SSIS - SQL Server Integration Services

Visual Studio

Tobacco Company Revenue development

By,
Sunil Elangovan

Tables of Content

Introduction

Data warehouse architecture

Schema structure for data warehouse

The case studies to be covered:

Data Sources:

Technologies Used:

Implementation and design process:

Designing of data warehouse:

Extraction transformation loading (ETL):

Back room: preparing the data

Case Studies

- **Case study 1**
- **Case study 2**
- **Case study 3**

Conclusion

List of References

Appendix A: Screenshot of code

Introduction:

Data generation is occurring at a rapid rate. According to (Villars, Olofson, & Eastwood, 2011) **in 2010, the world generated over 1ZB of data and reached 7ZB of data by 2014.All of this data creates new opportunities if used wisely. Now business decisions are not taken on the basis of experience anymore they are taken by analyzing the data they have therefore it becomes essential for companies to maintain data warehouses and keep collecting data for future use. This also introduces the concept of data-lake which means the NO-SQL database for the easy collection of data.**

Data warehouse and data management are the foundations of Business intelligence. It is the process to provide relevant and reliable information to the right people at the right time.

This project implements data warehousing concepts of gathering data from different sources and also displays its business intelligence capabilities to help improve the tobacco business in America. The report covers following steps.

1) **Collecting the datasets from different sources**
2) **Designing the data warehouse**
3) **Designing the Extraction-Transformation-Load steps**
4) **Development of ETL**
5) **Loading data into data warehouse**

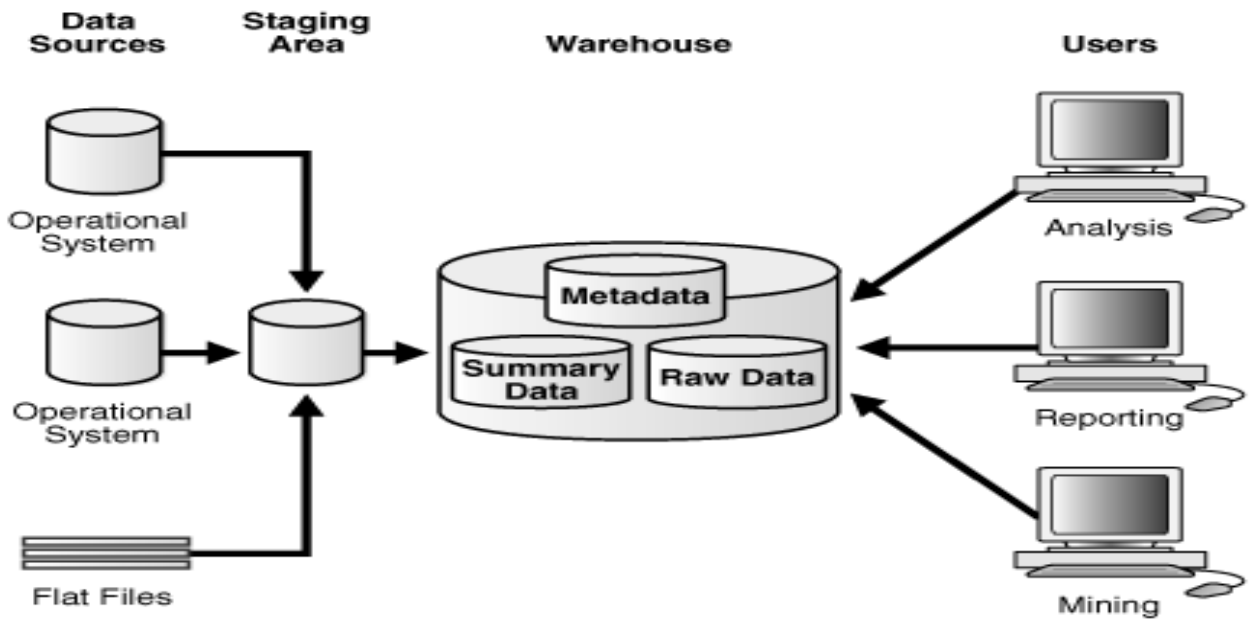

Figure 1: Data warehouse model

In the operational systems we can store all day to day transactions similarly in flat files all the unstructured data is stored all is raw form of data which has all the redundancies and anomalies which needs to be removed this is done through the ETL process in which the data is mines and transformed into structured format which is further dumped into data warehouse. this data now can be used for reporting and analysis and to run business queries to find out relevant information.

Online analytical processing (OLAP) data is typically stored in Star Schema or in snowflake schema in a relational data warehouse or a special purpose data management system. Dimensions are derived from the dimensions tables and measures are derived from the fact tables. The elements of the dimensions can further be organized into hierarchy's and create parent child relationships between different elements of dimensions.

Data Warehouse Architecture:

For this project I have used star schema which is the Ralph Kimball's Approach also known as bottom-up approach because it is the simplest approach and it is apt according to my idea and study.

According to (Ralph & Margy, 2011) schema consists of one or more fact tables and dimension tables around the facts. The fact table generally consists of all the foreign keys in the schema i.e. primary key of all the dimensions and the data which can be measured and is quantitative in nature for example in my case revenue, sales, sample value etc. these are measure which means facts of fact tables.

Dimensional tables have comparatively less data than the data in fact tables. Each dimension table has one primary key which is the foreign key for the fact table. In my case, the primariary key of dimension's are revenue_key(the revenue that was generated), gender_key(who generated the revenue), sales_key, product_key(the product that has been sold), state_key(which state has maximum sales), education_key, pattern_key(pattern of people smoking), year_key(the date when the product was sold), sentiment_key(tells the sentiment score as per twitter data).

In star schema there is no relationship between different dimensions and in my study I don't have any relationship between my dimensions also normalization is not required in kimball's approach that's why I used star schema inspite of snowflake schema.

Schema structure for Data warehouse:

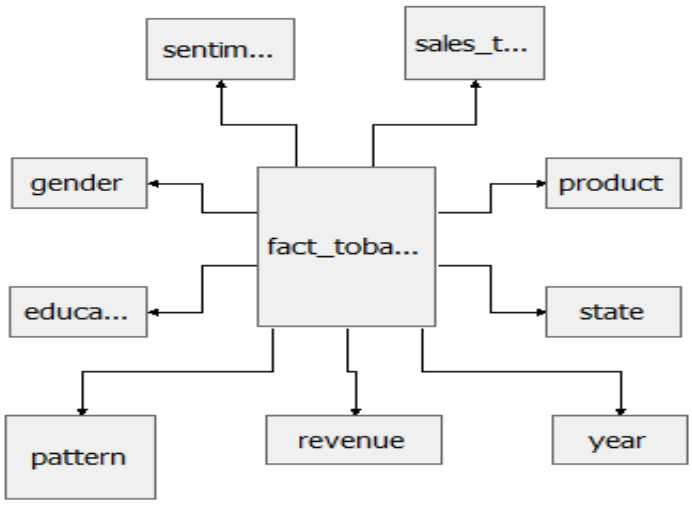

Figure 2 : schema structure for data warehouse

The case studies to be covered:

a) **Revenue from different products state wise.**

b) **Comparing Sales value and survey sample value with respect to the level of education**

c) **Comparing sentiment Score with the consumption pattern**

Data Sources:

To build a business focused data model we need to choose good sources of data which has least errors and data provided is in proper formats. To build data warehouse for tobacco product consumption I have used:

Data Source 1: heathdata.gov - https://www.healthdata.gov/dataset/samhsa-synar-reports-youth-tobacco-sales-0 (Structured CSV file).

Data Source 2: centers for disease control and prevention - https://www.cdc.gov/tobacco/data_statistics/surveys/nyts/ (Structured CSV format).

Data Source 3: kaggle.com - https://www.kaggle.com/anjalichappidi/tobacco-consumption (structured CSV File).

Data Source 4: cdc.gov - https://www.cdc.gov/mmwr/preview/mmwrhtml/mm6120a3.htm (Scraped data from HTML format to CSV format – Semi-structured data)

Data Source 5: For this I have extracted twitter data to perform sentiment analysis product wise and to know which product has maximum positive rating by people.

Technologies Used:

Programming Languages:

a) **R for extraction of data through twitter API and Rvest package to scrape data from web**

b) **Sql for creating facts and dimensions**

Database management tools

a) **Sql server management studio and SSIS for creating facts and dimensions.**

Additional add-ons/tools

a) **SPSS for Statistical analysis for statistically analyzing data.**
b) **Sql Server Analysis services is the tool to deploy the cube and browse the dimensions and facts.**
c) **Microsoft power BI tool is used for better visualization of the case studies.**
d) **Datum box api**
e) **Word cloud tool**
f) **Open refine is used to refine the data.**

Implementation and design process:

Here we have used the bottom up approach. In this process data marts are created first for all the reporting and analysis for business queries then they are added to the data warehouse. kimball's approach has four main processes as follows:

a) **Select the business process: Here the business process is to study revenue values, comparing sales value with the level of education and sentiment score with consumption pattern.**
b) **Declare the grain: After selecting the business process next step is to describe the grain of the project. The grain is the perfect model of the dimensions on which the model is based on. In my project the grain is the revenue.**
c) **Identify the dimensions: this is the most integral part of data warehousing as these are the foundation of making fact tables. The dimensions are the place where all the data is stored.**
d) **Identify the fact table: after identifying the dimensions the next step is to identify the keys and measurements that needs to be included in the fact table**

Designing of data warehouse:

In my project first I created database named "Tobacco" in Mysql Server Studio Management and also created all the dimension tables and fact tables that were required for the project.

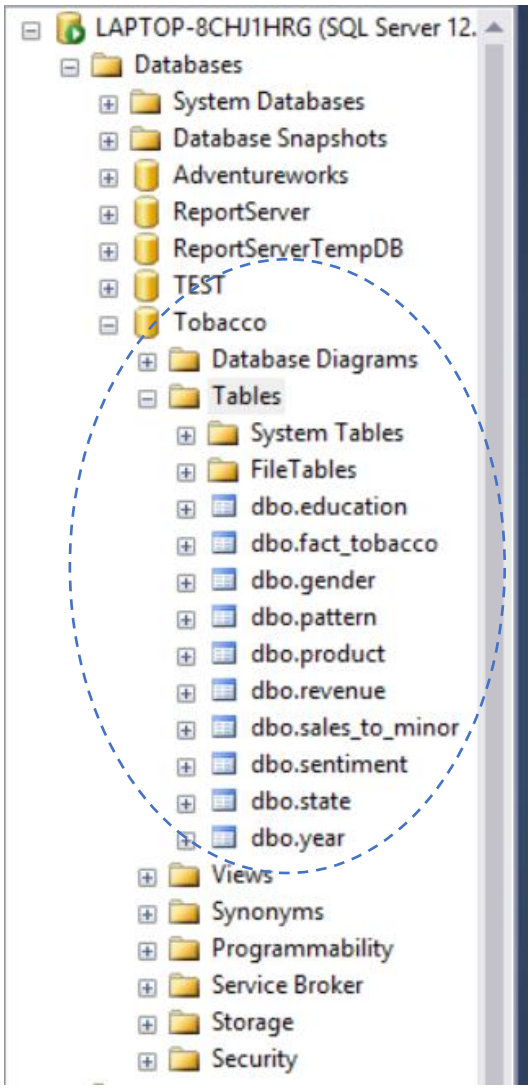

Figure 3 : tables in Mysql

Next step is to create the processes on Sql Server integrated services tool and create the workflow to populate the data in the database tables.

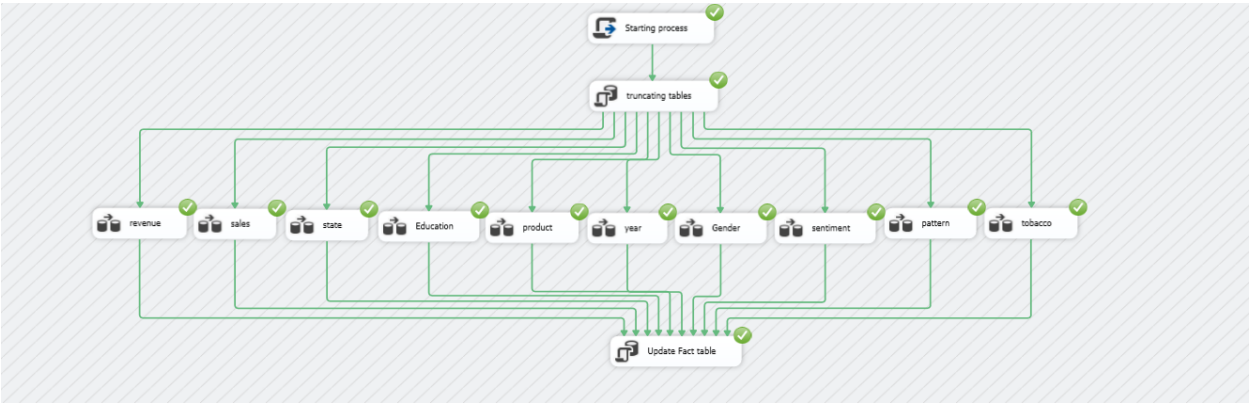

Figure 4 : SSIS workflow for project

This shows the flow of data in the data warehouse in SSIS. As we can see first it runs a C# script which says "about to start the process", then is the truncate script through which all the tables are truncated. In the third step all the dimension tables are populated with the data. The green ticks in the process shows that the process is successful and the dimension tables are populated successfully with all the data and in the last step the data is fetched into the fact table and it gets updated too.

Results | Messages

	year_key	state_key	revenue_key	sales_key	pattern_key	sales_to_minor	revenue(in_million_$)	smoke_everyday	smoke_someday	former_smoker	never_smoked	Sample_v
1	8	1	1	766	164	21	68.3	0.194	0.052	0.539	0.215	6
2	8	2	2	767	210	29	28.4	0.198	0.063	0.484	0.254	15
3	8	3	3	768	196	12	166.1	0.198	0.021	0.609	0.173	16
4	8	4	4	769	165	22	83.3	0.222	0.037	0.499	0.242	11
5	8	5	5	770	184	21	612.1	0.136	0.056	0.546	0.262	9
6	8	6	6	771	200	18	59.6	0.171	0.057	0.529	0.243	1
7	8	7	7	772	192	58	120.6	0.153	0.056	0.509	0.282	0
8	8	51	8	773	211	33	17.5	0.151	0.066	0.613	0.17	0
9	8	9	9	774	172	7	1000	0.18	0.04	0.519	0.261	28
10	8	10	10	775	169	20	85.1	0.197	0.039	0.545	0.219	8
11	8	11	11	776	177	23	32.4	0.158	0.037	0.544	0.261	24
12	8	12	12	777	204	12	25	0.158	0.045	0.546	0.252	11
13	8	13	13	778	174	26	457.2	0.17	0.061	0.519	0.25	25

Query executed successfully. | LAPTOP-8CHJ1HRG (12.0 SP1) | LAPTOP-8CHJ1HRG\Parija... | Tobacco | 00:00:01 | 660 rows

Figure 5 : Fact Table

After populating data into datawarehouse, we can go to the sql server analysis services in which we first have to upload the data sources, data source views, make all the dimensions and make the cube. Also we can create relations between the attributes and hierarchies of the elements of the dimensions.

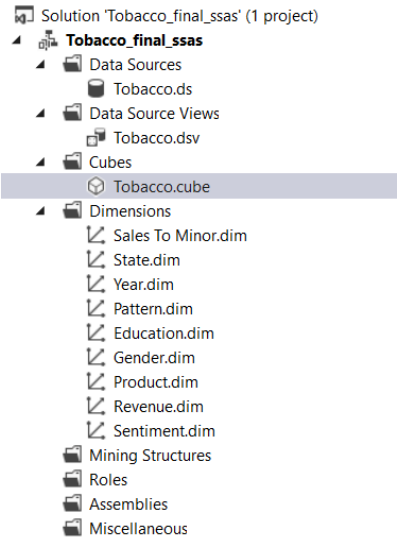

Figure 6 : setting up SSAS

The below image shows the structure of the cube.

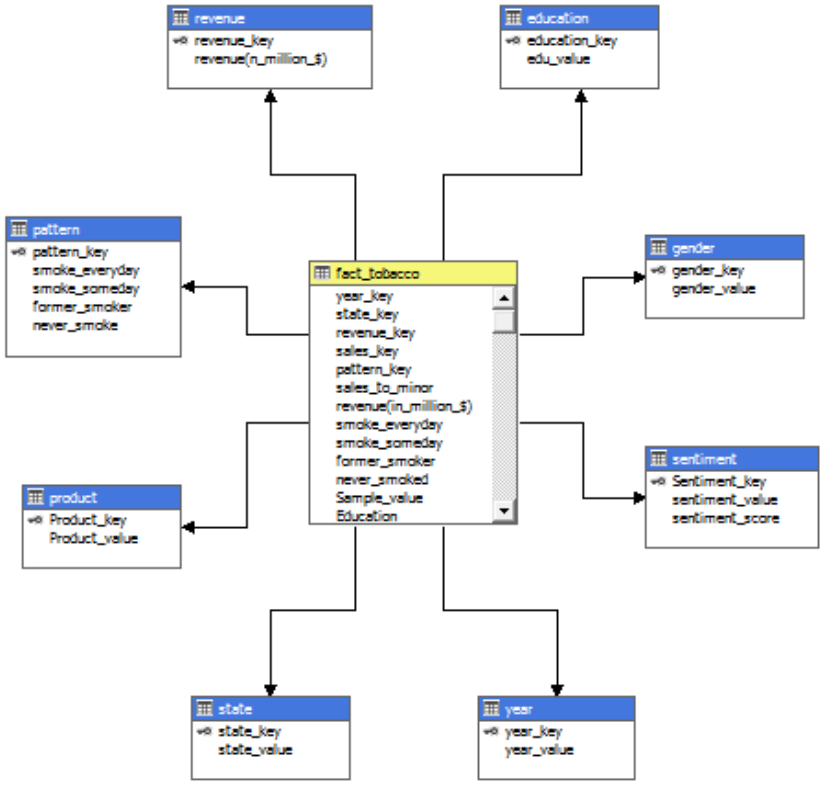

Figure 7 : star schema for project

Next step is to deploy the cube.

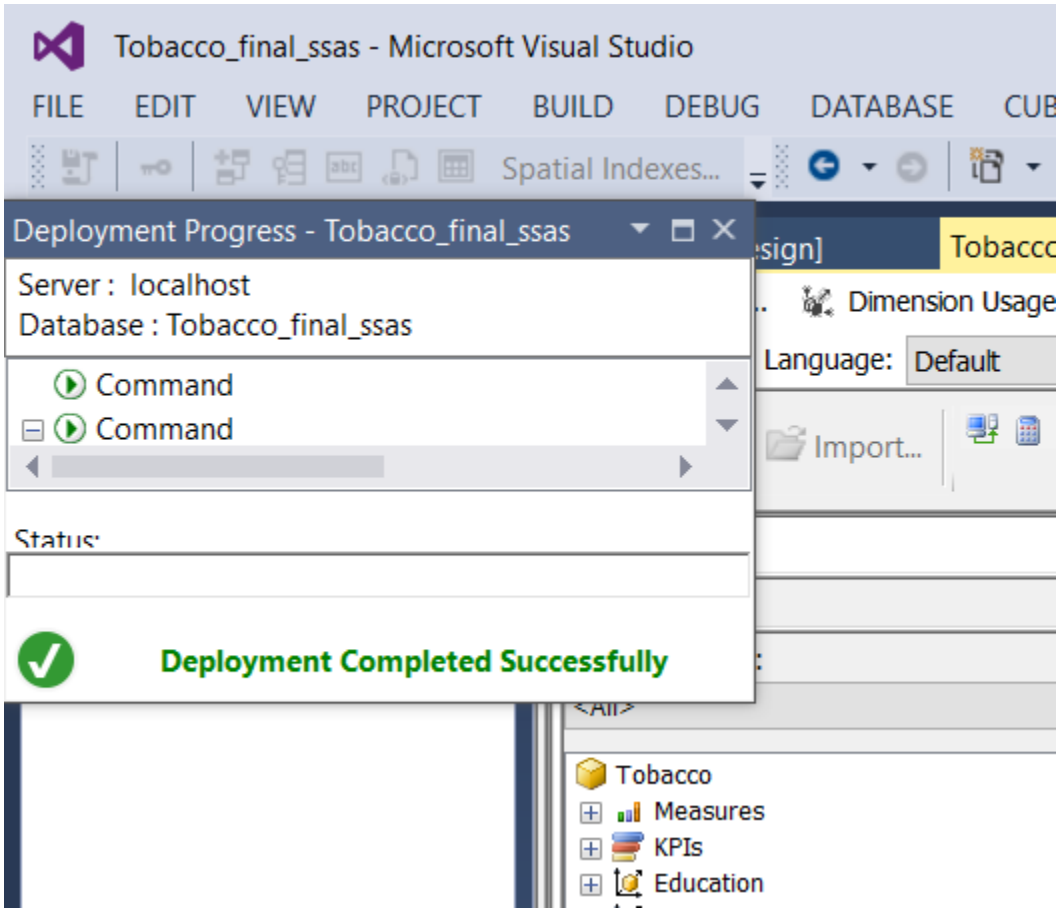

Figure 8 : deploying the cube

This shows that the cube has been deployed successfully .Now we can browse the cube and perform the business queries.

Extraction transformation loading (ETL):

Figure 9 : ETL Process

According to (Vassiliadis, Simitsis, & Skiadopoulos, 2002) **during this period data warehousing design is concerned with two tasks which are practically executed in parallel i.e. collection of requirements and analysis of the structure and content of the existing data sources and their intentional mapping to the common data warehouse model.**

The first step of the ETL is the extraction or collection of data. This is the most challenging part as real world data is full of redundancies, noise and is dirty. The data can be extracted in the excel, csv, json etc formats. The next step is transformation of data in which the raw data is converted into format which can be used. The last step is to load data into data warehouses where it can be used for analyzing business queries.

Back room: preparing the data

Extraction: in my project I have used all three formats of data i.e. structured, semi-structured and unstructured. For my structured datasets I downloaded the csv files from healthdata.gov, kaggle.com and cdc.gov. for the unstructured data I web scraped the revenue of states from tobacco products from the cdc.gov site using rvest package in CRAN repository (Wickhan, 2015). For the unstructured data I fetched tweets as per products and performed sentiment analysis on them to know which product is famous among people.

Cleaning: For cleaning structured and semi structured sources I used open refine and Microsoft excels and for unstructured data I used R programming. In R programming I used tm package to clean the tweets.

Transformation: the sentiment of tweets was in the form of strings which needs to be transformed to the numeric values for further analysis for this I used R programming.

Delivering - In this the dimension tables and the fact tables are loaded through SQL Server Integration Services and the cube is also deployed.

Case Studies

Case Study 1: Revenue from different products state wise

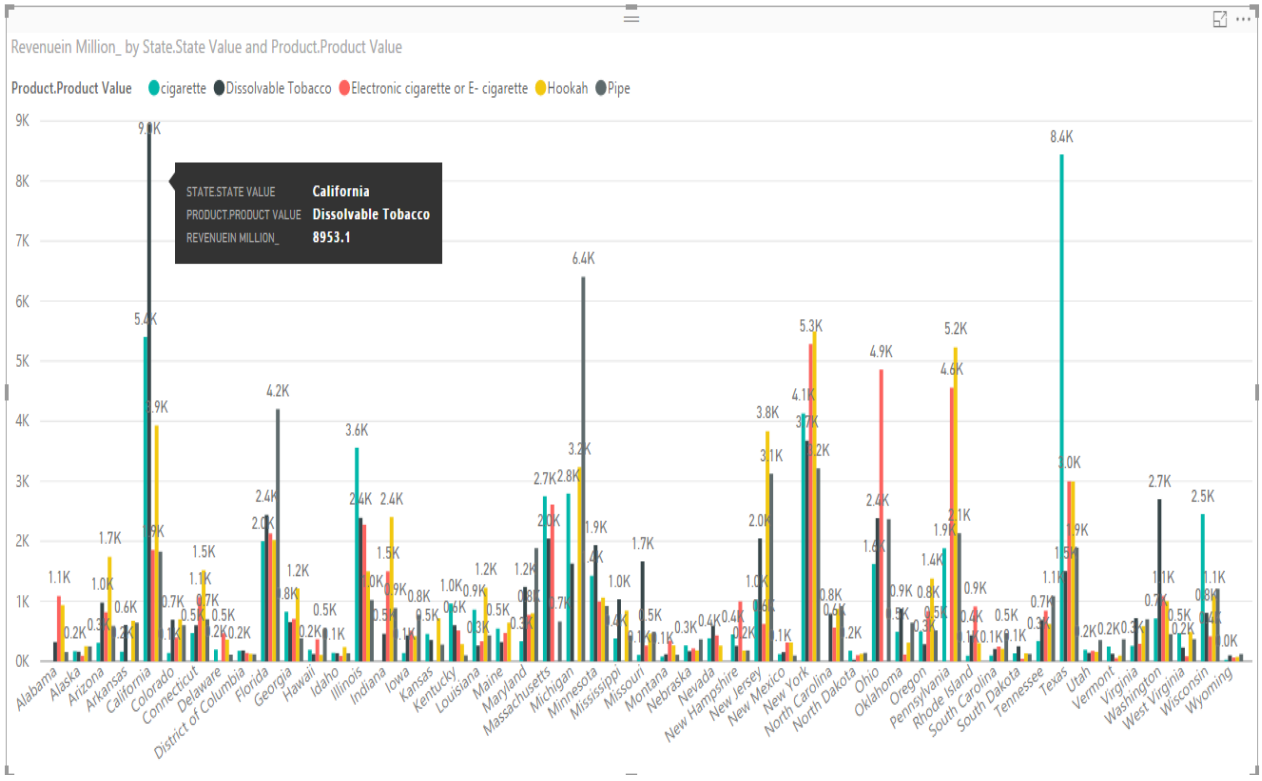

Figure 10 : case study 1

This case study shows the maximum revenue generated by a product from each country. This can help the company to decide which product they want to sell in which country to earn big revenue's. This graph shows that California has maximum revenue i.e. 8953.1 million dollars on Dissolvable tobacco.

Case study 2: Comparing Sales value and survey sample value with respect to the level of education

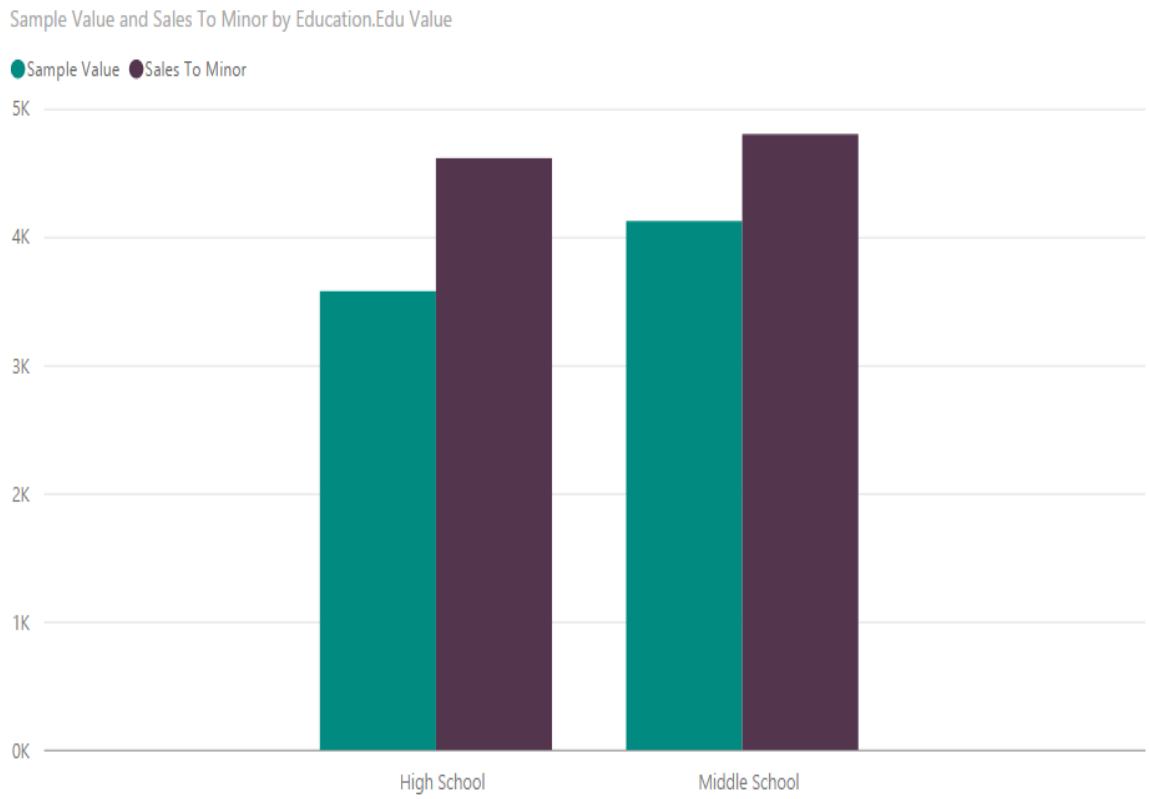

Figure 11 : case study 2

In this case study we are comparing sales values and survey sample values of tobacco product consumption with respect to the level of education. In this we can see that the consumption of tobacco is more in middle school as per both sales values and the sample survey values. Therefore we can say that middle school children are more prone to tobacco products use and can serve as a great market.

Case Study 3: Comparing sentiment Score with the consumption pattern

Figure 12 : case study 3

For this case study we have used the twitter sentiment score for each product and compared the total score with the smoke everyday group to contrast relationship between them. And it can be clearly seen that hookah has maximum positive points and this can also be seen in smoke everyday group that highest percentage of people like to have hookah every day.

Statistics:

Descriptive Statistics

	Mean	Std. Deviation	N
Revenue(million$)	369.033	433.5279	660
Survey_value	12.156	13.0856	660

Correlations

		Revenue (million$)	Survey_value
Revenue(million$)	Pearson Correlation	1	.041
	Sig. (2-tailed)		.296
	Sum of Squares and Cross-products	123856710.3	152208.608
	Covariance	187946.450	230.969
	N	660	660
Survey_value	Pearson Correlation	.041	1
	Sig. (2-tailed)	.296	
	Sum of Squares and Cross-products	152208.608	112842.647
	Covariance	230.969	171.233
	N	660	660

Figure 13 : statistical analysis

To test the outcomes of the project, I conducted a statistical test to get the correlation between revenues and survey_value. The revenues is in millions and the survey_value is the value of percentage of youth having tobacco. I used Pearson's correlation to generate a relation between revenue and the percentage of youth having tobacco. After conducting the survey it is concluded that there is indeed a positive correlation between revenue and survey_value.

The Pearson correlation score for survey_value is .041, which is not that significant overall but it does indicate that more revenue means more consumption of tobacco products. This test was conducted to see if the data is correct after the deployment of the cube.

Hence we can say that, revenue is directly related to more sale percentage of consumption of tobacco.

Conclusion:

Through this data warehouse project I wanted to tell how tobacco companies can increase their revenue state wise using my first case study, through second case study it is clear that middle school children are easy targets for tobacco companies to sell their products and the last case study shows that as per twitter sentiment score and smoke everyday percentage hookah is most liked by people.

Appendix A for screenshots of code:

```
install.packages("plyr")
install.packages("rvest")
install.packages("tm")
install.packages("xml2")
library(rvest)
library(xml2)
library(plyr)
library(stringr)
require("tm")
require("xml2")


html <- read_html("https://www.cdc.gov/mmwr/preview/mmwrhtml/mm6120a3.htm")
judges <- html_table(html_nodes(html, "#table-4"))
View(judges)
write.csv(judges, file ="revenue.csv" )
```

```
install.packages("twitteR")
install.packages("RCurl")
install.packages("RJSONIO")
install.packages("stringr")
install.packages("tm")
install.packages("wordcloud")


library(twitteR)
library(RCurl)
library(RJSONIO)
library(stringr)
library(tm)
library(wordcloud)


#verfiyng the twitter API
my_key='JFOa8FyAcbun3eokNoUMh1zA6'
my_secret = 'qMLG4w6nHYmQ7RRDl5XYfncbQUJgjWq8IPh1VcgklRgPWiWsM1'
atoken = '2782378207-VZiXtycPsz2rKN2R19ST8RqiFMs1uJR62ZJsA90'
asecret = 'iQKTSNpSlPGCqgZg8G46mKxSNmtXmatPvwO1edisUJuxO'

options(httr_oauth_cache=T)

setup_twitter_oauth(my_key, my_secret,atoken,asecret)
getSentiment <- function (text, key){
  text <- URLencode(text);
  #save all the spaces, then get rid of the weird characters that break the API, then convert back the URL-encoded spaces.
  text <- str_replace_all(text, "%20", " ");
  text <- str_replace_all(text, "%\\d\\d", "");
  text <- str_replace_all(text, " ", "%20");

  if (str_length(text) > 360){
    text <- substr(text, 0, 359);
  }
  ##############################################
  data <- getURL(paste("http://api.datumbox.com/1.0/TwitterSentimentAnalysis.json?api_key=", key, "&text=",text, sep=""))

  js <- fromJSON(data, asText=TRUE);

  sentiment = js$output$result


  return(list(sentiment=sentiment))

}
#The clean.text() function

clean.text <- function(some_txt)
{
  some_txt = gsub("(RT|via)((?:\\b\\W*@\\w+)+)", "", some_txt)
```


```
  return(list(sentiment=sentiment))

}
#The clean.text() function

clean.text <- function(some_txt)
{
  some_txt = gsub("(RT|via)((?:\\b\\W*@\\w+)+)", "", some_txt)
  some_txt = gsub("@\\w+", "", some_txt)
  some_txt = gsub("[[:punct:]]", "", some_txt)
  some_txt = gsub("[[:digit:]]", "", some_txt)
  some_txt = gsub("http\\w+", "", some_txt)
  some_txt = gsub("[ \t]{2,}", "", some_txt)
  some_txt = gsub("^\\s+|\\s+$", "", some_txt)

  # define "tolower error handling" function
  try.tolower = function(x)
  {
    y = NA
    try_error = tryCatch(tolower(x), error=function(e) e)
    if (!inherits(try_error, "error"))
      y = tolower(x)
    return(y)
  }

  some_txt = sapply(some_txt, try.tolower)
  some_txt = some_txt[some_txt != ""]
  names(some_txt) = NULL
  return(some_txt)
}

# harvest tweets
tweet_txt <- searchTwitter("@cigarette",n=100, lang="en")


tweet_txt = sapply(tweets, function(x) x$getText())
tweet_clean = clean.text(tweet_txt)
tweet_num = length(tweet_clean)
tweet_df = data.frame(text=tweet_clean, sentiment=rep("", tweet_num), stringsAsFactors=FALSE)
sentiment = rep(0, tweet_num)
for (i in 1:tweet_num)
{
  tmp = getSentiment(tweet_clean[i], "b34bf514405eee1f46e7b752dec3b108")

  tweet_df$sentiment[i] = tmp$sentiment

  print(paste(i," of ", tweet_num))

}
```

References

Bibliography

Vassiliadis, P., Simitsis, A., & Skiadopoulos, S. (2002). Conceptual Modeling for ETL Processes.

Villars, R. L., Olofson, C. W., & Eastwood, M. (2011, June). Big Data : What It Is and Why You Should Care. *IDC* . framingham: AMD.

Kimball, R. and Ross, M., 2011. *The data warehouse toolkit: the complete guide to dimensional modeling*. John Wiley & Sons.