite = pd.DataFrame(df, columns=['ORDERNUMBER','QUANTITYORDERED','PRICEEACH',
                                    'ORDERLINENUMBER','SALES','ORDERDATE','STATUS',
                                    'QTR_ID','MONTH_ID','YEAR_ID','PRODUCTLINE',
                                    'MSRP',
                                    'PRODUCTCODE','CUSTOMERNAME','PHONE',
                                    'ADDRESSLINE1',
                                    'ADDRESSLINE2','CITY','STATE','POSTALCODE',
                                    'COUNTRY',
                                    'TERRITORY','CONTACTLASTNAME',
                                    'CONTACTFIRSTNAME',
                                    'DEALSIZE'])
# TIME INFO

time = pd.DataFrame(newreadcsv, columns=['QTR_ID','MONTH_ID','YEAR_ID'])

t = time.drop_duplicates(subset='MONTH_ID',keep="first")
t.insert(0, 'TIMEID', range(5900, 5900 + len(t)))
# PRODUCT INFO

productDetails = pd.DataFrame(newreadcsv, columns=['PRODUCTCODE', 'MSRP'])

pro = productDetails.drop_duplicates(subset='PRODUCTCODE',keep="first")
pro.insert(0, 'PRODUCTLINEID', range(80000, 80000 + len(pro)))
```

The above code depicts how python segregates a raw table into different dimension tables and a single fact table.

Visualization

Dimension Modelling i.e. designing the star or snowflake schema and populating the data warehouse is not sufficient. We need a tool to visualize the data. High level Managers in any company are interested in the overall statistics of the company. A marketing Vice President is interested in knowing how much did a new product generate. A marketing manager is interested in the sales statistics of the company whereas a financial controller is interested in knowing what expenses were incurred by the company. In short, all of them are interested in determining the performance of the company rather than daily transactions [4].

IBM Cognos

For this case study, we used IBM Cognos for visualization of the data. We loaded the csv's generated in the previous schema and built relationship between all entities as defined in the final schema. We added the cardinality for each relationship as seen in Figure 5.

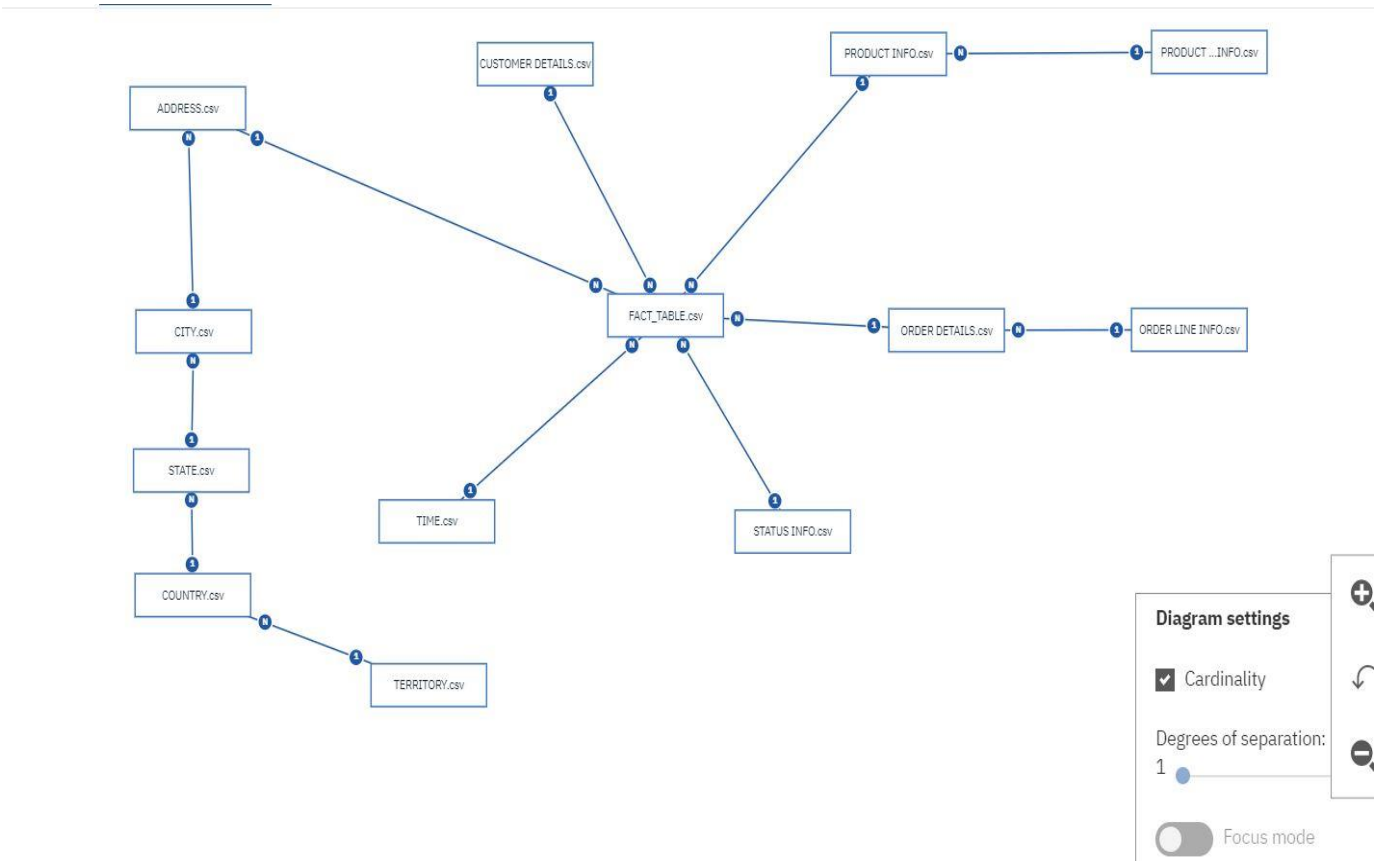


Figure 5: Snowflake schema in IBM Cognos

We then created a dashboard and added different tabs to it to visualize the key performance indicators as indicated in figure 6,7,8,9 and 10 [5].

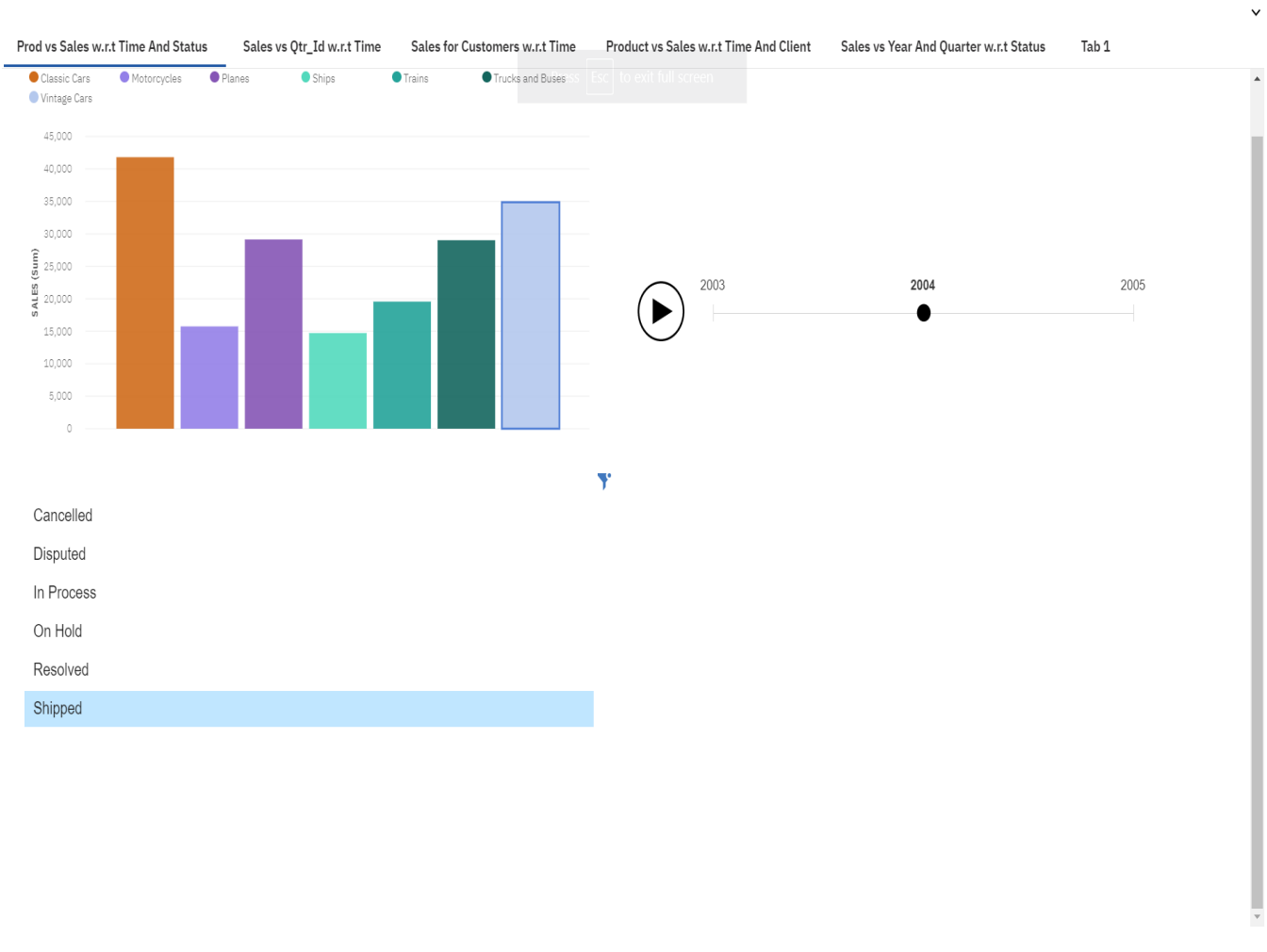


Figure 6: Product Line vs Sales with respect to Status and Year

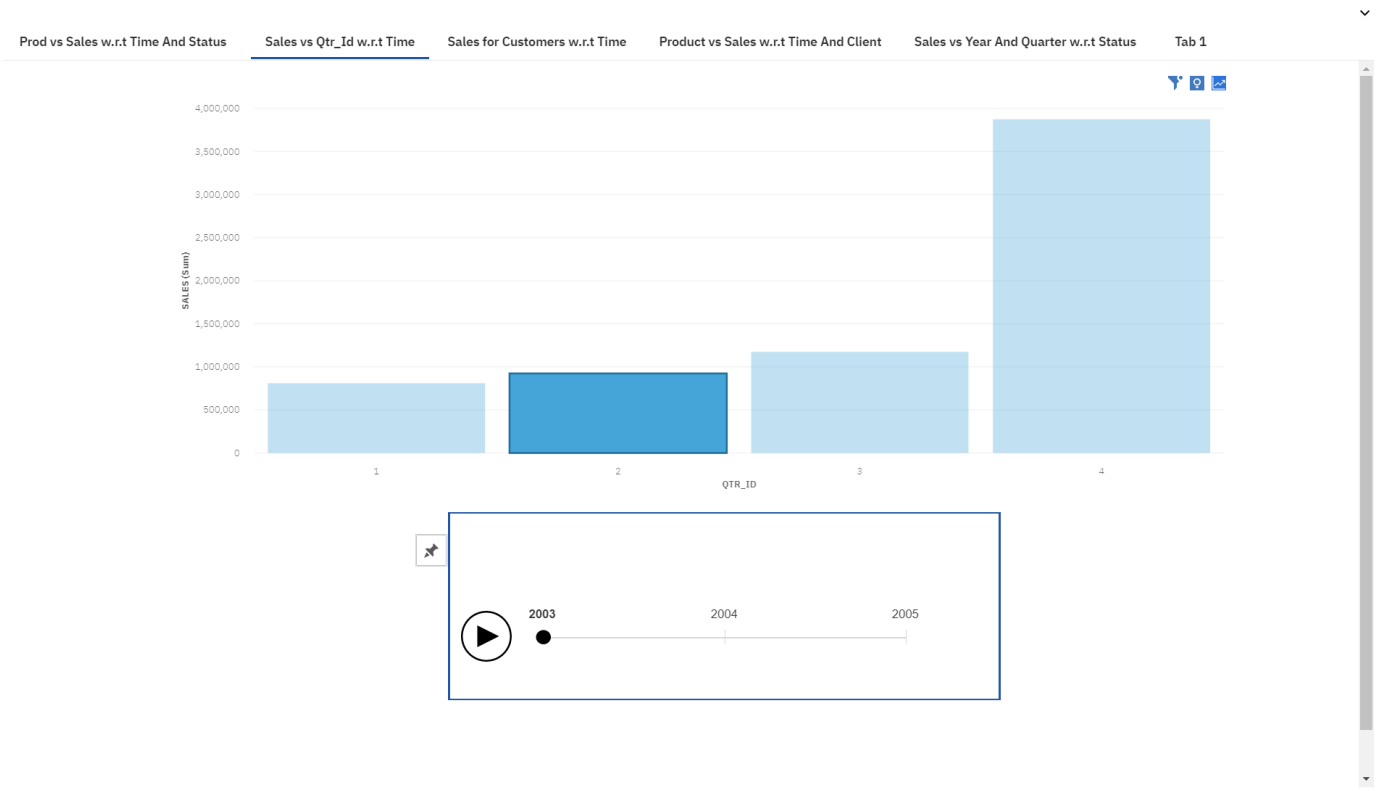


Figure 7: Sales vs Quarter Id with respect to year

Figure 8: Sales for Customers with respect to time(year)

Figure 9: Product Line vs Sales with respect to year and a particular company

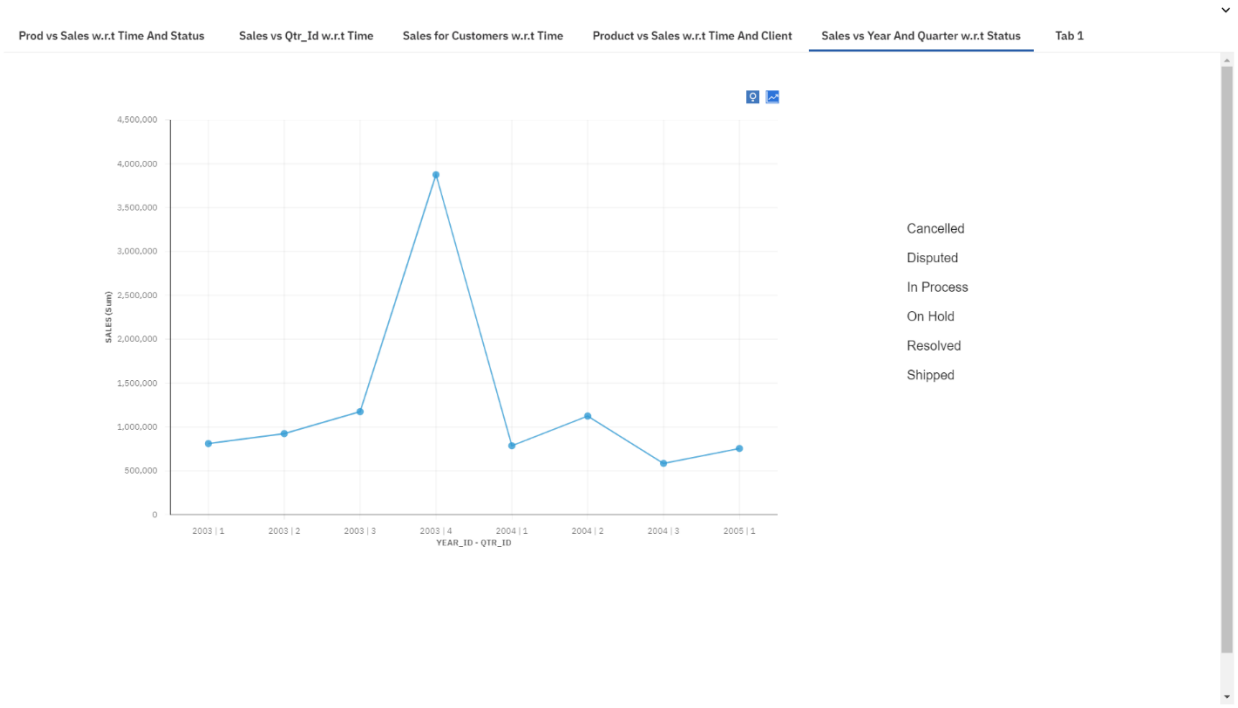


Figure 10: Sales vs Time(Year and Quarter id) with respect to Order Status

References

[1]"3 Ways to Build An ETL Process", *Panoply*, 2020. [Online]. Available: https://panoply.io/data-warehouse-guide/3-ways-to-build-an-etl-process/. [Accessed: 24- Mar- 2020].

[2]E. Kautzner and E. Kautzner, "What is a Data Warehouse and how does it deliver value? | Minerra", *Minerra*, 2020. [Online]. Available: https://www.minerra.net/business-analytics/what-data-warehouse-how-deliver-value/. [Accessed: 24- Mar- 2020].

[3]"Data Warehousing - Concepts - Tutorialspoint", *Tutorialspoint.com*, 2020. [Online]. Available: https://www.tutorialspoint.com/dwh/dwh_data_warehousing.htm. [Accessed: 25- Mar- 2020].

[4]"What is Business Intelligence? BI Definition", *OLAP.com*, 2020. [Online]. Available: https://olap.com/learn-bi-olap/olap-bi-definitions/business-intelligence/. [Accessed: 27- Mar- 2020].

[5]"IBM Cognos Analytics", *IBM Cognos Analytics*, 2020. [Online]. Available: https://www.ibm.com/ca-en/products/cognos-analytics. [Accessed: 27- Apr- 2020].

[6]"Data Warehousing Fundamentals", *Google Books*, 2020. [Online]. Available: https://books.google.ca/books?id=n2nIM0l1TQ0C&pg=PA91&lpg=PA91&dq=Managers+think+of+the+business+in+terms+of+business+dimensions&source=bl&ots=sUR-HRNUfT&sig=ACfU3U3vaSp-gm-wu18ljwZ2YW9TGffdlg&hl=en&sa=X&ved=2ahUKEwjS99jSk8foAhXTYDUKHejhBlgQ6AEwCXoECA4QLg#v=onepage&q=Managers%20think%20of%20the%20business%20in%20terms%20of%20business%20dimensions&f=false. [Accessed: 31- Mar- 2020].