Data Warehousing and Business Intelligence Project

on

Mobile Phone Sales

Manish Chauhan
18141889

MSc/PGDip Data Analytics – 2019

Submitted to: Sean Heeney

National College of Ireland
Project Submission Sheet – 2019/2020
School of Computing

Student Name:	Manish Chauhan
Student ID:	18141889
Programme:	MSc Data Analytics
Year:	2019
Module:	Data Warehousing and Business Intelligence
Lecturer:	Sean Heeney
Submission Due Date:	12/04/2019
Project Title:	Mobile Phone Sales

I hereby certify that the information contained in this (my submission) is information pertaining to my own individual work that I conducted for this project. All information other than my own contribution is fully and appropriately referenced and listed in the relevant bibliography section. I assert that I have not referred to any work(s) other than those listed. I also include my TurnItIn report with this submission.

ALL materials used must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is an act of plagiarism and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	January 21, 2020

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Table 1: Mark sheet – do not edit

Criteria	Mark Awarded	Comment(s)
Objectives	of 5	
Related Work	of 10	
Data	of 25	
ETL	of 20	
Application	of 30	
Video	of 10	
Presentation	of 10	
Total	of 100	

Project Check List

This section capture the core requirements that the project entails represented as a check list for convenience.

- ☒ Used LaTeX template
- ☒ Three Business Requirements listed in introduction
- ☒ At least one structured data source
- ☒ At least one unstructured data source
- ☒ At least three sources of data
- ☒ Described all sources of data
- ☒ All sources of data are less than one year old, i.e. released after 17/09/2017
- ☒ Inserted and discussed star schema
- ☒ Completed logical data map
- ☒ Discussed the high level ETL strategy
- ☒ Provided 3 BI queries
- ☒ Detailed the sources of data used in each query
- ☒ Discussed the implications of results in each query
- ☒ Reviewed at least 5-10 appropriate papers on topic of your DWBI project

Mobile Phone Sales

Manish Chauhan
18141889

January 21, 2020

Abstract

Data warehouse provides deep insights for decision making, which has increasingly become essential for every industry. Companies now take their decisions based on Business intelligence rather than on the data. Data warehouse makes it easy for them by storing the historical data and providing visual representation of it.On this basis ,this research contains 3 companies' historical data about mobile phones from 2010 to 2018 to observe their sales and revenue patterns by creating a Data warehouse. With the use of quarterly sales and revenue data, it shows how much revenue Apple, Samsung and Nokia generate from their Mobile phone sales. Moreover, it also shows the partial reason behind Nokia's market share loss and why Apple's revenue is very high compared to that of Samsung and Nokia. This data warehouse focuses on comparing the sales and revenue of Apple, Samsung, and Nokia mobiles, how their annual performance had been for the last 9 years and which model affected the sales most.

1 Introduction

Mobile phones or now more commonly called smartphones have become part of our day to day life and are used extensively by everyone.Almost everyone who has access to basic amenities has mobile phones.They are sold every day and the revenue generated is enormous.Moreover, as technology is becoming increasingly smaller, Mobile phones have started to replace laptops and computers.They have already become the hottest technology in the market.That is why I choose mobile phones sales as the dataset for my Data warehousing project.Apple Inc., Samsung, and Nokia are some of the big players in the mobile industry.

A data warehouse is an expensive, enterprise-wide endeavour with significant organizational impacts.Data warehousing creates changes that resonate throughout the entire organization, and it demands broad-based and lasting support.It requires the sponsorship and support of senior management, managers in the business units and IT.There must be a substantial initial and ongoing commitment of financial and human resources.This commitment must be made while recognizing that the greatest benefits from data warehousing usually occur later rather than immediately.Together, all three organizational factors were found to be significant in the research model and together they provide organizations with effective mechanisms for increasing widespread support for warehousing, addressing politics, and ensuring that the necessary resources are provided(Wixom & Watson (2001)).

The data warehouse (DW) has been considered as a vital technology for modern decision support systems (DSS) for organizations. Indeed, the DW offers efficient capabilities for supporting decision-makers.Despite the valuable efforts and attempts from researchers devoted to developing DW approaches, several DW projects failed. In fact, researchers agree that any successful attempt to develop a DW should consider two features namely i) the construction of the Data warehouse multidimensional schema by using a hybrid approach relying on user requirements and data sources; and ii) the use of semantic resource to overcome the heterogeneity issues complicating the DW construction process (Elamin et al. (2018)).

Data warehousing is the coordinated, architected, and periodic copying of data from various sources, both inside and outside the enterprise, into an environment optimized for analytical and informational processing (Hammergren & Simon (2009)).

The Data warehouse created shows the quarterly unit sales of mobile phones and the revenue generated from their sales from the year 2010 to 2018 for Apple Inc., Samsung, and Nokia.My main motive of choosing the project on this topic was to show how the sales and revenue for the major mobile phone manufacturers have changed during the course of the time.

The project captures three requirements that are further explained by using non-trivial BI queries in Tableau.

(**Req- 1**) The first requirement shows the quarterly sales and revenue comparison of Apple, Samsung, and Nokia.Though the unit sales might appear similar for all three brands, the revenue generated is different because of the major difference in the selling price of the mobiles.

(**Req- 2**) The second requirement shows the annual performance of Apple,Samsung and Nokia and in which year the companys' performance is affected the most.

(**Req- 3**) The last requirement shows the best performing model among Apple, Samsung, and Nokia mobile phones.This has been based on the number of units sold for a model.

2 Data Sources

Data collection is done from multiple sources and is briefly mentioned below.

2.1 Source 1: Statista

The Statista mobile phones dataset is downloaded from the below sources.This dataset shows the quarterly unit sales data for Apple, Samsung and Nokia phones from 2010 till 2018.

For Nokia phones, the dataset for unit sales is available from 2012. The missing dataset for the year 2010 and 2011 has been imputed based on linear regression.

- Nokia mobile phone units sold from Q1 2012 to Q3 2017. Publication date is April 2018.

`https://www.statista.com/statistics/299180/mobile-phone-shipments-worldwide-by-nokia/`

Source	Type	Brief Summary
Statista	Structured	Mandatory source for the project
Apple Investors Website	Unstructured	Quarterly revenue of smartphones
Nokia Investors Website	Unstructured	Quarterly revenue of smartphones

Table 2: Summary of sources of data used in the project

- iPhone units sold from Q1 2007 to Q4 2018. Publication date is Nov 2018.

`https://www.statista.com/statistics/263401/global-apple-iphone-sales-since-3rd-quarter-2007/`

- Samsung mobile phone units sold from Q1 2010 to Q4 2018. Publication date is Jan 2019.

`https://www.statista.com/statistics/299144/samsung-smartphone-shipments-worldwide/`

- Samsung mobile quarterly revenue from Q1 2011 to Q4 2018. Publication date is Jan 2019.This sheet has revenue data according to the business segment.Mobile phone data is extracted from it using Python. Linear Regression has been used to impute the missing revenue values in SPSS.

`https://www.statista.com/statistics/630434/samsung-quarterly-revenue-by-segment/`
All these above datasets are less than 1 year old and fulfill the criteria for data selection.

2.2 Source 2: Apple Investors website

The quarterly revenue data for the iPhone sales is scraped using the Data Miner chrome extension.To scrap the data, we have to select the required data and then right-click to select Get Similar(Data Miner) option.This creates the recipe in the Data Miner and then it can be downloaded as .CSV or .XLSX file.

The dataset on the official Apple website was scraped from the 10-K Annual report.It contains quarterly revenue for different Apple products.

Below link contains the drop down to choose the Annual 10-K forms from where the data has been scraped. To see the form, choose the 'Annual' option in SEC Groupings list and the year in year dropdown.

`https://investor.apple.com/investor-relations/sec-filings/default.aspx`

The 'Notes Detail' section in the Annual report contains the 'Segment Information and Geographic data - Net Sales by Product(Details)' from where the data has been scraped.

`https://d18rn0p25nwr6d.cloudfront.net/CIK-0000320193/71ac2994-85af-426b-982a-8fcc71d6fe52.html#`

`https://d18rn0p25nwr6d.cloudfront.net/CIK-0000320193/bc9269c5-539b-4a69-9054-abe7849c4242.html#`

2.3 Source 3: Nokia Investors website

Nokia quarterly revenue data from 2010 to 2018 was scraped from Nokia Investors website using Data Miner chrome extension. The data present there is in Euros. So, the data is converted into Dollars according to the conversion rates for that quarter using Python.

The fact that majority of the Nokia's revenue is generated by the mobile phone sales, that is why this has been considered a source of data for Nokia mobile phone revenue.

https://www.nokia.com/about-us/investors/

https://www.nokia.com/sites/default/files/files/nokia_results_2018_q4.pdf

https://www.nokia.com/sites/default/files/files/nokia_results_2018_q3.pdf

After collecting the data from these sources, some of the data has been manually taken from Wikipedia and processed via Python code to put it into data files.

For Example - The data about Mobile phone models and operating software used in them was taken from wikipedia and processed against the quarter or year in which they were launched. It was also assumed that the revenue generated in the year or quarter was generated by the mobile phone launched in that year or quarter.

3 Related Work

Its not the first time that data about Mobile Phones have been collected and analyzed. This has been done many times but for different purposes. This section talks about the previous researches and articles written in the field of Mobile Phone sales.

3.1 Factors affecting buying behaviour of Mobile Phones

There has been four different researches by different people on the factors and motivations behind the consumer buying behaviour of Mobile phone sales.First one,dated April 2005, focuses on the factors affecting Consumer choice of Mobile Phones in Finland in two different studies(Karjaluoto et al. (2005)). The second one,dated October 2013, focuses on the factors affecting consumer buying behaviour of Mobile Phone Devices in Hawassa town(Shanka (2013)).The third one, dated June 2018, focuses on the Motivations behind the purchasing decisions of Smart Phones by collecting the data using sampling methods covering various socioeconomic and cultural groups(Hossain et al. (2018)). The last one,dated January 2019, focuses on the factors influencing purchase intention of Cellular phones among the University students in Bangladesh(Rakib (2019)).

All these research are limited to a particular geographic location and so are very narrow in scope. My work is different from them in the sense that they focuses only on the consumers but not on manufacturers and the revenue.

3.2 Impact of Price on Customer Satisfaction in Telecom sector

This research shows the impact of price on customer satisfaction in the Telecom sector, mainly Mobile Phones(Hussain (2019)). It shows the direct positive relationship of price with customer satisfaction using statistical approach. We have used the same statistical approach of Linear Regression to predict the missing values in our research but our research focuses on the mobile phone units sold and revenue generated by sale.

3.3 Apple and Samsung sales

Some people have before collected the mobile phone sales data for Apple and Samsung. But either they have analyzed both the companys' data separately or worked on an understanding of the sales pattern of that company. None have done the comparison of the sales and revenue patterns of Mobile phones before.

For Example Samsung sales data by Abhijeet Pratap dated 2 December 2018 shows the Samsung sales data by Business division, namely Consumer Electronics, IT and Mobile communications and Device solutions, etc.

- https://notesmatic.com/2018/12/samsung-revenue-by-geographical-segment/

Apple products mainly iPhone, iPad and Mac sales data from 2005 till 2018 to show how many units of iPhones, iPads, and Macbooks are sold by Apple for that time period.

- https://notesmatic.com/sales-of-apple-products-by-year/

My work is entirely unique in that I have compared the sales and revenues of smartphones of Apple, Samsung, and Nokia from the year 2010 to 2018. This will add to the people's understanding of how the 3 big competitors in the smartphone market are performing.

4 Data Model

The data model here consists of one fact table and multiple dimension tables. In a multidimensional model, numeric values are the objects of analysis.Every numeric measure depends on a set of dimensions, which provide the context for the measure.

The dimensions together are assumed to uniquely determine the measure.Each dimension is described by a set of attributes (Chen (2001)).

4.1 Normalizing a Data Model

The purpose of creating a normalized data model is to ensure that the data will not be duplicated within the database and to ensure data consistency in its simplest form. This means designing a data model in which the design is flexible, meaning that the data components can be associated in different contexts without redundancy(Laberge (2011)).

Mobile phone sales data warehouse is in 3NF since it does not contain duplicate data and eliminates the dependencies on non-key fields by putting them in separate tables.

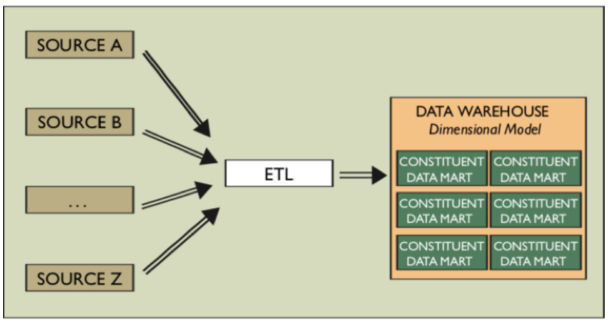


Figure 1: Data Warehouse Design(Jukic (2006)).

4.2 Fact Table

The fact table contains the measures.It preserves the statistical data related to the subject.A fact table contains the foreign key for Ids to establish a relationship with Dimension tables.A fact table can have multiple foreign keys.In the star schema, the fact table is surrounded by dimension tables.

This Data warehouse consists of only 1 fact table, namely FactSales.

4.2.1 FactSales

This table stores the quarterly sales and revenue data for the three brands of smartphones, which are part of the project.It contains Sales_Id, Source_RowID, Company_ID, Product_Id, Software_Id, Quarter_Id, Year_Id, Unit_Sales_Million and Revenue_Million_Dollar as columns.

Sales_Id is the Primary key here.Company_ID, Product_Id, Software_Id, Quarter_Id and Year_Id columns have foreign keys in them to connect them with Dimension table primary keys.Foreign keys are essential to maintain referential integrity.

FactSales is a transactional fact table since the measures are at their finest level of granularity and the transaction level is the Sales_Id. The measures in fact table are Additive(Aggregated) and can be aggregated by all dimensions of the fact table.

Aggregation requires defining a proper operator to compose the measure values characterizing primary fact instances into measure values characterizing each secondary fact instance.

Definition 4. Given a fact scheme f, measure mj∈M is said to be aggregable on dimension dk∈Dim(f) if ∃(mj, dk, Ω)∈S, non-aggregable otherwise. Measure mj is said to be additive on dk if ∃(mj, dk, 'SUM')∈S, non-additive otherwise.

As a guideline, most measures in a fact scheme should be additive. An example of additive measure in the sale scheme is qty sold: the quantity sold for a given sales manager is the sum of the quantities sold for all the stores managed by that sales manager (Golfarelli et al. (1998)).

4.3 Dimension Table

Dimension tables are integral companions to a fact table.The dimension tables contain the textual context associated with a business process measurement event.They describe the who, what, where, when, how, and why associated with the event (Kimball & Ross (2013)).

This Data warehouse contains 5 dimension tables, namely DimCompany, DimProduct, DimSoftware, DimQuarter, and DimYear.

4.3.1 DimCompany

DimCompany store the company information. This dimension table contains Company_Id, Company_Name, Headquarter, and Level as dimensions. Company_Id is autogenerated and primary key.

4.3.2 DimProduct

DimProduct stores the mobile phone models information. This dimension table contains Product_Id, Model_Number, Product_Type, Manufacturer, DESCRIPTION, and Product_Name as dimensions. Product_Id is autogenerated and the primary key here.

4.3.3 DimSoftware

DimSoftware table stores the Operating System information originally used in the mobile at the time of launch. This dimension table contains Software_Id, OSVersion_Number, OS_Name, and Designed_By as dimensions. Software_Id is autogenerated and primary key.

4.3.4 DimQuarter

DimQuarter table was created to classify the data based upon quarters. This dimension table contains Quarter_Id, Quarter and Quarter_Name as dimensions. Quarter_Id is auto-generated and primary key.

4.3.5 DimYear

DimYear table was created to classify the data based upon Year. This dimension table contains Year_Id and Year as dimensions.Year_Id is auto-generated and primary key.

4.4 Star Schema

The star schema architecture is the simplest data warehouse schema. It is called star schema because it resembles a star, with points radiating from the center.

Star schema represents the multidimensional data model. The database consists of a single fact table and a single table for each dimension. Each tuple in the fact table consists of a pointer (foreign key - often uses a generated key for efficiency) to each of the dimensions that provide its multidimensional coordinates, and stores the numeric measures for those coordinates. Each dimension table consists of columns that correspond to the attributes of the dimension(Dayal & Chaudhuri (1997)).

The star schema in this Data warehousing project looks like one below.

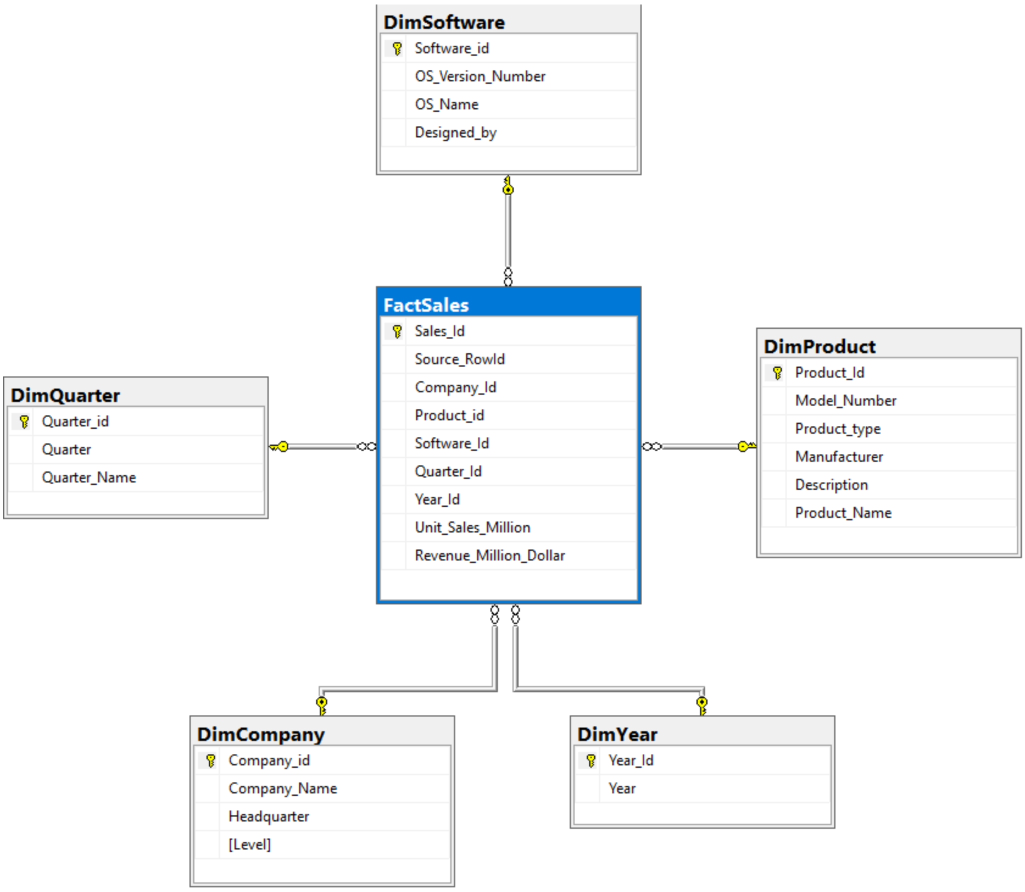


Figure 2: Star Schema for Mobile Sales Data Warehouse

5 Logical Data Map

The physical implementation can be a catastrophe if it is not carefully architecture before it is implemented. Just as with any other form of construction, you must have a blueprint before you hit the last nail. Before you begin developing a single ETL process, make sure you have the appropriate documentation, so the process complies logically and physically with your established ETL policies, procedures, and standards. The logical data map describes the relationship between the extreme starting points and the extreme ending points of ETL system (Kimball & Caserta (2004)).

Table 3: Logical Data Map describing all transformations, sources and destinations for all components of the data model illustrated in Table below

Column Name	Destination	Data Type	Type	Transformation
Company_Id	DimCompany	Int	Dimension	Auto generated value
Company_Name	DimCompany	Varchar(50)	Dimension	Populated based on mobile model
Headquarter	DimCompany	Varchar(50)	Dimension	Populated based on mobile Model
Level	DimCompany	Varchar(20)	Dimension	Populated based on mobile Model
Product_Id	DimProduct	Int	Dimension	Auto generated value
Model_Number	DimProduct	Varchar(50)	Dimension	
Product_Name	DimProduct	Varchar(50)	Dimension	
Product_Type	DimProduct	Varchar(50)	Dimension	
Manufacturer	DimProduct	Varchar(50)	Dimension	Populated based on mobile model
Description	DimProduct	Varchar(400)	Dimension	
Quarter_Id	DimQuarter	Int	Dimension	Auto generated value

Continued on next page

Table 3 – *Continued from previous page*

Column Name	**Destination**	**Data Type**	**Type**	**Transformation**
Quarter	DimQuarter	Varchar(20)	Dimension	All characters after and including removed to separate it from year
Quarter_Name	DimQuarter	Varchar(50)	Dimension	Populated based on Quarter
Year_Id	DimYear	Int	Dimension	Auto generated value
Year	DimYear	Int	Dimension	Separated from Quarter Year column
Software_Id	DimSoftware	Int	Dimension	Auto generated value
OS_Version_Number	DimSoftware	Varchar(50)	Dimension	
OS_Name	DimSoftware	Varchar(50)	Dimension	
Designed_By	DimSoftware	Varchar(50)	Dimension	Populated based on OS_Name
Sales_Id	FactSales	Int	Fact	Auto-generated value
Source_RowId	FactSales	Int	Fact	
Unit_Sales_Million	FactSales	Float	Fact	Comma removed and linear regression used to populate values based on revenue
Revenue_Million_Dollar	FactSales	Float	Fact	Comma removed in value, $ sign removed, and linear regression used to predict missing values.

6 ETL Process

Data from multiple sources are extracted and transformed to match the data warehouse schemas and loaded first into staging tables and then into Dimension and Fact tables.

ETL is often a complex combination of process and technology that consumes a significant portion of the data warehouse development efforts and requires the skills of Business Analysts, Database Engineers, and Application Developers.It is not a one-time event as new data is loaded into the data warehouse periodically-monthly, daily and hourly.

Extraction

Structured Dataset has been directly downloaded from Statista, but the unstructured data is scraped from Apple Inc., and Nokia Investors website using Data Miner Chrome plugin.

To scrap data using Data Miner, Recipes are created by opening the web page. Another easy way to scrap the data is to select some data and right-click on Get similar(Data Miner) option and it will automatically bring similar data to the CSV file. The .csv file can then be downloaded.

Transformation

The downloaded dataset is transformed, cleaned and staged into staging tables. Python code has been written to clean the dataset.Execute Process Tasks have been created to automate the cleaning process in SSIS. Units sold' dataset for all three companies has a single Execute process task.For the revenue dataset, three Execute Process task has been created, one for each mobile brand.This involves complex coding of Python because quarterly data was present in different files for each year. Spaces and $' was removed using Python to clean the data and all revenue data were merged to create a single file.

Transformation is done in each step of the SSIS using Data Conversion toolbox item. Finally, all the sources data is merged into a single CSV file to be loaded into the Staging table.

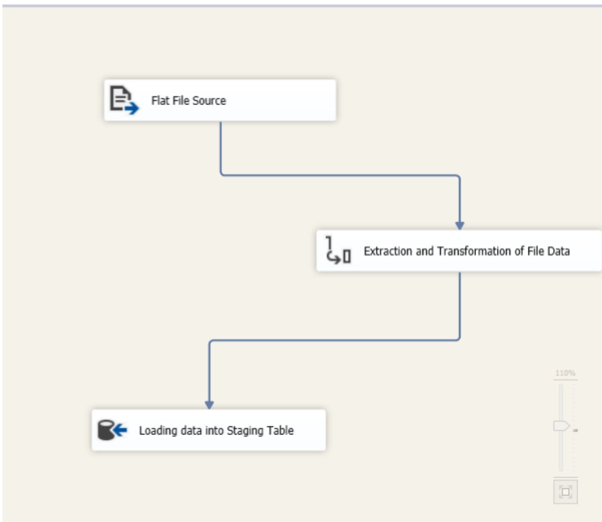


Figure 3: Data Transformation

Staging

There are 2 levels of staging being performed. First one by using Data Flow Task node, which loads the data from the CSV file into Source_Data_Staging table. Second is by loading the data from Source_Data_Staging table to Source_Data_Staging_Final table using Execute Flow task. During this step, data is further cleaned to remove null values and cast the data from String to Int for the required columns.

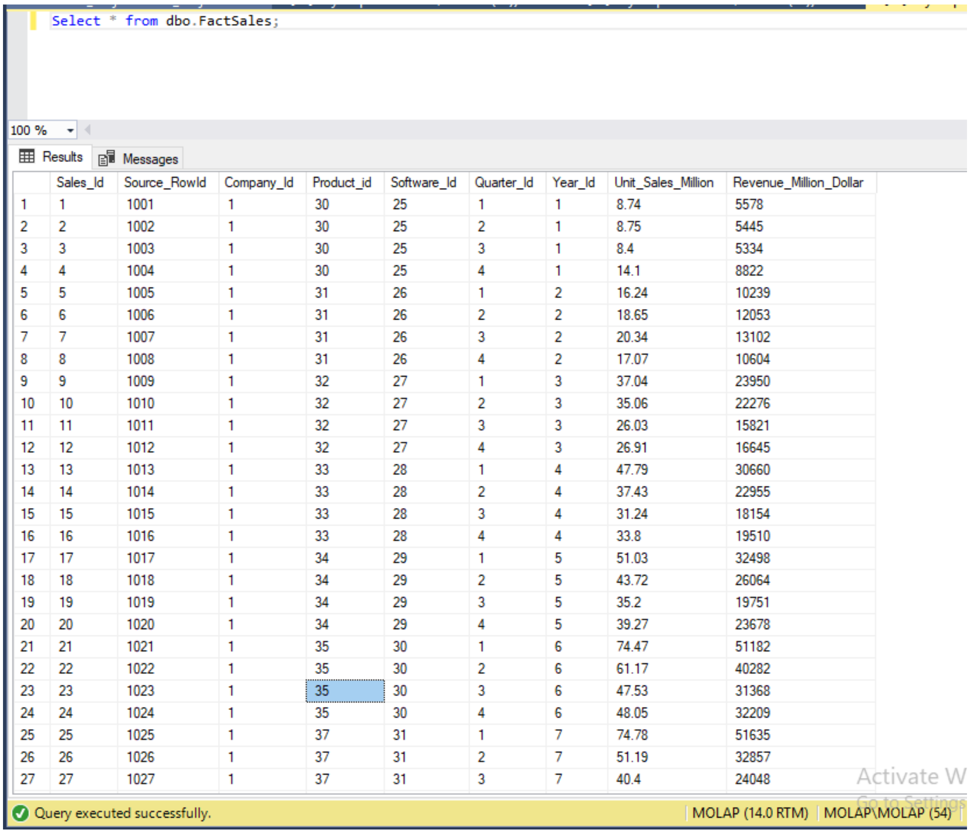
Loading

Once the data is loaded into Staging tables, Data is then loaded into each of Dimension tables from Source_Data_Staging_Final table using SQL query which selects the distinct rows from the staging table.

Once the dimension loading is performed, the primary key column value of the dimension table is updated into both the staging tables using Execute SQL Task.

After loading all the dimension tables, Fact table, FactSales is loaded from Source_Data_Staging_Final table using Data flow task. All the required columns are mapped in the OLE DB Source and Destination flow tasks.

There are no duplicate rows in the FactSales table. It follows 3rd Normal form.



```sql
Select * from dbo.FactSales;
```

100 %

Results | Messages

	Sales_Id	Source_RowId	Company_Id	Product_id	Software_Id	Quarter_Id	Year_Id	Unit_Sales_Million	Revenue_Million_Dollar
1	1	1001	1	30	25	1	1	8.74	5578
2	2	1002	1	30	25	2	1	8.75	5445
3	3	1003	1	30	25	3	1	8.4	5334
4	4	1004	1	30	25	4	1	14.1	8822
5	5	1005	1	31	26	1	2	16.24	10239
6	6	1006	1	31	26	2	2	18.65	12053
7	7	1007	1	31	26	3	2	20.34	13102
8	8	1008	1	31	26	4	2	17.07	10604
9	9	1009	1	32	27	1	3	37.04	23950
10	10	1010	1	32	27	2	3	35.06	22276
11	11	1011	1	32	27	3	3	26.03	15821
12	12	1012	1	32	27	4	3	26.91	16645
13	13	1013	1	33	28	1	4	47.79	30660
14	14	1014	1	33	28	2	4	37.43	22955
15	15	1015	1	33	28	3	4	31.24	18154
16	16	1016	1	33	28	4	4	33.8	19510
17	17	1017	1	34	29	1	5	51.03	32498
18	18	1018	1	34	29	2	5	43.72	26064
19	19	1019	1	34	29	3	5	35.2	19751
20	20	1020	1	34	29	4	5	39.27	23678
21	21	1021	1	35	30	1	6	74.47	51182
22	22	1022	1	35	30	2	6	61.17	40282
23	23	1023	1	35	30	3	6	47.53	31368
24	24	1024	1	35	30	4	6	48.05	32209
25	25	1025	1	37	31	1	7	74.78	51635
26	26	1026	1	37	31	2	7	51.19	32857
27	27	1027	1	37	31	3	7	40.4	24048

Query executed successfully. | MOLAP (14.0 RTM) | MOLAP\MOLAP (54)

Figure 4: FactSales Data

After FactSales loading, the Sales_Id is updated into both the staging tables using SQL Script.

Cube Deployment

Cube deployment is also automated in the end using Analysis Service Execute DDL Task. A separate node for Analysis services for Dimension and Fact is created.

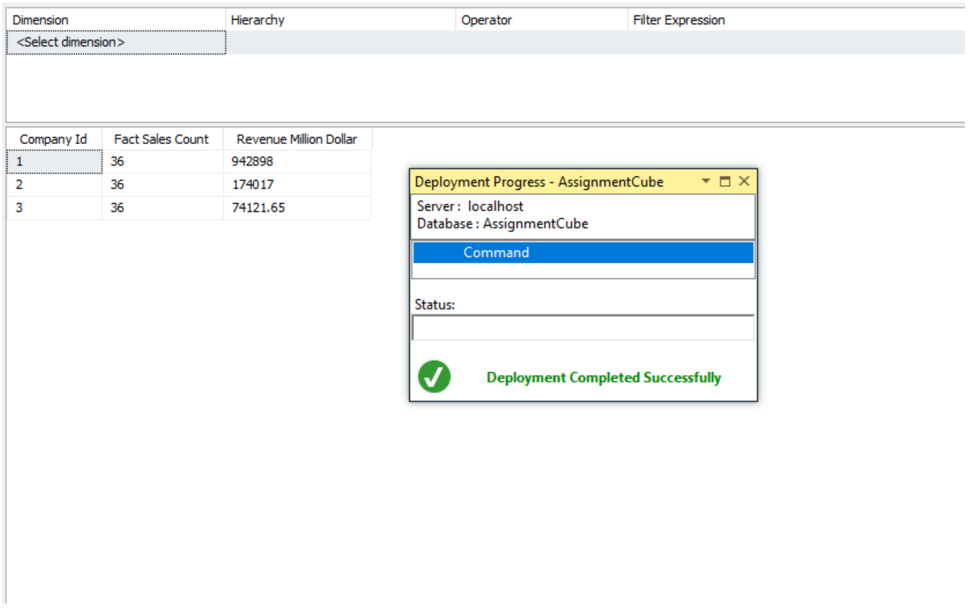


Figure 5: Deployed cube

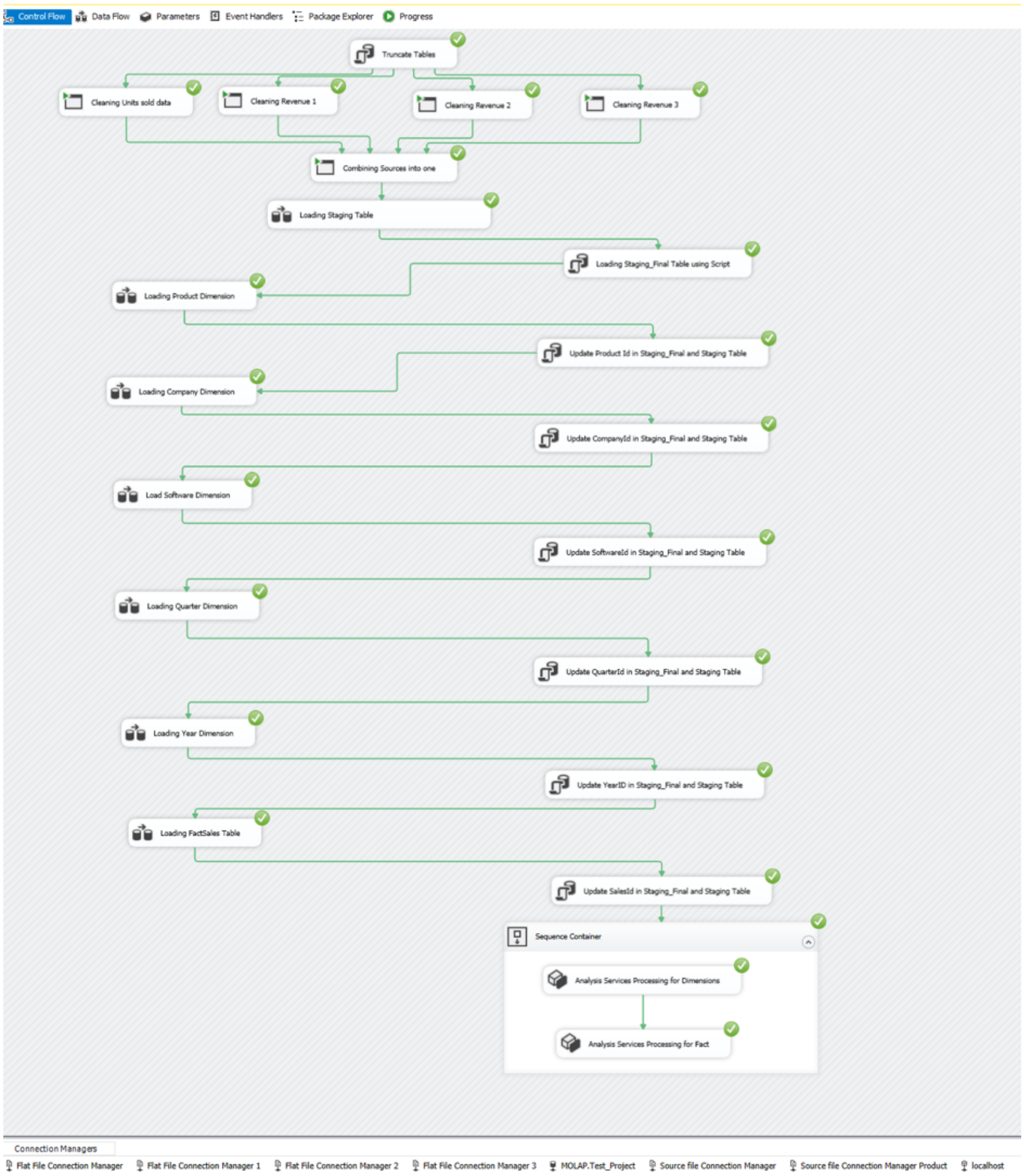


Figure 6: SSIS Workflow

7 Application

BI's major objectives are to enable interactive and easy access to diverse data, enable manipulation and transformation of these data, and provide business managers and analysts the ability to conduct appropriate analyses and perform actions(Turban et al. (2008)).

The Business Intelligence queries to fulfil the business requirements are given below.

7.1 BI Query 1: Quarterly Sales and Revenue of Apple, Samsung, and Nokia

For this query, the contributing sources of data are Statista for Unit Sales Quarterly data and Apple Investors and Nokia Investors website for Quarterly revenue data. Here, it has been assumed that the model launched in the year has contributed most to the revenue of that year.

The general findings are that even though the unit sales of all three competitors appear similar, the revenue generated is quite different. Apple mobile phones generate the highest revenue followed by Nokia and Samsung from the year 2010 to 2018. This huge gap in revenue can be attributed to the high selling price of Apple mobile phones.

These findings are illustrated in Figure 7.

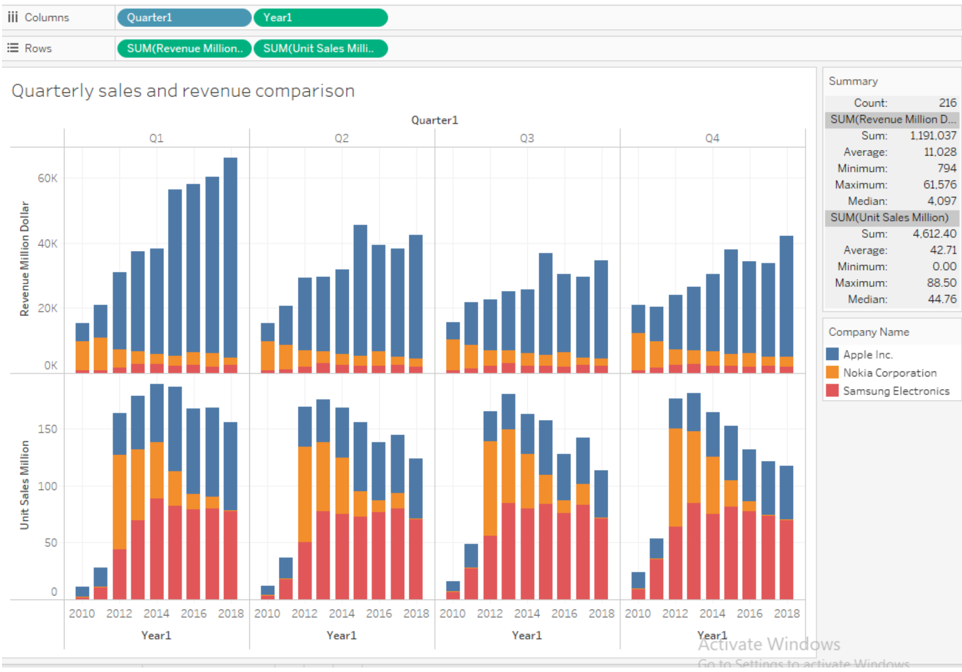


Figure 7: Results for BI Query 1

7.2 BI Query 2: Annual Performance of Apple, Samsung, and Nokia

For this query, the contributing sources of data are Statista for Unit Sales Yearly data and Apple Investors and Nokia Investors website for Yearly revenue data.

The general findings are that Apples revenue and unit sales keep on increasing steadily until 2015. Then because of the decrease in unit sales, the revenue comes down slightly. This decrease in sales can be attributed to iPhone XS or XS Max models since they were launched in 2015.

For Nokia, the steep decrease in sales started in the year 2012 and it never went up. Nokia Lumia 900 and 920 were launched in 2012.

For Samsung, the unit sales steadily increased till 2013 and stabilize after then. But because of low price models, it's revenue never reached near to that of Apple.

These findings are illustrated in Figure 8.

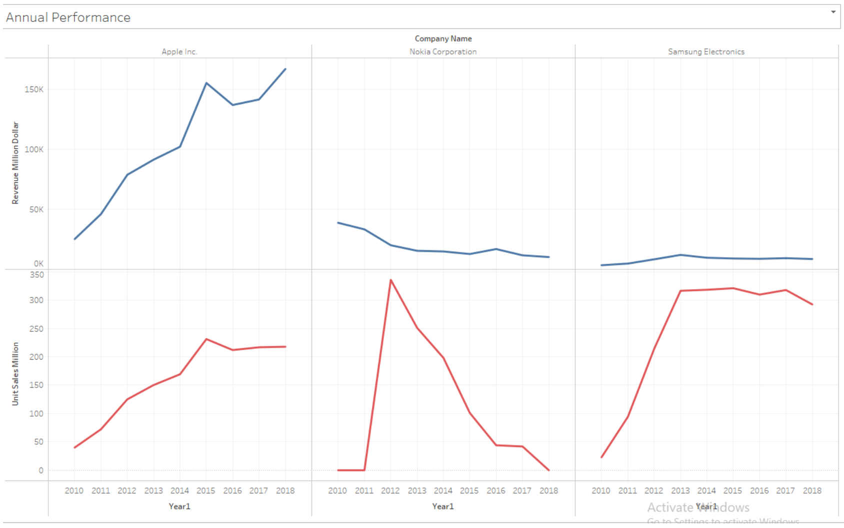


Figure 8: Results for BI Query 2

7.3 BI Query 3: Best Performing model among Apple, Samsung, and Nokia Mobiles

For this query, the contributing sources of data are Statista for Unit Sales data and Wikipedia for the mobile model information data.

The general findings are that Apple's iPhone 6S or 6S Plus has been the best performing model among all mobile phones based on the units sold. We cannot consider revenue to find out the best performing model because of the huge difference in selling prices of different models.

These findings are illustrated in Figure 9.

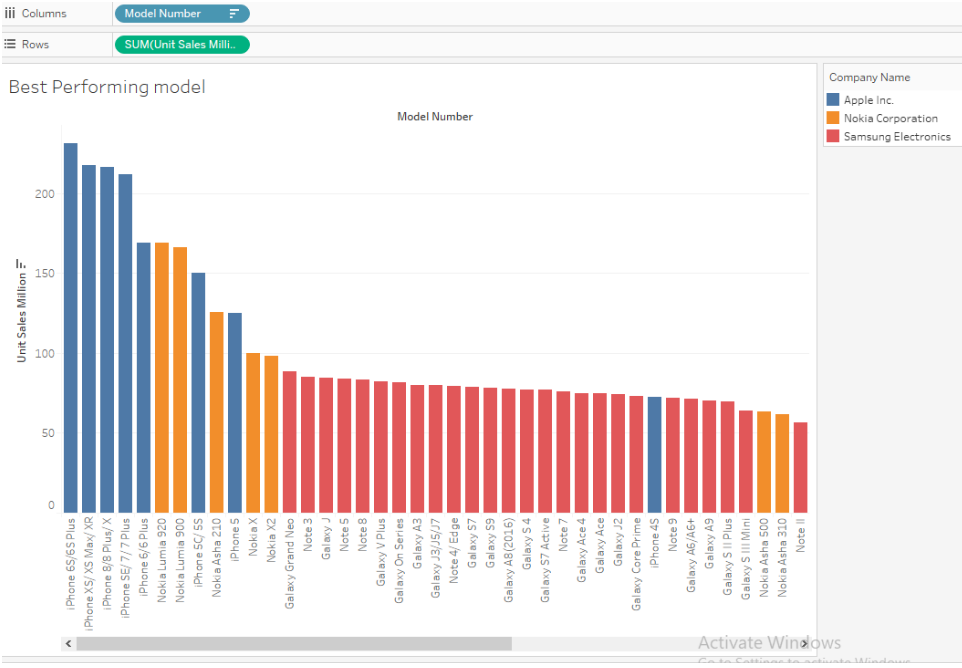


Figure 9: Results for BI Query 3

7.4 Discussion

The Business Intelligence queries mentioned above provide a detailed insight into the sales and revenue of Apple Inc., Samsung, and Nokia. Queries are very helpful in telling us how the annual performance of the three companies was from 2010 to 2018. Queries also let us know how the revenue and sales of the 3 companies have changed from 2010 to 2018. And last but not the least, queries also tell us which the best performing model is.

This data warehouse could partially answer some of the big questions people are looking for. This could tell us why Nokia lost its market share of Mobile phones. BI query tells us that the problem with Nokia starts in the year 2012 and since then it has not recovered. There are for sure other factors responsible for Nokia' s slump, but this is one of the major factors and we cannot ignore it.

The related work mentioned in 3.3 does not represent data warehouse, but it is just the yearly data about Apple and Samsung products. Neither it performs any comparison of different mobile companies nor it draws any insights from it. On the other hand, Mobile Sales Data Warehouse performs a comparison of sales and revenue of three major Mobile giants.

This data warehouse has certain limitations. It does not automatically performs imputation of missing values. SPSS has been used to predict the missing revenue values using Linear Regression.It analyses only three companys data about sales and revenue from the year 2010 to 2018. But these limitations can be easily overcome. More variables and factors can be added in the data warehouse easily to make it a big source of information. For Example We can add a greater number of mobile companies to make it a bigger warehouse. We can also add more historical data for the previous years to expand the scope of the study. Further, we can add other variables such as profit from Mobile phone sales to make a better comparison of the performance.

8 Conclusion and Future Work

Mobile Phone sales data warehouse was created to compare the performance of Mobile phone companies. The data warehouse fully answers the requirement of creating it. But as discussed, its scope can be extended to include more information. It could provide us better insights if we add more information in it.

Future work can focus on analyzing what features have changed or newly added in the mobile phones of Apple, Samsung and Nokia from 2010 to 2018 and then analyze the selling pattern to describe what all type of features a customer likes before buying a mobile phone. On that basis, future prediction about the selling behaviour of a mobile phone can be made. We can also consider adding more historical data. More the data, better the insights. We could also add the percentage of revenue a company generates from selling different products and then it can actually do a better comparison of company' s performance based on one type of product. We could also present the sales data based on Geographic regions.

Overall, this data warehouse provides enough information to compare Apple Inc., Samsung, and Nokia based on the mobile phone market. But there is always scope for improvement.

Appendix

Python code example for Cleaning Units Sold Data

```
import pandas as pd
import numpy as np
data_file = str(input("Enter the data file name"))
sheet_name = str(input("Enter the sheet name"))
brand_name = str(input("Enter the brand name"))
missing_values = ["NaN",""," ","n/a","na"]
df = pd.read_excel("/Users/manish/Documents/Apple_data/Done_files/" +
data_file,
sheet_name=sheet_name, na_values= missing_values)
def model_name():
    pass

new = df["Quarter Year"].str.split(" ", n=1, expand=True)
df["Quarter"]= new[0]
df["Year"] = new[1]
df["Year"] = df["Year"].str.replace("'",'20')

model = {'Apple': 'iPhone', 'Samsung':'Smart phone','Nokia':'Phone'}

if brand_name == 'Apple':
    df["Brand Name"] = 'Apple'
elif brand_name == 'Samsung':
```

```python
    df["Brand Name"] = 'Samsung'
elif brand_name == 'Nokia':
    df["Brand Name"] = 'Nokia'
else:
    pass

df["Model"] = df["Brand Name"].map(model)


#Dropping unnamed column
df.drop(df.columns[df.columns.str.contains('Unnamed', case=False)], axis=1,
inplace=True)
df = df.drop(['Quarter Year'], axis=1)
df = df[['Quarter','Year','Brand Name','Model','Units sold']]
cnt = 0
for row in df['Units sold']:
    try:

        if type(df.loc[cnt,'Units sold']) == str:
                df.loc[cnt,'Units sold'] = np.mean
        else:
            pass

    except ValueError:
        pass
    cnt+=1

df['Units sold'] = df['Units sold'].astype(float).astype(int)
output_file = str(input("Enter the output file name"))
writer = pd.ExcelWriter("/Users/manish/Documents/Apple_data/cleaned/"+
output_file, engine='xlsxwriter')

df.to_excel(writer)
writer.save()
```

SQL Code to update dimension Ids in Staging Table

```sql
--Query to update ProductId in Staging final Table
Update dbo.Source_Data_Staging_Final
  set dbo.Source_Data_Staging_Final.Product_Id = dbo.DimProduct.Product_Id
  FROM dbo.Source_Data_Staging_Final
 INNER JOIN dbo.DimProduct ON dbo.Source_Data_Staging_Final.Model_Number
 = dbo.DimProduct.Model_Number

--Query to update ProductId in Staging Table
Update dbo.Source_Data_Staging
  set dbo.Source_Data_Staging.Product_Id = dbo.DimProduct.Product_Id
  FROM dbo.Source_Data_Staging
  INNER JOIN dbo.DimProduct ON dbo.Source_Data_Staging.Model_Number
  = dbo.DimProduct.Model_Number
```

```sql
--Query to update SoftwareId in Staging final Table
Update dbo.Source_Data_Staging_Final
  set dbo.Source_Data_Staging_Final.Software_Id= dbo.DimSoftware.Software_Id
  FROM dbo.Source_Data_Staging_Final
  INNER JOIN dbo.DimSoftware ON
  dbo.Source_Data_Staging_Final.OS_VERSION_NUMBER=
  dbo.DimSoftware.OS_VERSION_NUMBER

--Query to update SoftwareId in Staging Table
Update dbo.Source_Data_Staging
  set dbo.Source_Data_Staging.Software_Id = dbo.DimSoftware.Software_Id
  FROM dbo.Source_Data_Staging
  INNER JOIN dbo.DimSoftware ON
  dbo.Source_Data_Staging.OS_VERSION_NUMBER =
dbo.DimSoftware.OS_VERSION_NUMBER

--Query to update CompanyId in Staging final Table
Update dbo.Source_Data_Staging_Final
  set dbo.Source_Data_Staging_Final.Company_Id =
  dbo.DimCompany.Company_Id
  FROM dbo.Source_Data_Staging_Final
  INNER JOIN dbo.DimCompany ON
  dbo.Source_Data_Staging_Final.Company_Name = dbo.DimCompany.Company_Name

--Query to update CompanyId in Staging Table
Update dbo.Source_Data_Staging
  set dbo.Source_Data_Staging.Company_Id =
  dbo.DimCompany.Company_Id
  FROM dbo.Source_Data_Staging
  INNER JOIN dbo.DimCompany ON
  dbo.Source_Data_Staging.Company_Name =
dbo.DimCompany.Company_Name

--Query to update QuarterId in Staging final Table
Update dbo.Source_Data_Staging_Final
  set dbo.Source_Data_Staging_Final.Quarter_Id =
  dbo.DimQuarter.Quarter_Id
  FROM dbo.Source_Data_Staging_Final
  INNER JOIN dbo.DimQuarter ON
  dbo.Source_Data_Staging_Final.Quarter = dbo.DimQuarter.Quarter

--Query to update QuarterId in Staging Table
Update dbo.Source_Data_Staging
  set dbo.Source_Data_Staging.Quarter_Id = dbo.DimQUarter.Quarter_Id
  FROM dbo.Source_Data_Staging
  INNER JOIN dbo.DimQuarter ON
  dbo.Source_Data_Staging.Quarter = dbo.DimQuarter.Quarter
```

Video Link:

https://www.youtube.com/watch?v=DXuOoecou5A

References

Chen, Z. (2001), *Intelligent Data Warehousing: From Data Preparation to Data Mining*, 1st edn, CRC Press.

Dayal, U. & Chaudhuri, S. (1997), 'An overview of data warehousing and olap technology', *ACM Sigmod Record* **26**, 65–74.

Elamin, E., Alzaidi, A. & Feki, J. (2018), 'A semantic resource based approach for star schemas matching', *International Journal of Database Management Systems* **10**, 15–28.

Golfarelli, M., Maio, D. & Rizzi, S. (1998), 'The dimensional fact model: A conceptual model for data warehouses', *International Journal of Cooperative Information Systems* **07**(1), 215–247.

Hammergren, T. C. & Simon, A. R. (2009), *Data Warehousing For Dummies, 2nd Edition*, 2nd edn, Wiley Publishing, Inc.

Hossain, M., Rahman, M. M. & Jahan, N. (2018), 'Motivations behind the purchasing decision of a smart phone', **12**, 141–150.

Hussain, R. (2019), 'International journal of research impact of price on customer satisfaction; mediating role of consumer buying behaviour in telecom sector', *International Journal of Research* **6**, 150–165.

Jukic, N. (2006), 'Modeling strategies and alternatives for data warehousing projects', *Communications for the ACM* **49**, 83–88.

Karjaluoto, H., Karvonen, J., Kesti, M., Koivumki, T., Manninen, M., Pakola, J., Ristola, A. & Salo, J. (2005), 'Factors affecting consumer choice of mobile phones: Two studies from finland', *Journal of Euromarketing* **14**.

Kimball, R. & Caserta, J. (2004), *The Data Warehouse ETL Toolkit*, 3rd edn, Wiley Publishing, Inc.

Kimball, R. & Ross, M. (2013), *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling, Third Edition*, 3rd edn, John Wiley & Sons, Inc.

Laberge, R. (2011), *The Data Warehouse Mentor: Practical Data Warehouse and Business Intelligence Insights*, 1st edn, McGraw-Hill Education.

Rakib, M. R. H. K. (2019), 'Factors influencing purchase intention of cellular phones among the university students in bangladesh', **11**, 92–101.

Shanka, M. S. (2013), 'Factors affecting consumer buying behavior of mobile phone devices', *Mediterranean Journal of Social Sciences* **4**, 103 – 112.

Turban, E., Sharda, R., Aronson, J. E. & King, D. (2008), *Business intelligence: A managerial approach*, Pearson Prentice Hall Upper Saddle River, NJ.

Wixom, B. H. & Watson, H. J. (2001), 'An empirical investigation of the factors affecting data warehousing success', *MIS Quarterly* **25**, 17–41.