CSCI-5408 Data Management, Warehousing, and Analytics

Winter 2020

Case Study Report

CSG 11

Guli, Tanu B00839890
Muni, Deep B00828375
Rokhjavan, Raouf B00826953

Contents

1 Summary of the Dataset 3
2 Business Intelligence – BI 4
2.1 What is the BI? 4
2.2 Development of BI Systems 4
2.2.1 What are the Steps 4
2.2.2 Business Understanding 5
2.2.3 Selection of data sources 5
2.2.4 Multidimensional Modeling and DWH Creation 5
2.2.5 ETL Process 6
2.2.6 Data Analysis 6
2.2.7 Visualization and Report 7
3 The Case Study – Sale Records 7
3.1 Introduction 7
3.2 Multidimensional Modeling and DWH Creation 7
3.3 ETL Process 9
3.4 Data Analysis and OLAP queries 9
4 Cognos Analytics 10
4.1 What is Cognos Analytics? 10
4.2 Data Analysis and Reporting in Cognos Analytic 11
5 References 13

Figure 1 Order details 3
Figure 2 Product details 3
Figure 3 Customer details 3
Figure 4 Deal details 3
Figure 5 Star Schema 5
Figure 6 Snowflake Schema 6
Figure 7 ETL Process in a BI System 6
Figure 8 Dimensions Tables of DWH 8
Figure 9 ERD based on Snowflakes Schema for DWH of Case Study 8
Figure 10 Snowflake schema in Cognos Analytics 11
Figure 11 Total Sales per country per product line for a quarter in a particular year 11
Figure 12 Order Status per Year, per ProductLine, and Customer 12

1 Summary of the Dataset

The data for this case study is taken from Kaggle [1]. The data set contains details about customers, orders, and products of a business. In total there are 2823 entries and 25 attributes in this data set.

Regarding the details of orders, there are 307 unique order numbers. For each order, a different type of product is bought. The details regarding each product such as quantity ordered, selling price and total sale (quantity order X selling price) for that product is maintained in the dataset. Apart from this, the order date and status of the order is also maintained. There are 6 types of status for an order, viz, Shipped, Resolved, Cancelled, On Hold, Disputed, and In Process.

ORDERNUMBER	QUANTITYORDERED	PRICEEACH	ORDERLINENUMBER	SALES	ORDERDATE	STATUS	QTR_ID	MONTH_ID	YEAR_ID
10100	30	100	3	5151	01-06-2003 00:00	Shipped	1	1	2003
10100	50	67.8	2	3390	01-06-2003 00:00	Shipped	1	1	2003
10100	22	86.51	4	1903.22	01-06-2003 00:00	Shipped	1	1	2003
10100	49	34.47	1	1689.03	01-06-2003 00:00	Shipped	1	1	2003

Figure 1 Order details

The business has maintained the records of the products. There are seven product lines, viz, Vintage Cars, Classic Cars, Trucks and Buses, Trains, Ships, Planes, and Motorcycles. These product lines comprise of 109 unique product codes. For each product code, the MSRP (Manufacturer's Suggested Retail Price) is also maintained.

PRODUCTLINE	MSRP	PRODUCTCODE
Vintage Cars	170	S18_1749
Vintage Cars	60	S18_2248
Vintage Cars	92	S18_4409
Vintage Cars	41	S24_3969

Figure 2 Product details

The customers who bought the above products are also maintained in this data set. There are 92 unique customers having 2823 orders in total. A customer's details such as address, phone number and contact name are stored along with their name.

CUSTOMERNAME	PHONE	ADDRESSLINE1	ADDRESSLINE2	CITY	STATE	POSTALCODE	COUNTRY	TERRITORY	CONTACTLASTNAME	CONTACTFIRSTNAME
Online Diecast Creations Co.	6035558647	2304 Long Airport Avenue		Nashua	NH	62005	USA	NA	Young	Valarie
Online Diecast Creations Co.	6035558647	2304 Long Airport Avenue		Nashua	NH	62005	USA	NA	Young	Valarie
Online Diecast Creations Co.	6035558647	2304 Long Airport Avenue		Nashua	NH	62005	USA	NA	Young	Valarie
Online Diecast Creations Co.	6035558647	2304 Long Airport Avenue		Nashua	NH	62005	USA	NA	Young	Valarie

Figure 3 Customer details

To understand the deal size of the orders/products bought, the final attribute "Deal size" stores this detail. The value for this attribute is small, medium and large.

The deal size is assigned accordingly to the sale of a product.

- Small => Sale is less than 3000
- Medium => Sale is between 3000 and 7000
- Large => Sale is greater than 7000

DEALSIZE
Small
Medium
Large

Figure 4 Deal details

2 Business Intelligence – BI

2.1 What is the BI?

Each organization has valuable sets of data in different forms which are gathered in different ways. Most organizations analyze their current or historical data to identify useful patterns and support business strategies. BI is a technology infrastructure for gaining maximum information from the available data to improve the business process. It provides believable information to help to make an effective business decision; in fact, the decision support system is part of BI, and it uses data and models to support management decisions. Typical BI infrastructure components are solutions for gathering, cleansing, integrating, analyzing, and sharing data. The most common kinds of BI systems are:

- Executive Information System - EIS
- Decision Support System - DSS
- Management Information Systems - MIS
- Geographical Information Systems - GIS
- Online Analytical Processing and Multidimensional Systems – OLAP
- Customer Relationship Management Systems - CRM

BI is based on Data Warehousing technology which gathers information from a range of a company's operational systems. Data loaded to the DWH is usually well-integrated and cleaned which allows to produce credible information in a way that is called "One version of the truth". In other words, it shows another aspect of the business which was not explicitly understandable on their own.

2.2 Development of BI Systems

2.2.1 What are the Steps

As mentioned earlier, a BI system is an infrastructure gathering and analysis of data. According to this notion, we need to follow some steps to set up a BI system.

1. Business Understanding
2. Selection of data sources
3. Multidimensional Modeling and DWH Creation
4. ETL Process
5. Data Analysis
6. Visualization and report

2.2.2 Business Understanding

BI starts with business understanding and answering questions like that How will I make the money? How will I sell the products? How will I manufacture products and deliver the service? Which region does my business plan to operate? Who are my customers? Answering these questions leads us to "Business Model" which describes different aspects of a business. The understanding of the business model leads to business strategies.

2.2.3 Selection of data sources

After finding business strategies, an organization has to find which information it has that fits into the analytical systems which helps it to make business decisions. The source of this data is heterogeneous; it means data can come from questionnaires, transactions, sensors, barcodes, websites which can be stored in many places like sales, stores, operation management, HR, and finance department.

2.2.4 Multidimensional Modeling and DWH Creation

Data Modeling includes designing data warehouse in details, and it follows principals and patterns established in data warehousing. Data warehouse is a data repository that makes operational and other data accessible in a form that is readily acceptable for decision support and other user applications. Data Warehousing refers to the process of design and analysis of a data warehouse to consolidate data from many sources in one large repository. The most common OLAP models are Star and Snowflakes. After finding a model for DWH, it's time to set it up. There are many different models and techniques for creating DWH.

There are two types of multidimensional schema: **Star Schema** and **Snowflake Schema**.

In star schema, the relationship is between the fact table and dimension tables. The tables in this schema are not normalized till a higher level.

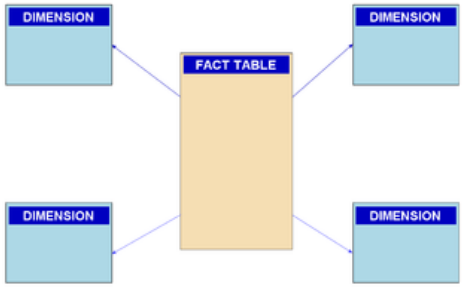

Figure 5 Star Schema

In snowflake schema, the relationship is not only between the fact table and dimension tables but it exists between a dimension table and another dimension table. The tables in this schema are normalized till a higher level.

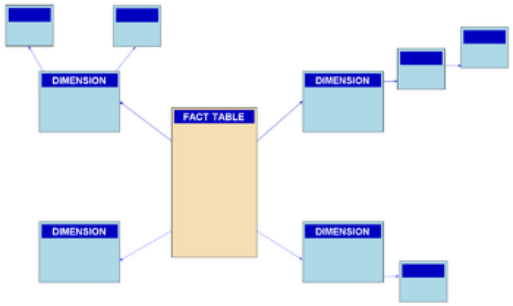

Figure 6 Snowflake Schema

2.2.5 ETL Process

This process stands for Extraction, Transformation, and Loading. The purpose of this process is data integration. The main problem during ETL is data quality because data comes from different data sources, so redundancy, inconsistency, missing data, etc. shall be addressed by ETL. Accordingly, data must be gathered from relevant sources, they should be cleaned and filtered, and they must be arranged into meaningful patterns using different tools. Some of the important concepts in this process are:

Semantic Integration which corresponds to eliminate mismatches between two data sources that share same semantic

Load, refresh, purge that is when loading data, periodically refresh and purge too old data.

Metadata management is when keeping track of source, loading time, and other information for all data in the warehouse.

Heterogeneous Source is focused on data that must be accessed from a variety of source formats and repositories. They may be heterogeneous in the data model, data management, data structure, data semantics, etc.

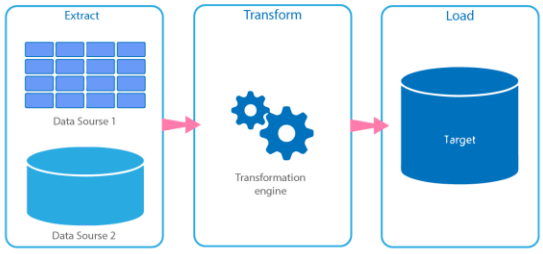

Figure 7 ETL Process in a BI System

2.2.6 Data Analysis

After loading data into the DWH, we need to analyze it and extract business decisions based on that. Data can be analyzed in many different ways; for example, it can be analyzed through OLAP complex SQL queries, or based on spread-sheet style operations, or based on a multidimensional view of data which is called descriptive analysis. A common operation is to aggregate a fact over one or more dimensions like finding total sales, finding total sales for each city/state/country, or finding the top five brands ranked by sales. However, there are other tools and techniques for data analysis like predictive analysis, decision analysis, waiting line analysis, network analysis, etc. Here in this report, we focus on

OLAP queries known as OLAP cubes. The type of OLAP queries are **Roll-up, Drill-down, Slice, Dice, and Pivot or Rotate** queries.

2.2.7 Visualization and Report

When data is analyzed and important factors were extracted, it is time for presenting them in an understandable form. In this way, high ranked people in an organization can perceive it, and make a business decision based on that. There are so many different ways to represent data like maps and GIS, multidimensional and 2D representation, tables and graphs, virtual reality, dashboards, storytelling, etc.

3 The Case Study – Sale Records

3.1 Introduction

In our case study, after considering the above-mentioned steps for developing a BI system, we planned to reflect all of those concepts in our work. The case study is based on the sales records of a company in different locations, time periods, for different customers. After understanding the business and its parameters based on the provided data set, we created a data model for the measure we wanted to analyze and create a DWH based on that. Then we extracted, transformed/cleaned, and loaded our data set to the DWH. For the next step, we analyzed the data in our DWH based on OLAP cube queries. The last step of the case study was the illustration of analyzed data by means of visualizations and reporting methods that were provided by Cognos Analytics.

3.2 Multidimensional Modeling and DWH Creation

As we mentioned in the summary of the dataset, we found that the provided dataset shows the sales result of a company. Accordingly, we choose the sales as a fact in the company that can be analyzed, and it can be used as a measure for business decisions. On the other hand, we discerned some fields in our dataset which were introducing 4 dimensions of business related to the sales fact. These dimensions are geographical dimension, customer dimension, product dimension, and time dimension.

After specifying the fact and dimensions in our dataset, we had to choose the type of schema for connecting the dimensions to the fact table. We decided to choose snowflake schema for this purpose for the following reasons:

- This model is normalized, so it will reduce the amount of redundancy and consequently possible inconsistency in our DWH.
- Normalization reduces the usage of storage in our DWH because we keep just one copy of data in a table, and refer to it with a foreign key.
- Snowflake model illustrates the relationship between facts and components of a dimension clearly; additionally, it prevents having bloated tables on each dimension which is common in a star schema.

The geographical dimension talks about the location of the sales including city, state, country, and territory. The customer dimension provides information about customers that sales related to. The product dimension gives us the information related to the product and its type for each sale. Finally,

the time dimension specifies the time of each sale including date, month, quarter, and year. The following table shows each dimension and its tables:

	Time Dimension	Customer Dimension	Product Dimension	Geographical Dimension
Tables	*Years*	*Customer*	*ProductLine*	*Territory*
	Quarters		*Product*	*Country*
	Months			*State*
	OrderDate			*City*

Figure 8 Dimensions Tables of DWH

To create the final multidimensional model based on snowflake schema for the DWH, we had to connect each dimension to the fact table. According to this notion, we connected time dimension using the OrderDateID of OrderDate table, customer dimension using the CustomerID of Customer table, product dimension using ProductID of Product table, and geographical dimension using CityID of City table as foreign keys to the sales table. We also created Orders table to extend the Sales fact table because we wanted to exclude those fields which were not directly related to the sales fact table like price of each product, quantity of each order, and status of each order. The final ERD for our data warehouse is as follows:

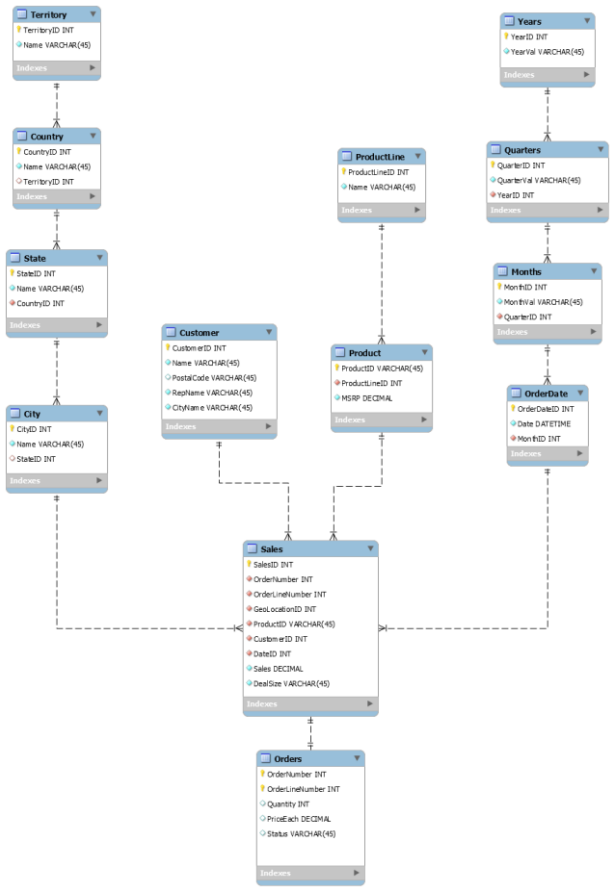

Figure 9 ERD based on Snowflakes Schema for DWH of Case Study

3.3 ETL Process

As we mentioned earlier, ETL stands for extraction, transformation, and load, and includes fetching data from heterogeneous sources of data, modifying and cleaning them based on our DWH fact and dimension tables, and finally load them into the DWH. In this case study, the extraction part was easy because the source of data was given to us as a web link on “kaggle.com”, so we did not need to fetch the data sets in different types and from different sources. To transform the dataset, we used regular expression in python. For each row in the data, we do the following steps:

- Removed any whitespace from the beginning and end of the column values
- Replaced empty strings or null values to N/A (Not Applicable)
- For TERRITORY column, replaced japan value with APAC.
- For CONTACTFIRSTNAME and CONTACTLASTNAME columns, concatenated tho string values with space in between.
- For ORDERDATE column, changed the date and time values to python’s datetime object like “%m/%d/%Y %H:%M”.
- Removed PHONE, ADDRESSLINE1, and ADDRESSLINE2 columns because we think these data are not useful in our business analysis, and they just add the size of data we stored in our DWH. We keep the CITY, and POSTALCODE for each CUTOMER to uniquely identify them in our DWH. To extract more detailed information about them, we can refer to the original data sources.

After Data cleaning and transforming data, we needed to load it to the DWH that we had created before. For this process, we developed a python script that read cleaned data set each row at a time and populated dimension tables and fact table in our MySQL database. To populate the DWH, we select one dimension and start from the table which is the leaf of the dimension. Because it does not depend on other tables, we can insert data in it directly, and get the ID of inserted data to connect it to the next table. As an example, for the geographical dimension and each row in our data set, first, we search for the TERRITORY filed in the Territory table, and if it does not exist, we will add it; otherwise, use its found ID for the Country table. For COUNTRY, STATE, and CITY columns, we do the same thing. We populate customer, product, and time dimensions in the same way. We keep CityID, CustomerID, ProductID, and OrderDateID in the Sales table as foreign keys. We also extend the Sales fact table with Orders table and put some data which are not necessarily related to the sales fact in that table like STATUS, QUANTITY, and PRICEEACH. We relate Orders and Sales table with OrderNumber and OrderLineNumber. In the end, we add SALES and DEALSIZE columns to the corresponding row on the Sales fact table. The final result would be similar to the snowflake model that we proposed in the previous section.

3.4 Data Analysis and OLAP queries

After loading the data to the DWH, we can issue different queries and extract related some analytical data based on them. As we mentioned in the previous parts, there are many different types of OLAP queries that can reveal valuable analysis of facts in a business. In our case study, we consider geography, customer, product, and time as different dimensions of the business and sales as a measure of business operation, so we can develop some OLAP queries as follows:

- **Roll-Up Queries & Drill-Down Queries** - Going from fine/coarse granularity and go to the other side of the dimension
    - Show the annual Sales per City, State, Country, and Territory. (Roll-up)
    - Show the annual sales per Product, Product Line. (Roll-up)

    - Show a specific Customer's sales for Year, Quarter, Month. (Drill-down)
- **Slice Queries** - Keep one dimension of the business fixed, Change the others
    - Select a specific ProductLine and based on that ProductLine show the total Sales for each Territory in each Year. we can drill down and roll up also.
    - Select a specific Year and based on that Year show the total Sales of each ProductLine for each Territory. We can also drill down and roll up in different dimensions.
    - Select a specific Territory/Country/Customer and based on that Territory/Country/Customer show the total sales in each year for each Product.
- **Dice Queries** - Keep two or more dimensions of the business fixed, Change others.
    - Select a specific ProductLine/Product, and a specific Territory/Country/State/City, and show the Sales based on the Year.
    - Select a specific Customer, and a specific Year/Qrt/Month, and show the Sales based on Product/ProductLine.
    - Select a specific Product, and a specific Year/Qrt/Month, and show the Sales based on 10 Customers who bought the most.
    - Select a specific Product/ProductLine, and a specific Customer, and show the Sales based on the Year/Qrt/Month
    - Select two specific Product/ProductLine, and show the sales based on each Territory/Country/City and Year/Qrt/Month
    - Select specific Product/ProductLine, specific Territory/Country/City, and specific Year/Qrt/Month, and show Sales based on Customers
- **Pivoting or Rotate Queries** - Change from 3D to 2D Visualization
    - For each Year, show Sales for each Year/City/Customer (dimension)
    - For each Country - Show Sales for each Year/Product/Customer (dimension)
    - For each Customer - show Sales for each Year/Product/ProductLine
    - For each Product- show Sales for each Year/Territory/Customer

In the above-mentioned queries, when we add "/" we means exclusive "or". It means we choose one of them.

4 Cognos Analytics

4.1 What is Cognos Analytics?

For reporting and visualization of the available dataset, we have used Cognos BI, which is a web-based business intelligence tool by IBM. It works on the IBM Watson artificial intelligence tool. Cognos provides dashboards and stories to visualize insights and analysis using chart, table, data player, and graphs. Also, the visualization act as an efficient tool that helps businesses in identifying market trends as well as identifying areas of the business to improvise. Cognos Analytics provides templates that contain predefined layouts and grid lines for easy arrangement and alignment of the visualization in a view.
Cognos provides feature to import external data and to create a data module. So, we have divided SalesData.csv into multiple csv files and then connected them to create a snowflake scheme in Cognos BI.

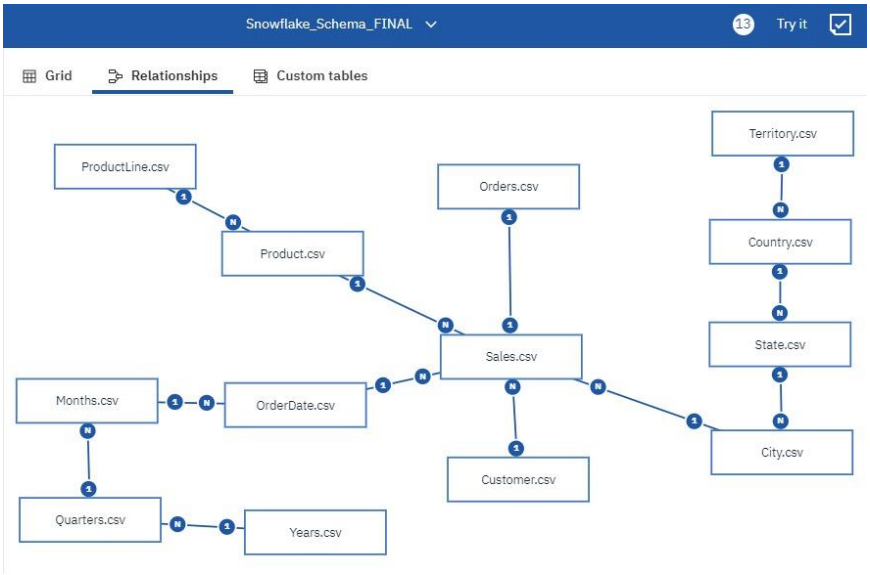

Figure 10 Snowflake schema in Cognos Analytics

4.2 Data Analysis and Reporting in Cognos Analytic

Dashboard 1: This dashboard displays the total sales, total order numbers for a particular quarter in a financial year. The year and quarter are represented using the data player. The pie chart displays the distribution of orders over product lines that helps business to identify the top-selling product lines for a specific period. Also, the bar graph denotes the countries where the sales were focused. Moreover, we can select a particular country and visualize the contribution of it in the total sales and which product lines were ordered. So, In the below representation, the company had a total of 1.11 million sales and 319 orders in the third quarter of the year 2004. Out of 319 orders, the maximum orders were placed for classic cars. The USA contributed the highest amount in the total sales whereas Austria was at the least.

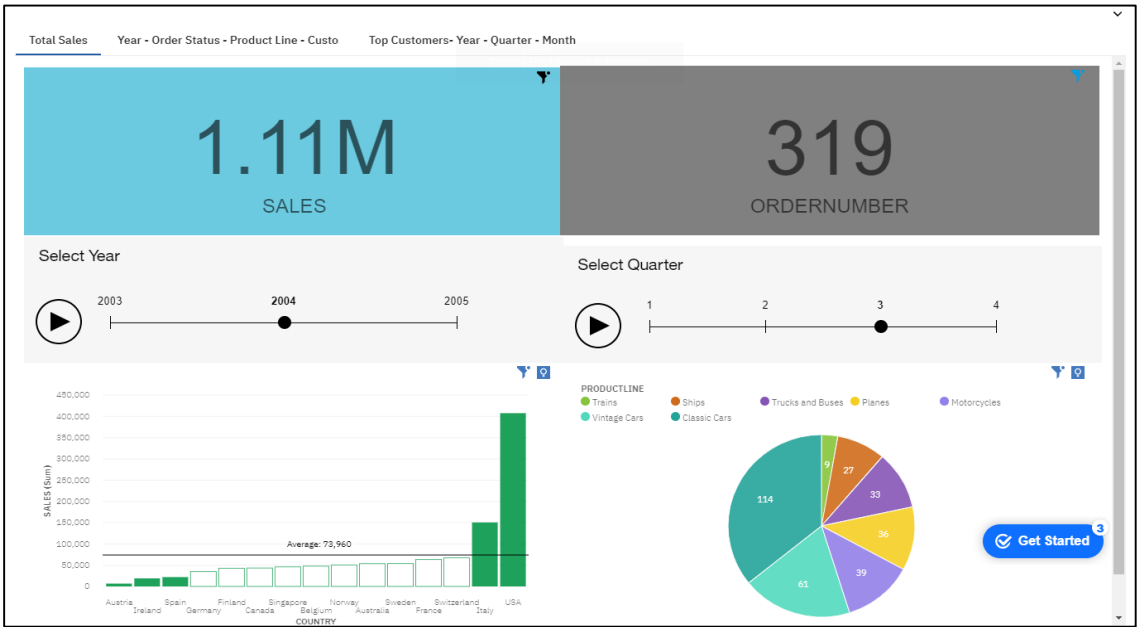

Figure 11 Total Sales per country per product line for a quarter in a particular year

Dashboard 2: The dashboard helps businesses to identify the total number of orders in a particular status for a financial year. Also, the pie chart denotes the product lines, which were in the selected state and the packed bubble on the left end corner of the screen represents the customer who placed that order and the country they belong to. So, In the below dashboard for the year 2005, there were five orders in the resolved status. These orders were placed by Toys4Grownups.com and EuroShoppingChannel customers, which belongs to the USA and Spain, respectively.

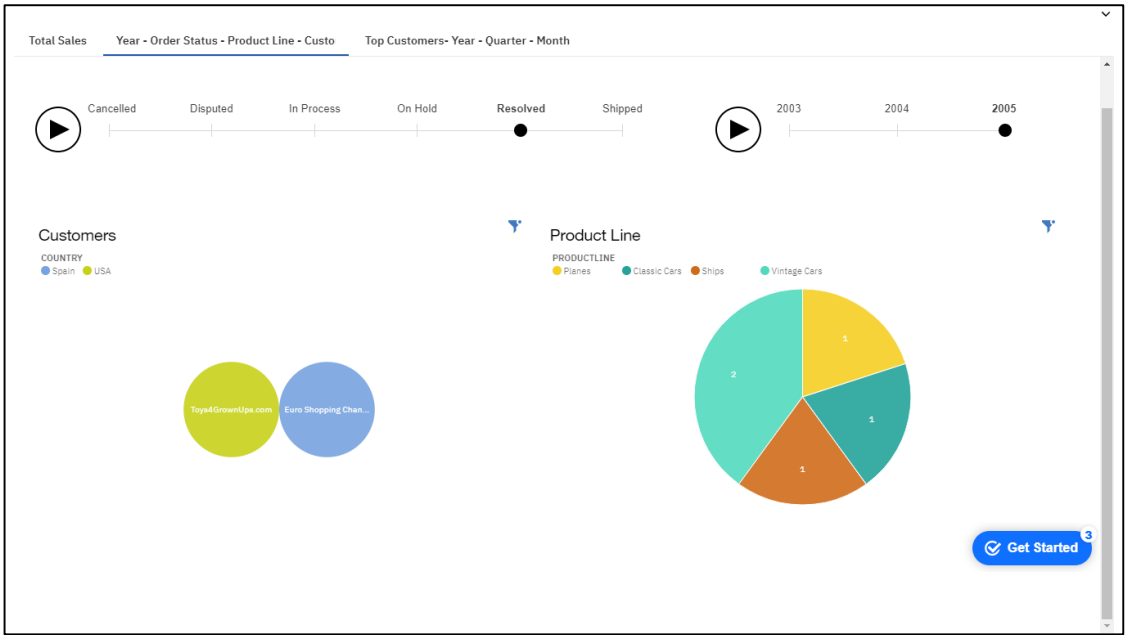

Figure 12 Order Status per Year, per ProductLine, and Customer

5 References

[1] "Sample Sales Data", Kaggle.com, 2020. [Online]. Available: https://www.kaggle.com/kyanyoga/sample-sales-data. [Accessed: 01- Mar- 2020]

[2] Cognos reference - “Cognos Analytics - Overview - Canada: IBM,” IBM Cognos Analytics. [Online]. Available: https://www.ibm.com/ca-en/products/cognos-analytics. [Accessed: 25-Mar-2020].

[3] IBM Knowledge Center. [Online]. Available: https://www.ibm.com/support/knowledgecenter/en/SSEP7J_11.1.0/com.ibm.swg.ba.cognos.cbi.doc/welcome.html. [Accessed: 25-Mar-2020].