A close up of a logo

Description automatically generated

CSCI 5408

Data Management, Warehousing, and Analytics

Case Study Report

April 3rd, 2020

Group 3

|  |  |
| --- | --- |
| **Name** | **Banner ID** |
| Aditya Patel  Abhinav Mandava  Naitik Prajapati | B00843467  B00841453  B00856836 |

Table of Contents

[Introduction 4](#_Toc36845200)

[Purpose 4](#_Toc36845201)

[Business Intelligence (BI) 4](#_Toc36845202)

[BI Process 4](#_Toc36845203)

[BI Tools [2] 5](#_Toc36845204)

[Source Data 5](#_Toc36845205)

[Extract, Transform, Load (ETL) 6](#_Toc36845206)

[Extract 6](#_Toc36845207)

[Transform 6](#_Toc36845208)

[Transformations we performed 6](#_Toc36845209)

[Transformation script 9](#_Toc36845210)

[Load 11](#_Toc36845211)

[IBM DB2 11](#_Toc36845212)

[Data Warehouse 11](#_Toc36845213)

[Schema 12](#_Toc36845214)

[Analytics, Visualization and Reporting 12](#_Toc36845215)

[Time Series Analysis 13](#_Toc36845216)

[Sales by Customers 14](#_Toc36845217)

[Orders and Order Status by Sale Deal Size 15](#_Toc36845218)

[Sales by Product Line 15](#_Toc36845219)

[Sales by Country 16](#_Toc36845220)

[Conclusion 16](#_Toc36845221)

[References 16](#_Toc36845222)

Table of Figures

[Figure 1: BI Process [1] 4](#_Toc36845233)

[Figure 2: Key Restructuring 6](#_Toc36845234)

[Figure 3: Cleaning 7](https://dalu-my.sharepoint.com/personal/ad723057_dal_ca/Documents/Case_Study_Report_CSCI5408.docx" \l "_Toc36845235)

[Figure 4: Cleaning and Format Revision 7](https://dalu-my.sharepoint.com/personal/ad723057_dal_ca/Documents/Case_Study_Report_CSCI5408.docx" \l "_Toc36845236)

[Figure 5: Cleaning and Format Revision 7](#_Toc36845237)

[Figure 6: Dates dimension 7](#_Toc36845238)

[Figure 7: Fact Table 7](#_Toc36845239)

[Figure 8: Products dimension 8](#_Toc36845240)

[Figure 9: Orders dimension 8](#_Toc36845241)

[Figure 10: Customers dimension 8](#_Toc36845242)

[Figure 11: Locations dimension 8](#_Toc36845243)

[Figure 12: Dimension Generation Script 9](#_Toc36845244)

[Figure 13: Fact Table Generation script 10](#_Toc36845245)

[Figure 14: OLTP Vs. OLAP 11](#_Toc36845246)

[Figure 15: Snowflake Schema [4] 12](#_Toc36845247)

[Figure 16: Monthly Sales 13](#_Toc36845248)

[Figure 17: Monthly Sales by Deal Size 13](#_Toc36845249)

[Figure 18: Sales by Customers 14](#_Toc36845250)

[Figure 19: Customers by Country 14](#_Toc36845251)

[Figure 20: Orders and Order Status by Customers, Deal Size of the Sales 15](#_Toc36845252)

[Figure 21: Sales by Product Line 15](#_Toc36845253)

[Figure 22: Sales by Country 16](#_Toc36845254)

# Introduction

Any business checks the profits, along with the profits they also look for: how sales happened, where they are getting more profits, where they are behind, and the demand for a particular product in a specific region for particular time span. Based on the analytics, they can decide how to improve the business from past learnings, or they can predict future outcomes. **Business Intelligence (BI)** helps to achieve these results.

BI helps business units, managers and other operational workers to make better business decisions backed up with accurate data. These decisions will incorporate new business opportunities, cost-cutting and identify inefficient processes that need a new process.

# Purpose

The purpose of the case study is for a better understanding of sample sales on automobiles to make appropriate business decisions based on various key performance index (KPI). We extracted sample sales data from the Kaggle. We performed transformations on the raw data to store in a data warehouse then executed business queries to derive more meaningful information to make business decisions. The detailed explanation is as follows:

# Business Intelligence (BI)

Business Intelligence (BI) is a series of technologies, architecture and processes that help in converting all the raw data into meaningful information that helps businesses in driving profit. It is a suite of services to change information into significant knowledge and intelligence. BI has a direct impact on the organization’s strategic, tactical and operational business decisions.

## BI Process

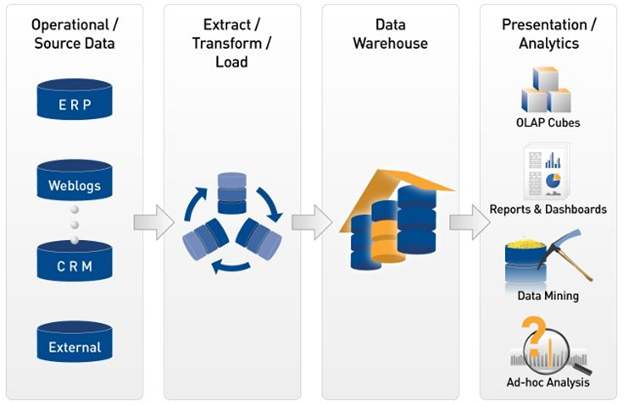


Figure 1: BI Process [1]

It starts with extracting data from various sources, which include daily databases such as ERP, Weblogs or CRM, or external files like CSV files. Also, we could have data spread across multiple heterogeneous systems or homogeneous systems. In our case, data is available on a single homogeneous system. After that, the transformation stage is taken into consideration, to make data consistent and easily understandable, because raw data consists of redundancy, duplication, which we need to remove before proceeding further. In this phase, we generated dimension and fact table. After completing this phase, data is loaded into the data warehouse to perform operations. Finally, data available in the data warehouse is taken and transform it into visualization using BI tools. Now, using the BI system, users can ask queries, requests, ad-hoc reports or conduct any other analysis.

## BI Tools [2]

Business intelligence (BI) tools are types of application software that collect and process large amounts of unstructured data from internal and external systems, including books, journals, documents, health records, images, files, email, video, and other business sources. It can also include data visualization software for designing charts, as well as tools for building BI dashboards and performance scorecards that display business metrics and KPIs to bring company data to life in easy-to-understand visuals. There are number of business intelligence tools available in market, following are some of them. There are number of business intelligence tools available in market, following are some of them.

* SAP BI
* MicroStrategy
* Datapine
* Oracle BI
* Tibco Spotfire
* Qlik Sense
* Tableau Desktop
* Microsoft Power BI
* IBM Cognos

## Source Data

The summary of the source data is as follows:

* Source: [link](https://www.kaggle.com/kyanyoga/sample-sales-data)
* Description: Sample Sales Data, Order Info, Sales, Customer, Shipping, etc., used for Segmentation, Customer Analytics, Clustering and more. Inspired for retail analytics. This was originally used for Pentaho DI Kettle, but the owner found the set could be useful for Sales Simulation training. Originally written by María Carina Roldán, Pentaho Community Member, BI consultant (Assert Solutions), Argentina.
* Version downloaded: Version 1
* Visibility: Public
* Downloadable file format: CSV
* Date created: 23 Nov. 2016
* Last modified: 23 Nov. 2016
* License: [CC0: Public Domain](https://dalu-my.sharepoint.com/personal/ad723057_dal_ca/Documents/•%09https:/creativecommons.org/publicdomain/zero/1.0)
* Owner: [Gus Segura](https://www.kaggle.com/kyanyoga/datasets)
* Size: 516 KB
* Rows: 2823
* Columns: 25
* Sparse[[1]](#footnote-2) columns: ADDRESSLINE2andSTATE
* Dirty[[2]](#footnote-3) columns: ORDERDATE and POSTALCODE

## Extract, Transform, Load (ETL)

### Extract

The extraction is the process of reading data from the source. In our case, extraction was just reading the CSV file. We used pandas.read\_csv(filename) of the Pandas library to extract data from CSV file.

### Transform

We used transformations [3] like:

* **Cleaning:** Mapping N/A and blank field to null or date format consistency, etc.
* **Deduplication:** Identifying and removing duplicate records
* **Format revision:** Character set conversion, unit of measurement conversion, date/time conversion, etc.
* **Filter:** Selecting only certain rows and/or columns – e.g. creation of dimensions
* **Key restructuring:** Establishing key relationships across tables or creating new relations if necessary

### Transformations we performed

* We dropped ADDRESSINE2 as it was too sparse to give any useful insights
* We also dropped TERRITORY because we felt no requirement of it.

A screenshot of a cell phone

Description automatically generated A picture containing drawing

Description automatically generated

Figure 2: Key Restructuring

dd/mm/yyyy

mm/dd/yyyy

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

Figure 3: Cleaning

Figure 4: Cleaning and Format Revision

A screenshot of a cell phone

Description automatically generated A screenshot of a cell phone

Description automatically generated

Figure 5: Cleaning and Format Revision

#### Filter

A screenshot of a cell phone

Description automatically generated

Figure 6: Dates dimension

A screenshot of a cell phone

Description automatically generated

Figure 7: Fact Table

A screenshot of a cell phone

Description automatically generated

Figure 8: Products dimension

A screenshot of a cell phone

Description automatically generated

Figure 9: Orders dimension

A screenshot of a cell phone

Description automatically generated

Figure 10: Customers dimension

A screenshot of a cell phone

Description automatically generated

Figure 11: Locations dimension

The dimensions and fact table we see above have been generated using Pandas library, which is an excellent library for data transformations (written in Python).

### Transformation script

#### Dimension generation script:

Figure 12: Dimension Generation Script

#### Fact table generation script:



Figure 13: Fact Table Generation script

### Load

Once we are done transforming the data, we have to load the data into a data warehouse.

We used IBM DB2 warehouse for this case study.

### IBM DB2

IBM DB2 is relational database which is also available on IBM Cloud. IBM also provides IBM DB2 Warehouse which internally uses IBM DB2 database but is enhanced with features like autoscaling, analytics toolset, and reporting. IBM DB2 Warehouse on cloud is a paid service and has no trial version available. However, IBM DB2 has a free trial on cloud and since we intend to use IBM Cognos as our BI tool, we chose DB2 so that the process is homogenous with respect to the vendor, i.e. IBM.

## Data Warehouse

* Data Warehouse is a major component in the process of BI for any company.
* DW is a subject-oriented, integrated, time-variant, and non-volatile collection of data which is in the form that makes analytics, reporting, and decision making convenient.
* DWs are central repositories of integrated data from one or more disparate sources. They store current and historical data in one single place that are used for creating analytical reports for workers throughout the enterprise.
* DW is structured, organized and optimized for analytical needs, retrieval efficiency thereby being insightful.

A screenshot of a cell phone

Description automatically generated

Figure 14: OLTP Vs. OLAP

### Schema

Initially we chose star schema as our DW schema but after incorporating suggestions by Dr. Dey, we changed it to snowflake. We chose snowflake because:

* They have little to no redundancy, so they are easier to maintain and change.
* It can simplify complex relationships (many: many)
* Can give fine granularity. For instance, the decomposing customers and introducing locations table can give fine grain relations.

A screenshot of a cell phone

Description automatically generated

Figure 15: Snowflake Schema [4]

## Analytics, Visualization and Reporting

Now that the data are cleaned, transformed and loaded into the data warehouse, we can use these data for analysis and gain useful insights which help in making strategic decisions. For performing analytics and reporting insights, using a robust tool comprising intuitive visualizations would make it easy for the users to understand how the business is faring, identify potential risks, formulate mitigation and contingency plans, and do whatever is necessary to keep the business in a good shape.

This case study uses IBM Cognos Analytics 11.1 for analyzing the data and creating visualizations to serve analytic queries. Cognos is a web-based business intelligence suite which includes tools for analytics, reporting, score carding, and monitoring metrics. In this case study, the following visualizations are created which answer the respective analytic queries.

### Time Series Analysis

* A line graph showing monthly sales and forecasted sales:

A screenshot of a cell phone

Description automatically generated

Figure 16: Monthly Sales

* A line graph showing monthly sales in each of the deal size categories:

A picture containing game

Description automatically generated

Figure 17: Monthly Sales by Deal Size

### Sales by Customers

The packed bubble in Figure 17 shows the amount of sales generated by each of the customers.

The word cloud shows the amount of sales generated by each customer and the size of the deal with that customer.

A screenshot of a cell phone

Description automatically generated

Figure 18: Sales by Customers

The map visualization below shows the spread and density of customer base across different countries.

A screenshot of a cell phone

Description automatically generated

Figure 19: Customers by Country

### Orders and Order Status by Sale Deal Size

The bar graphs in visualization in figure-18 show the order statuses and number of orders in a particular status along with the deal size. The packed bubble shows the order status by customer

A screenshot of a social media post

Description automatically generated

Figure 20: Orders and Order Status by Customers, Deal Size of the Sales

### Sales by Product Line

The following visualization shows the sales in each product line, deal sizes, total sales and total products sold

A screenshot of a social media post

Description automatically generated

Figure 21: Sales by Product Line

### Sales by Country

The following visualization shows the sales by country.

A screenshot of a cell phone

Description automatically generated

Figure 22: Sales by Country

# Conclusion

To conclude, the case study helped us learn the importance of BI and how it can be useful to make meaningful business decisions. We got ourselves familiarized with the process of ETL, basic data warehouse design, analytics and creation of useful visualizations. Making intelligent analysis from a simple CSV file requires a lot of analytical skills and we believe, this case study helped us hone that skill.

# References

|  |  |
| --- | --- |
| [1] | J. K. a. B. Dietz, "Business Intelligence: A Primer," Open Source Business Resource, Sept. 2009. [Online]. Available: https://timreview.ca/article/284. [Accessed April 2020]. |
| [2] | E. G. Haije, "Top 15 Business Intelligence Tools: An Overview," Mopinion, 6 Nov. 2019. [Online]. Available: https://mopinion.com/business-intelligence-bi-tools-overview/. [Accessed April 2020]. |
| [3] | E. T. —. E. Database, "ETL Transform — ETL Database," [Online]. Available: www.stitchdata.com/etldatabase/etl-transform/. [Accessed April 2020]. |
| [4] | "dbdiagram.io - Database Relationship Diagrams Design Tool," 2020. [Online]. Available: dbdiagram.io. [Accessed April 2020]. |

1. Sparse means the column had many blank fields. [↑](#footnote-ref-2)
2. Dirty means the column had inconsistent format and unnecessary information [↑](#footnote-ref-3)