

# **Data Analytics on Telecom Dataset**

**Supervised by: Ma'am Rubina**



**BY**

**Fatima Niaz**

**Shehryar Tahir**

**QUAID I AZAM UNIVERSITY ISLAMABAD**

**Institute of Information Technology**

**2016-2020**

## Acknowledgment

First, we thank **Allah Almighty** for giving us the courage, knowledge, and ability to complete this project. Without the blessing and kindness of Him, we are nothing and have no power and ability to do such work.

We would like to express our deepest gratitude to **our parents** for a lot of support throughout the academic career and our mothers for her love, affection, patience, encouragement, and prayers.

We express the deepest gratitude to our kind supervisor **Miss Robina Rashid** who kept our morale high with appreciation and motivation. Her friendly behavior with us is a primary factor that we can finish the project confidentially. We are incredibly fortunate to have her as our supervisor.

We are hugely grateful to **Muhammad Farjad Khan**, who supports us and guides us from the start till the end of a project. He was always available whenever we tried to approach him. His valuable suggestions significantly improved this project, and without his precious guidance, we would have never been able to develop such a project.

It is our bounden duty to pay tributes to our worthy teachers and staff members of the Institute of Information Technology, Quaid-i-Azam University Islamabad for their help and support to develop an understanding of the project.

Last but not least we wish to avail ourselves of this opportunity, express a sense of gratitude and love to our beloved fellows either seniors, juniors, classmates, or university fellows for their moral, manual support, strength, and help and for everything which they did for us.

**Fatima Niaz**

**Shehryar Tahir**

## **Abstract**

Telecom companies are unaware of the volume of data that could, on proper analysis, can get deeper insights into customer behavior, preferences, interests, and service usage patterns. This is what Data analytics is for Telcos. With the increasing adoption of smartphones and growth in mobile Internet, Telcos today have access to exceptional amounts of data sources including – customer profiles, device data, network data, customer usage patterns, location data, apps downloaded. This project is developed to give any telecom companies a way of making better business decisions to improve their business. The project provides end-users intelligent reports which show the trend of data visually using techniques like graphs, histograms, and maps, etc. The application takes the telecom data and transform, clean, and load into the data warehouse. It involves the process of ETL (Extract, Transform, and Load) in which we can load the data from the sources into the staging area. There we can transform data according to our requirements. Then load it into the Data Warehouse. The data in the data warehouse is being stored in facts and dimension tables. The analysis will be carried out on the data this leads to data visualization. A visual dashboard and report will be generated on that data. These visualized data reports then presented to the company manager, end-users. From these visualized reports the company takes better business decisions.

## Preface

**Chapter 1** Includes the introduction to the project, the existing System, and the proposed solution, the scope of the System, project objectives, and risk analysis. Brief description of software and hardware resources used.

**Chapter 2** Includes requirement analysis, functional, and non-functional requirements. It also includes the system requirements of the project.

**Chapter 3** Includes system design, system architecture, use case diagram, activity diagram, class diagram, ERD, sequence diagram.

**Chapter 4** Includes a brief overview of the tools and technologies used to implement the project.

**Chapter 5** Includes the significant component and activates to implement the project. It includes the main steps of implementation.

**Chapter 6** Includes the test scenario and test cases of the System.

**Chapter 7** Includes the conclusion to this project and future work regarding this project.

## Table of Contents

<b>CHAPTER 1</b> .....	11
<b>What is data analytics?</b> .....	11
<b>1.1 Introduction</b> .....	11
1.1.1 How Data Analytics works .....	11
<b>2) Know the sources of data</b> .....	12
<b>3) Access, manage and store data</b> .....	12
<b>4) Analyze data</b> .....	13
<b>5) Make intelligent, data-driven decisions</b> .....	13
<b>1.2 Storage of Data</b> .....	13
<b>1.3 Existing System</b> .....	14
<b>1.4 Proposed System</b> .....	14
Purpose: .....	15
<b>1.5 Why telecom companies need data analytics?</b> .....	15
<b>1.5.1 Benefits Data Analytics can bring to telecoms:</b> .....	17
<b>1.6 Objectives</b> .....	17
<b>1.7 Scope</b> .....	18
1.7.1 Resource Identification .....	19
1.7.2 Hardware Resources .....	20
1.7.3 Software Resources.....	21
<b>1.8 Risk Analysis</b> .....	21
1.8.1 When to Use Risk Analysis .....	22
<b>CHAPTER 2</b> .....	23
<b>Requirement Analysis</b> .....	23
<b>2.1 Introduction</b> .....	23
<b>2.2 Requirement definition</b> .....	23
<b>2.3 Requirement Elicitation</b> .....	24
<b>2.4 Requirements Analysis</b> .....	25
<b>2.5 Requirements Specification</b> .....	26
<b>2.6 Requirement Validation</b> .....	27
<b>2.7 Functional Requirements</b> .....	28
<b>2.8 Non-functional Requirements</b> .....	29

2.8.1 Other Non-Functional Requirements .....	30
<b>2.8.1.1 Performance Requirements.....</b>	<b>30</b>
<b>2.8.1.2 Safety Requirements.....</b>	<b>32</b>
<b>2.9 System Requirements .....</b>	<b>32</b>
<b>CHAPTER 3.....</b>	<b>34</b>
<b>System Analysis and Design.....</b>	<b>34</b>
<b>3.1 Introduction.....</b>	<b>34</b>
<b>3.2 Analysis .....</b>	<b>34</b>
<b>3.3 Design .....</b>	<b>35</b>
3.3.1 Design and Implementation Constraints .....	35
<b>3.4 System Architecture.....</b>	<b>35</b>
<b>3.5 Use Case Diagrams.....</b>	<b>36</b>
3.5.1 Elements of the Use Case Diagrams .....	37
3.5.2 User Classes and Characteristics.....	37
<b>3.6 Activity Diagram.....</b>	<b>40</b>
<b>3.7 Class Diagram.....</b>	<b>42</b>
<b>3.8 ER Diagram.....</b>	<b>43</b>
<b>3.9 Sequence Diagram.....</b>	<b>44</b>
<b>CHAPTER 4.....</b>	<b>45</b>
<b>Tools and Languages.....</b>	<b>45</b>
<b>4.1 SQL Server Management Studio (SSMS) .....</b>	<b>45</b>
4.1.1 SSMS System Requirements .....	46
<b>4.2 MS Visio.....</b>	<b>47</b>
4.2.1 System requirements for MS Visio.....	47
<b>4.3 Microsoft Word.....</b>	<b>48</b>
<b>4.4 SQL (Structured query language).....</b>	<b>49</b>
<b>4.5 VMware workstation pro.....</b>	<b>50</b>
<b>4.6 Erwin Data Modeler .....</b>	<b>51</b>
<b>4.7 Tableau Software.....</b>	<b>52</b>
4.7.1 Requirements for Tableau Desktop.....	53
<b>CHAPTER 5.....</b>	<b>54</b>
<b>System Implementation.....</b>	<b>54</b>
<b>5.1 Implementation Phase.....</b>	<b>54</b>

<b>5.2 Objectives of Implementation Phase .....</b>	<b>54</b>
<b>5.3 Project Objectives and Goals.....</b>	<b>55</b>
<b>5.4 Getting the data.....</b>	<b>55</b>
<b>5.5 Data Warehouse.....</b>	<b>56</b>
5.5.1 Data Warehouse Architecture .....	57
5.5.2 How does a Data Warehouse Work .....	58
5.5.3 Benefits of Data Warehouse .....	58
<b>5.6 ETL.....</b>	<b>59</b>
5.6.1 Extracting the data .....	59
5.6.2 Transforming the data .....	59
5.6.3 Loading the data.....	59
5.6.4 How ETL Works.....	60
<b>5.7 Data Analysis.....</b>	<b>60</b>
<b>5.8 Data Visualization .....</b>	<b>61</b>
5.8.1 How Data Visualization Works .....	62
5.8.2 Benefits of Data Visualization .....	62
5.8.3 Tool Used for Visualization.....	64
<b>CHAPTER 6.....</b>	<b>66</b>
<b>Testing.....</b>	<b>66</b>
<b>6.1 The Importance of Data Warehouse Testing .....</b>	<b>66</b>
<b>6.2 ETL Testing.....</b>	<b>66</b>
6.2.1 ETL Testing Process .....	66
6.2.2 Types of ETL Testing .....	68
• <b>Metadata Testing:</b> .....	68
• <b>Data Completeness Testing:</b> .....	68
• <b>Data Quality Testing:</b> .....	68
• <b>Data Transformation Testing:</b> .....	69
• <b>ETL Regression Testing:</b> .....	69
• <b>ETL Integration Testing:</b> .....	69
6.2.3 ETL Test Scenarios and Test Cases .....	69
6.2.4 Types of ETL Bugs.....	73
<b>6.3 Test Scenarios and Test Cases.....</b>	<b>73</b>
<b>CHAPTER 7.....</b>	<b>77</b>

<b>Conclusive Discussion and Future Work.....</b>	<b>77</b>
<b>7.1 Conclusion.....</b>	<b>77</b>
<b>7.2 Future Work.....</b>	<b>77</b>



## **List of Tables and Figures**

<b>Table and Figure</b>	<b>Page Number</b>
Table 1.1 Hardware Resources	20
Table 1.2 Software Resources	21
Figure 3.1 System Architecture Diagram	36
Figure 3.2 Detailed view of System Architecture	36
Figure 3.3 Use Case Diagram	39
Figure 3.4 Activity Diagram of System	41
Figure 3.5 Class Diagram	42
Figure 3.6 ERD	33
Figure 3.7 Sequence Diagram	44
Figure 4.1 SQL server management studio	47
Figure 4.2 MS Visio Home Page	48
Figure 4.3 MS Word Home Page	49
Figure 4.4 SQL Queries	50
Figure 4.5 VMware Workstation Home Page	51
Figure 4.6 Erwin Data Modeler	52
Figure 4.7 Tableau Desktop	53
Figure 5.1 Datasheet	56
Figure 5.2 Data Warehouse Architecture	57

Figure 5.3 ETL(Extract, Transform and Load)	60
Fig 5.4 Tableau Dashboard	65
Fig 6.1 ETL Testing Stages	67
Fig 6.2 ETL Testing	68

# CHAPTER 1

## What is data analytics?

### 1.1 Introduction

This project is about Data analytics on telecom dataset reporting and analysis. Data analytics is the science of analyzing data to form conclusions about that information. Many of the techniques and processes of knowledge analytics are automated into mechanical processes and algorithms that employment over data for human consumption.

Data analytics techniques can reveal trends and metrics that might rather be lost within the mass of data. This information can then be went to optimize processes to extend the general efficiency of a business or system. Data can be analyzed for insights that lead to better decisions and strategic business moves. Due to the advancement in tools and technologies, BI allows companies to gather data from the operational data stores, enterprise resource planning (ERP), and customer relationship management (CRM) and from multiple sources to prepare it for analysis and reporting. This may consist of dashboards and data visualizations that enable people to view the details of data graphically and diagrammatically to make better business decisions.

#### 1.1.1 How Data Analytics works

Before businesses can put data to work for them, they should consider how it flows among a multitude of locations, sources, systems, owners, and users. There is the following process involved in data analysis involves several different steps to taking charge of this "data fabric" that includes traditional, structured data along with unstructured and semi-structured data:

- Determine the data requirements
- Identify data sources.
- Access, manage and store the data.

- Analyze the data.
- Make data-driven decisions.

### 1) Determine the data requirements

The first step is to determine the data requirements or how the data is grouped. Data could also be separated by age, demographic, or gender. Data values could also be numerical or be divided by category.

### 2) Know the sources of data

The second step in data analytics is the process of collecting it. This can be done through a spread of sources like computers, online sources, cameras, environmental sources, or personnel.

- **Streaming data** comes from the Internet of Things (IoT) and other connected devices that flow into IT systems from wearables, smart cars, medical devices, industrial equipment, and more. You can analyze this data as it arrives, deciding which data to keep or not keep, and which needs further analysis.
- **Social media** data stems from interactions on Facebook, YouTube, Instagram, etc. This includes vast amounts of data in the form of images, videos, voice, text, and sound – useful for marketing, sales, and support functions. This data is often in unstructured or semistructured forms, so it poses a unique challenge for consumption and analysis.
- **Publicly available data** comes from massive amounts of open data sources like the US government's data.gov, the CIA World Factbook, or the European Union Open Data Portal.
- **Othe data** may come from data lakes, cloud data sources, suppliers, and customers.

### 3) Access, manage and store data

Modern computing systems provide the speed, power, and flexibility needed to quickly access massive amounts and types of data. Along with reliable access, companies also need methods for integrating the data, ensuring data quality, providing data governance and storage, and preparing

the data for analytics. Some data may be stored on-premises in a traditional data warehouse – but there are also flexible, low-cost options for storing and handling data.

#### **4) Analyze data**

Once the data is collected, it must be organized so it is often analyzed. The organization may happen on a spreadsheet or other sort of software that will take statistical data. With high-performance technologies like grid computing or in-memory analytics, organizations can choose to use all their data for analyses. Another approach is to determine upfront which data is relevant before analyzing it. Either way, data analytics is how companies gain value and insights from data.

#### **5) Make intelligent, data-driven decisions**

The data is then cleaned up before analysis. This means it is scrubbed and checked to ensure there is no duplication or error, and that it is not incomplete. This step helps correct any errors before it goes on to a data analyst to be analyzed. Well-managed, trusted data leads to trusted analytics and delegated decisions. To stay competitive, businesses need to seize the full value of data and operate in a data-driven way – making decisions based on the evidence presented by data rather than gut instinct. The benefits of being data-driven are clear. Data-driven organizations perform better, are operationally more predictable, and are more profitable.

## **1.2 Storage of Data**

As we know, our project is purely data related, and it covers the two main domains of information technology, i.e. database and business. The main product of our project is the data warehouse. Before the invention of the computer, data was stored in the filing cabinet. After the arrival of computer floppy disk and hard drives are used to store data. As I mention above that data is precious, so it has to be stored in a special place where the possibility of leakage and loss is zero.

The computerized databases started in 1960 when the use of computers became famous among private organizations. In 1970-72, E.F. Codd gave the idea of the relational database model in his paper which leads to the new concept of database among the people. In 1976 P. Chen proposed another model of database known as Entity relation or ER. Because of this model designers then focus on the data application rather than logical table structure. The advent of the Internet prompted exponential development of the database business. Normal work area clients started to utilize

customer server database frameworks to get to PC frameworks that contained inheritance information.

In a data analysis system, the key component is a data warehouse. So they provide a better way to collect and store data. In data warehouse data comes from multiple sources. Every source has its format of data, so data is organized first then store in the data warehouse. So we take telecom data from multiple sources and also generate some of it in the same format to increase the number of records. So, the next step is to make a business report from that data which can help the executive, manager, and several end users to make decisions effectively.

### **1.3 Existing System**

Information technology has grown to an extent where telecommunication companies have become part of our daily lives. For the majority of the urban people, the day starts and ends with watching the phone, either they'll be texting, calling people, or surfing the web. observing it from an information technology (IT) perspective, there is so much data being generated. Data flows into these servers owned by the telecom companies at a rate that is beyond our imagination. Terabytes and petabytes of data are pouring into organizations every single day. Every call that we make creates a call detail record in their Call Detail Record (CDR) servers. for example, Pakistan has about 4 major telecommunication companies and a couple of smaller companies. for instance, one among the companies has about 6 million active customers, and if on a mean each customer makes 2 calls a day, then the server has been loaded with 12 million call detail records. Just imagine the world population, the number of calls, and therefore the amount of data stored a day. Data flows into their servers every day, every minute, and possibly every second. Why do these telecom companies need to store this huge amount of data? How can they use this data for his or her benefit? The Business world always searches for data, conducts surveys to know their customers better and to design better products and plans. But, here we see is an industry that has so much data and finding it hard to form meaning out of it.

### **1.4 Proposed System**

Telecommunication companies maintain and store a tremendous amount of data about the Customer information, phone calls, and the operations of their networks. Due to the improvement of computer systems and telecommunication technologies, this industry has expanded Rapidly.

Data mining helps to improve the quality of service, detect the customer communication type, determine deceitfulness activities, and make better use of Resource.

## **Purpose:**

In the data-driven solutions, the customer demographic attributes, such as gender play a core role that may enable companies to enhance the offers of their services. Individual base analysis of how many individual users a one telecom company has and which company has more users in a certain period. Based on the geographic area we will find which **region and city** contains more users of the particular telecom company to target the customer at the right time and place.

This project will have two significant parts, and the first part is based on designing a data warehouse. There comes a step of ETL, In doing so, Data is going to be cleaned and we have to extract the data on which we can perform analysis. After extraction, Display the results using a Visualization Software for Telecommunication Data.

## **1.5 Why telecom companies need data analytics?**

Because a telecom company has got to continue with breakthroughs in technology, stay ahead in one among the world's best industries, and cling to a slew of regulations, the role of data analytics in telecom is to supply companies with the simplest way to uncover insights from all their data. using a telecom data analytics solution that will help a corporation gain better insights will generate more profits. Such a system also will enable the corporate to remain one step before its competitors and better anticipate its customer's needs.

### **1. Telecom Data Analytics Allows Better Use of Investments**

With the easy-to-use data analysis that comes from a telecom data analytics solution, a corporation can make decisions supported data, not guesswork. The role of data analytics in telecom is to offer each company a unified view of their data across departmental lines. When data streams in from multiple data sources during the corporate, the corporate can cash in on all its team's suggestions to return up with the simplest solution for each challenge.

## **2. Give Data Analytics Top Priority for Telecom Success**

With data pouring in through multiple sources and from all the departments, a telecom service provider can use all of the skills of its employees to deal with issues, find innovative new workflows, and keep its customers happy. With the Necto Telecom data analytics system, employees will see the data displayed in stunning infographics in an easy-to-understand format. they're going to be ready to add their notes to the data to get even more insights into how the corporate is performing.

## **3. Data Analysis of Top Importance for Telecom**

A telecom company cant afford to be left behind when it involves data analysis. consistent with data analysis thought leader Gartner, telecommunications services will have spent nearly two-fifths of the world's total spending on smart data and cloud technology. If your company isn't on board with a world-class telecom data analytics system you'll be soon left with a dinosaur death blow for the data-needy telecom industry.

## **4. Connect the data Dots with Unified Business Insights**

In a telecommunications and media service provider, every department generates plenty of data. Without a unified data analysis, other departments cant learn from each other's insights into the data. Furthermore, without a centralized solution, the data will often differ from that of other departments, confusing even anger.

With a centralized data analytics solution, all the department's data merge into one central system. Each department can add its insight into the data gleaned, giving another perspective which will cause an answer more quickly. All the departments can learn from one another, increasing efficiency and boosting productivity.



## **5. Imagine a far better Future with Telecom Data Analytics**

To better understand the role that data analytics plays in telecom, imagine a system during which all the departments in your company can share insights and resources. All users can access the data consistent with the permissions previously set by the IT department, making this a governed and secure solution. Putting their collective heads together to lower service costs, drive expansion, and act immediately when key performance indicators (KPIs) change, your departments will function together smooth profit-generating machine.

### **1.5.1 Benefits Data Analytics can bring to telecoms:**

Even the simplest use of technology can dramatically improve your business' productivity and efficiency. There are a lot of reasons why organizations and companies adopt data analytical tools. This data analysis can help decision-makers to enhance and organization and to make effective decisions in the following ways.

- Making smarter investment decisions
- Visibility into the profitability of different departments
- Reducing fraud
- Improved risk management
- Increased sales
- Smoother network operation
- Enhanced customer experience and reduced churn rate
- Cutting off operations that drain the budget
- Making the operations more efficient
- The increased average revenue per user

## **1.6 Objectives**

With the insight generation capabilities of data analytics, businesses have scored heavily in securing competitive advantage, by knowing what is working and what is relevant to their customers.

The objectives of data analytics are:

- **Understanding and Targeting Customers:** Data analytics helps an organization understand its customers better, and helps it narrow down the target audience, thus improving its marketing campaign.
- **Taking Strategic Decisions:** With data analytics, businesses can make data-backed decisions. No need to fumble in the dark, when the huge volume of data provides practically every bit of information about your business, market, customers, industry, and competition.
- **Cost Optimization:** Costs can be better optimized when you have the data to know which elements are draining costs but not returning a high value. For instance, the healthcare sector can utilize data analytics to find the cause of a cost hike in healthcare facilities.
- **Improving Customer Experiences:** Take the case of retail. Data analytics can help retailers & wholesalers show customer reviews about the product's quality & delivery time, thus improving customer experiences in the retail buying process.

Data analytics plays a crucial role in taking business decisions. It enables an organization to accumulate and analyze large amounts of data to resolve problems surrounding the organization.

Then we perform the data analysis and visualization to generate reports and dashboards which can help the company make better and accurate business decisions. Once a data warehouse is ready then we can run different queries on data according to the requirements. So with the help of reports and dashboards, one can make healthy business decisions that can enhance the company and maximize the profit.

## **1.7 Scope**

Project scope is also one of the concerns of project planning that involves deciding and documenting a list of several project goals, deliverables, tasks, costs, and deadlines. The documentation of a project's scope, which is called a scope statement, terms of reference, or statement of work, explains the boundaries of the project, establishes responsibilities for each team member, and sets up procedures for how completed work will be verified and approved. During the project, this documentation helps the project team remain focused and on task. The scope

statement also provides the project team with guidelines for making decisions about change requests during the project. Please note, a project's scope statement should not be confused with its charter; a project's charter simply documents that the project exists.

- Data has augmented the demand of information management specialists so much that leading organizations such as Microsoft, Oracle Corporation, IBM, SAP, HP, EMC, and Dell have invested around \$15 billion on software companies having expertise in data analytics and management. Enterprises of all time are exploring different ways to fetch a business value from their collected data through data analytics solutions. They have moved their focus to data analytics as the principal source of getting this essential value and are engaged with data analytics solutions based on their technology capabilities. But our scope is limited to reporting, data analysis, and data visualization. In other words, we are working on the telecom dataset of Pakistan includes **Zong, Jazz, Warid, Ufone, and Telenor**. Our System has data of customers as input, extract from many sources in the form of CSV and fixed-width files, etc. after this we clean the data and store it in the data warehouse then we perform analysis on it. Simple data is now converted into information in the form of intelligent reports and dashboards, which can help the user to make well business decisions. In the data-driven solutions, the customer demographic attributes, such as gender play a core role that may enable companies to enhance the offers of their services. We'll check the total subscriber base on our data, how many customers are active and inactive, how old are our customers, and Yearly check customers in our data. We can find how many individual users are using more than one telecom company sim and which have company more users in a certain period. Based on the geographic area we can find which **region and city** contains more users of the particular telecom company to target the customer at the right time and place.

## 1.7.1 Resource Identification

Here are some points that we must keep in mind during the process of identification.

- Make sure that we have a clear brief before the project starts. The client and we should have a common understanding of what is included in the project. Confirm the scope with

whoever needs to approve the work going ahead. This is the definitive list of everything that is expected on the project. Of course, that might change as work progresses, but our need to have a starting point from which to plan your work and the people who should be involved. When we know what the scope of the project is, we can start to look at the resources required. Breakdown the work into the parts. As we start to build out your project plan, we'll need to draw on the subject matter experts from the team. They will help us identify the tasks to be done and give us an idea about the skills required to complete the tasks. When the project schedule comes together, we'll have a clear view of what kind of person is required to deliver each of those activities, even if we don't have exact names at this point.

- It's better to plan our resource needs. There is too much risk involved with waiting until we need a particular resource and then trying to book them. That person might be already fully committed to another project, or on vacation, or so on. This exercise of booking resources in advance is part of capacity planning. It's essential to make sure that we have a reliable flow of work through the organization. Capacity planning means that our resources have enough to do at all times. No one is waiting for work to come in, and everyone knows what the next quarter's commitments will be. Planning for our project helps our client have confidence that the project will take the length of time we've said, and helps the business manage the throughput of assignments.

We have to use the following resources to build this project.

### 1.7.2 Hardware Resources

<b>1</b>	<b>System</b>	<b>Intel® Core™ i5 CPU</b>
<b>2</b>	Processor	2.40 GHz
<b>3</b>	Hard Disk	750 GB
<b>4</b>	RAM	8 GB(minimum)

5	System Type	64Bit Operating System, x64-based processor
---	-------------	---

**Table 1.1**

### 1.7.3 Software Resources

1	Operating System	Microsoft Windows 10
2	Development Tools	SQL server management studio, Tableau,  Erwin Data Modeler, VMware Workstation Pro
3	Documentations	Microsoft Word, Microsoft Visio
4	Language	SQL

**Table 1.2**

### 1.8 Risk Analysis

The risk may involve two things i.e. an expectation of loss and a potential problem that may be or may not occur in the future. It may be caused due to the absence of information, control, or time. A possibility of suffering from loss in the software development process is called a software risk. Loss can be anything, an increase in production cost, development of poor quality software, not being able to complete the project on time. Software risk exists because the future is uncertain and there are many known and unknown things that cannot be incorporated in the project plan.

Now come on the risk analysis. Risk analysis is a component of risk management. Risk analysis is the evaluation of the risks link with a specific event or action. It is applied to projects, information technology, security issues, and any action where risks may be evaluated on a quantitative and qualitative basis. To carry out risk analysis, we have to find the expected threats that we may face and then calculate the probability that these threats will happen.

## 1.8.1 When to Use Risk Analysis

We may use risk analysis in the following circumstances:

- During the planning of a project, risk analysis helps us to foresee and neutralized the possible problems.
- When moving forward to the next step in the project is a question mark we use risk analysis.
- During the management of potential risks in the workplace, risk analysis will prove helpful.

As we know the key component of our project is a data warehouse so we face the errors related to the data. We also use countermeasures to eliminate these errors to get better performance of a system.

- **Data duplication:** As we know we have to build a data warehouse. So we are mentally prepared to play with numerous data. Telecom data could contain users that are registered on two different networks or could have a person that has two or more SIMs registered with the same network. So the possibility of duplication of data exists. To encounter this we have to clean and transform the data e.g. in ETL when we are integrating data that came from the different sources.
- **Data corruption:** There must be a situation when the company or user data is corrupt. In this particular case, we simply need new user data. The data cleansing technique of ETL works here.
- **Data Loss:** we can also face a rare or odd situation in which we may lose the data unintentionally. We can retrieve these data by running the backup files for the data.

## **CHAPTER 2**

# **Requirement Analysis**

### **2.1 Introduction**

The requirement is a statement that identifies a product or process operational, functional, or design characteristic or constraint, which is unambiguous and necessary for product or process stability. Requirement engineering, also called requirement analysis, contains tasks that go into determining the needs or conditions to meet for a new or altered product, taking account of possible conflicting requirements of different stakeholders, such as customers and users. Requirements must be testable, actionable, measurable, related to identified business needs, and completely defined.

Our efforts should be directed towards ensuring that our final product conforms to customers' needs rather than attempting to mold user expectations to fit requirements. Requirements should be specified. There are high chances of project failure if conditions are not defined because the customer does not get what he wants. So, requirement analysis is the most critical phase of any project.

We analyze each requirement by requirement engineering process. The activities involved in requirements engineering can be broken down into several subareas or processes, which generally consist of elicitation, analysis, specification, and validation of requirements. These activities are regarded as the most common and essential phases in requirements engineering. In this chapter, the requirements of the proposed system are discussed in detail to achieve the goals and objectives of the System.

### **2.2 Requirement definition**

Requirements definition is the most crucial part of a project. Incorrect, inaccurate, excessive, unclear, and implementable requirements result in schedule delays, wasted resources, customer dissatisfaction, and even project failure. Requirements, therefore, should be adequately defined. Requirement analysis should begin with business requirements. Business requirements are statements from a business point of view, describing what kind of needs, goals, or visions there

are to be accomplished, often to increase the value of a business operation. Business requirements are regarded as high-level requirements, which are initial problems that should be solved.

Then the next step is to define product requirements. Product requirements are requirements concerning some kind of product e.g. a system or software that is being developed, and describe its features. Product requirements may form a solution for a business requirement, how it's going to be accomplished, and therefore in such a case, the product requirement is a lower level requirement. Product requirements can be further categorized as functional and non-functional requirements.

Functional requirements describe the functionality and capability of what a system is required to do, while non-functional requirements describe restrictions and constraints which could affect solutions and how a system may operate, e.g. availability, maintainability, interoperability, portability, reliability, safety, security. The non-functional requirements could be even further categorized based on what they are concerning. Process requirements are defined after that. Process requirements describe the way of performing, how to handle a task, and how something should be done.

In software engineering, in the process of product development, a process requirement may define what programming language, tools, and software should be used, or sets other requirements that could restrict available options. The many types of requirements are affecting each other; they are linked together by relationships through different levels. Commonly high-level requirements are linked to several low-level requirements, even though they may be linked in the other direction as well. The way how requirements are transformed from a high level into more low-level ones is called requirements tracing; it's the comprehension of relationships between requirements on different levels. Tracing of requirements and their relationships on different levels will allow and improve certain things, e.g. improved progress measuring, calculation of benefits vs expenditures, impact analysis.

## **2.3 Requirement Elicitation**

Requirements Elicitation is also known as requirements gathering. This phase of requirement engineering includes the task of identifying various requirement types from stakeholders or project documentation. Requirements elicitation may also be described as collecting requirements or



requirements gathering. However, elicitation is the most commonly used terminology because one should not expect to receive all requirements in an accurate way by merely questioning the users and stakeholders what a system should do. The main intentions include getting an overview of a problem that is yet to be solved, e.g. by a system or software that is going to be made according to the purposes.

This is achieved by collecting requirements by various means and techniques from the Stakeholders involved must be able to question the stakeholders and get them to talk openly during interviews and meetings so that they can reveal their requirements. Before being able to start the elicitation work itself, the possible requirement sources must be known and identified, i.e. the stakeholders and equivalent people, like the customer and end-users which could be the target group for a new system. It is important to have as many sources and alternatives as possible, as it will potentially improve the results by the number of different viewpoints when gathering requirements across the board.

Different elicitation techniques are commonly used to collect the requirements; these may include stakeholder and expert interviews, meetings, observation, different inquiry methods, and surveys. Interviews can be held individually, but it would be advantageous to arrange group interviews or meetings where people can openly discuss their points of view regarding their needs for better overall awareness. The discussion between stakeholders may also result in the realization of possible conflicts, which would have a better opportunity to get solved because of the early phase in the development process. Here are some general technique guidelines for requirements elicitation and interviews; identify as many stakeholders as possible especially key persons, work with many types of stakeholders handling different areas of work, document everything about requirements, question stakeholders for reasons and purposes of their requirements, but don't ever judge requirements. The requirements elicitation phase is followed by the analysis and specification of the requirements.

## **2.4 Requirements Analysis**

Requirements analysis determines if the gathered requirements are clear, complete, free of contradictions, implementable, and consistent. The analysis also handles any ambiguous requirements that do not clearly state what needs to be implemented, which could create a loss of

resources and time if identified later in the testing development phase. Requirement analysis involves identifying the stakeholders and taking their needs into account to help them understand the implications of designing the new System, along with what modules are worth implementing and which modules are more cost-efficient, and then to create a software requirement specification document. To elicit or gather the stakeholders' requirements, different processes, such as developing a scenario or user stories and identifying the use case which is being used for the project can be utilized. To gather the requirements of the project from stakeholders, the first step for analysts is to identify the stakeholders. Stakeholders are people or organizations that are directly involved in the project or affected by the project activities. Steps to identify stakeholders are given as:

- Any person who operates the System
- Any person benefits from the System
- Any person who is involved in purchasing the System, directly or indirectly.
- People or organizations who are opponents of the System.
- Organizations are responsible for the system design.
- Organizations that regulate the financial aspects of the System.
- Organizations that regulate the security aspects of the System.

Once the stakeholders are successfully identified, interviews are conducted through different processes; the needs and requirements of the System are identified, and a requirements specification document is prepared. The document is then discussed with major stakeholders to identify any ambiguous requirements and understanding of the System.

## **2.5 Requirements Specification**

The requirement specification is the third phase of requirement engineering. This step involves documenting the requirements in various forms, including summary lists, visual documents, natural language documents, user stories, use cases, or process specifications. The requirements specification document contains a complete description of how the system is expected to perform. It consists of all the requirements required for project development.

The requirements specification practices may, therefore, include the making of several, but in general three, different types of documents; a system definition document, a system requirements specification, and a software requirements specification. The system definition document could be considered as high-level requirements documentation, which describes the System and its intentions on an overall level. This documentation's target audience is primarily users of the System and customers needing a complete understanding of the System.

The system requirements specification can be defined as a common specification document, which could be used in any kind of system engineering task. The software requirements specification (SRS) is generally in software engineering the most commonly used type of documentation, which describes the software that is going to be developed. The SRS can be defined as a collection of structured data, which reflects the requirements of the system software. The documentation should include information making it possible to assess development costs, risks, and the time needed for development.

## 2.6 Requirement Validation

It is the process carried out to satisfy the customers with the specified requirements. We aim to find what's wrong with the requirement and try to identify the problems regarding requirements. Validation activity consists of more than one task to be performed e.g. Sessions with clients, partners, and useful specialists to decide alleviation and issue-resolving plans for unacceptable requirements before the project move into the advanced stage. In doing so, different checks are carried out on the requirements. They are:

1. **Validity checks:** The requirements proposed by stakeholders should be compared with what the System needs to execute.
2. **Consistency checks:** Requirements in the document shouldn't conflict or different explanations of the same function.
3. **Completeness checks:** The document should be complete i.e. consist of all the requirements and restrictions.

4. **Realism checks:** To make sure that the requirements can be applied using the knowledge of existing technology.
5. **Verifiability:** Once requirements are written so that they can also be verified including a set of tests that signify that the System fulfills the specified requirements.

## 2.7 Functional Requirements

Functional requirements are those which define the performance and functionality of a system and subsystems. It is a document that listed the activities and operations that a system must be able to perform. The key fields which are the essential part of the functional requirements specification document are scope, Business process, functional requirements, data and integration, security requirements, and performance. For reading the functional requirements document a one should have an understanding of a system, but no specific technical required or essential. Functional requirements are product features or functions that software developers must apply to ensure users achieve their tasks. So, it's essential to make them clear both for the development team and the stakeholders. Generally, functional requirements explain system behavior under particular conditions.

Similarly, business intelligence systems also have a lot of functional requirements. Because it consists of many subsystem components that join in an effective way to produce an entire system. So the following are the functional requirements of the above-described System.

- **Data Extraction:** Extract data from different CSV and fixed-width files, flat files, Excel spreadsheets, or from other different sources.
- **Data Staging area:** It is also known as ETL(extract, transform and load) from the different data sources as explain above and then transmit to the data staging area where it is processed, cleaned up, and converted into the standard form before loaded it to the data warehouse.

- **Data warehouse:** Cleaned data then loaded into the data warehouse, data warehouse act as a repository for storing data from the different sources. The data in the data warehouse stores in the form of fact and dimension tables.
- **Data Analysis:** Data stored in the data warehouse is then processed for data analysis. We perform different business-related queries on that data to understand it in a better way and to make that data helpful for an organization to make meaningful decisions ahead.
- **Data Visualization:** That data then load into the business intelligence software i.e. power BI by Microsoft or Tableau. This software generates reports and creates a dashboard from the data that will provide the company with an easy way to understand the trend in the data. So with the help of these intelligent reports, one can make informed business decisions.
- **Data Presentation:** Now we can adopt different methods of presenting such reports. We can present it on the web application or the excel spreadsheet or even on our mobile phone.

## 2.8 Non-functional Requirements

In addition to the prominent features and functions that are part of the System, other requirements don't actually do anything but are very important characteristics nevertheless. These are called non-functional requirements.

Non-functional requirements explain how a system must act and establish constraints of its functionality. This type of requirement is also known as the System's quality attributes. These requirements can specify the criteria that can be used to judge the operation of a system rather than the particular behavior. Non-functional requirements cover all the remaining needs which are not part of functional requirements.

The following are some non-functional requirements.

- **Usability:** Degree to which a system is used to achieve quantified objectives in terms of use. It refers to the continence and practicality of use. As far as our Business Intelligence system is concerned, it has to be able and fit to use.

- **Reliability:** In sample words, reliability means that a system can respond and perform its planned operations in all scenarios and conditions. The System is available whenever it is needed.
- **Efficiency:** The efficiency of a system defines satisfaction of a purpose without miss use of an asset for example memory and transfer speed.
- **Performance:** The System must perform all basic functionalities successfully. The performance of the Business Intelligence system depends upon the efficiency of intelligent reports generates by the System.
- **Future Enhancement:** The world is not static it's always a dynamic. People, places, trends, sciences, requirements, and technology are evolving, dynamic, and changing continuously. So this dynamic nature of the world must be kept in mind during the development of any software. So due to evolution in everything a system needs some additional features in the future that are not a part of the System earlier. So we must design a system in such a flexible way that it can't resist the change. The System should design in such a way that it supports the enhancement.
- **Multi-platform delivery:** Users can access the same application on multiple platforms. This attribute is specified for the apps, web applications, and desktop apps.

## 2.8.1 Other Non-Functional Requirements

### 2.8.1.1 Performance Requirements

To assess the performance of a system, the following must be specified:

- Response Time
- Workload
- Scalability
- Platform

#### 1. Response Time:

In some cases, the system response times are identified as part of a business case, for example, a criminal's fingerprint needs to be identified while the criminal is still in custody (less than an hour). In some cases, the response time will be dictated by legal requirements although this is rare. For general applications asking users to what is an acceptable response time is like asking people how much salary they require! The whole process is simply a process of negotiation.

## **2. Workload**

Again the business case or existing procedure ought to be the beginning of the workload definition. Nonetheless, it isn't sufficient to express that "the framework ought to be equipped for supporting 80,000 clients" or "the framework ought to have the capacity to help 4 pages/sec". These announcements are regularly great measurements at an abnormal state the board level yet don't characterize the work that the framework must help. This is especially critical as the blend of exchange influences the execution. For instance, a DB framework may effortlessly deal with 10,000 read exchange for every hour except just 3,000 refresh exchanges for every hour.

The in all probability exchanges to indicate are the user started exchanges however care must be taken to consider every one of the users of the framework and the group forms. For instance, a framework may have outer clients, inward staff giving information passage and clump procedures, for example, backups. If the backup isn't finished medium-term then it might genuinely disturb the execution experienced by the users the following day.

## **3. Scalability**

In one regard adaptability is just indicated as the expansion in the System's workload that the System ought to have the capacity to process. The adaptability required is frequently determined by the lifespan and the development of the System. For instance, another (and henceforth juvenile) system could endure a startling development in prevalence and experience the ill effects of a huge increment in workload as it winds up famous with new users. Progressively develop systems which speak to enhancements for more established systems are probably going to have all the more precisely characterized workloads and hence be more averse to endure in this regard Make sure to determine that the reaction time prerequisites should even now be meet as the workload scales An issue with adaptability specification is that it may not be monetarily practical to test the versatility, as it frequently requires extra equipment. In this way, an option is either to lease in the extra

equipment with the end goal of the tests or to utilize an extrapolation method, for example, a reproduction show.

### **2.8.1.2 Safety Requirements**

#### **1. Corrupt data:**

There must be a situation when the user or company data is corrupt. So we cannot understand the corrupt data. To encounter this, we need new user data.

#### **2. Data duplication:**

There exists duplication in the data, this case analysis issue. To encounter this, we have to clean the data and transform the data.

#### **3. Data lose:**

We can face a situation where we lost the data unintentionally. We can retrieve these data by running the backup files for that data.

## **2.9 System Requirements**

System requirements include the software, hardware, and user interface requirements. It means that to access that application and use it effectively, you must have a system that fulfills the required specifications.

### **1. User Interface Requirements**

GUI (Graphical User Interface) along with meaningful frames and interface.

### **2. Software Requirements**

The operating system will be required. Microsoft Windows 8 and above.

### **3. Hardware Requirements**

Minimum hardware requirements are:

- 1.5 GHz CPU
- 2 GB RAM



- 256 HARD DISK
- 1024\*1024 of Display

# CHAPTER 3

## System Analysis and Design

### 3.1 Introduction

System analysis and design consist of planning the development of systems through understanding and identifying in detail what a system should do and how the subsystem components should be applied and work together. System analysts decipher business mess through analyzing the requirements of information systems and designing such systems by applying analysis and design tools and techniques. It is a systematic process that has phases such as planning, analysis, design, deployment, and maintenance.

### 3.2 Analysis

The analysis states what the System should do. It is an activity of gathering and explaining facts, recognizing the problems, and decomposition of a system into its components. System analysis is conducted for the motive of studying a system or its parts to identify its aims. It is a problem-solving technique that surpasses the System and ensures that all the components of the System work efficiently to fulfill their purpose.

This section addresses security considerations. Key security activities include

- Control the risk evaluation and use the results to increase the baseline security controls.
- Examine security requirements.
- Perform functional and security testing.
- Prepare inception documents for system certification and accreditation.

Although this section has the information security elements in sequential top-down order, the order of completion is not obligatorily fixed. Security examination of complex systems will need to be repeated until stability and completeness are attained.

### **3.3 Design**

System Design focuses on how to accomplish the objective of the System. The process of defining the architecture, components, and data of a system to satisfy specified requirements to complete a project. Before planning, we need to understand the old System thoroughly and determine how computers can best be used to operate efficiently.

#### **3.3.1 Design and Implementation Constraints**

As we are operating on the laptop, so hardware and memory limitations are always there. So to manage the hardware and produce a result in a given limited time.

To handle the complexity of data, we are forced to work on the data of one year only. So 2qssall the reports we are going to develop from the telecom data is consisting of a limited period. We also face a problem regarding the collection of data. A company doesn't want to compromise its confidentiality so often they prevent to share their data with anyone. In doing so, we can get data from the Internet.

### **3.4 System Architecture**

The system architecture is the conceptual model that characterizes the structure, behavior, and more perspectives of a system. An architecture depiction is a formal portrayal and representation of a system, sorted out such that underpins thinking about the structures and practices of the System. A system architecture can include system segments, the extended systems created, that will cooperate in executing the general System. System architecture passes on the enlightening substance of the components involving a scenario, the connections among those components, and the principles administering those connections. The compositional parts and set of relationships between these parts that an architecture depiction may comprise of equipment, programming, documentation, offices, manual techniques, or jobs played by associations or, on the other hand, individuals.

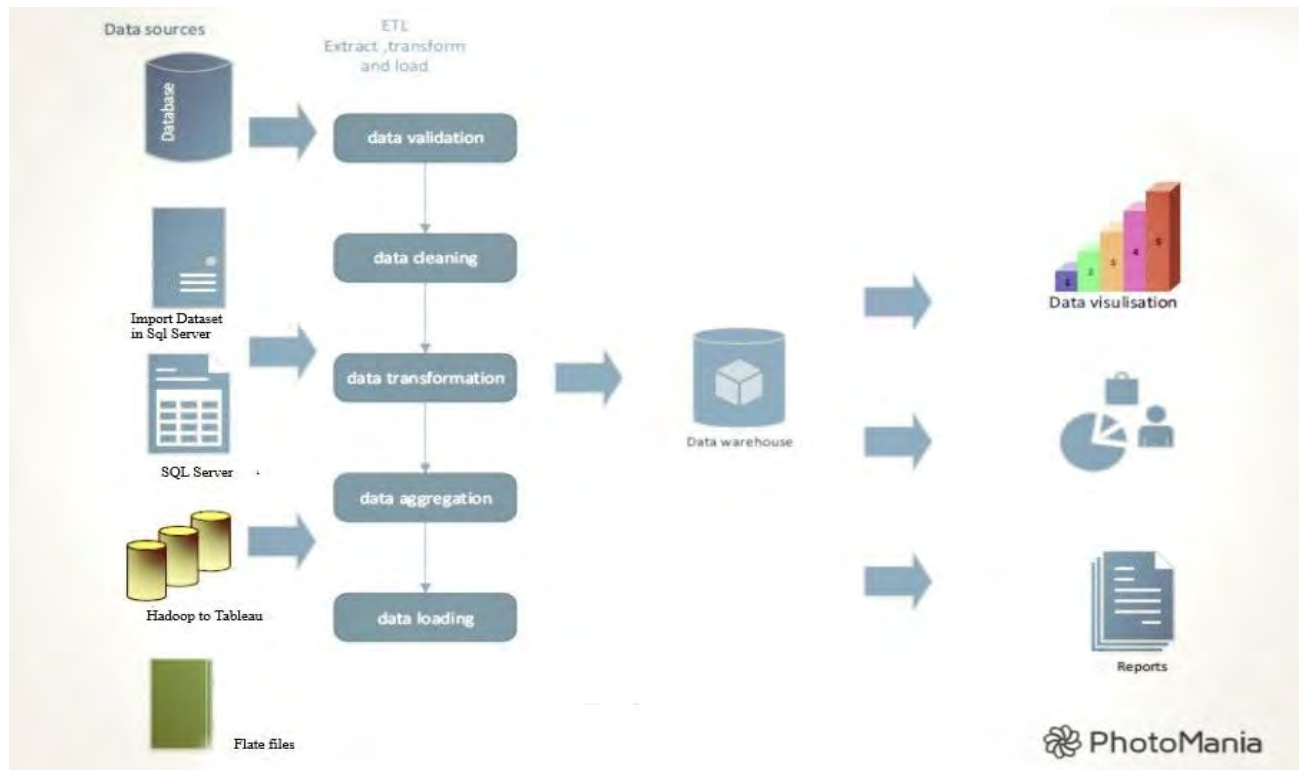


Fig 3.1 System architecture diagram

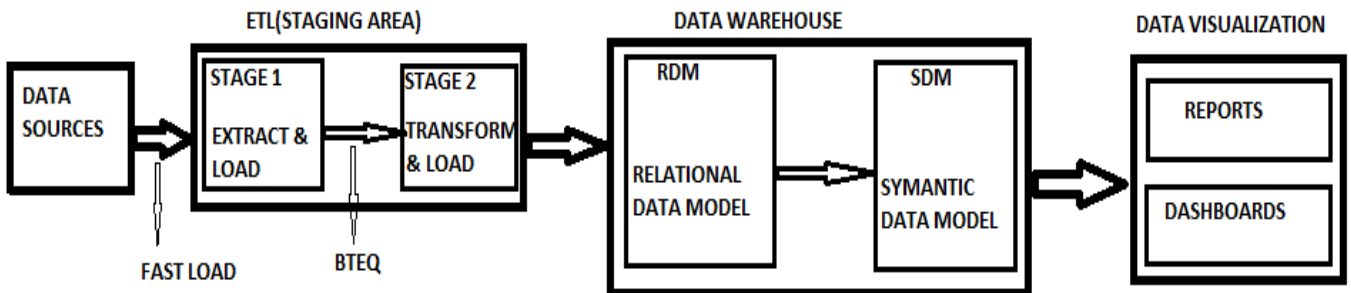


FIG 3.2 DETAILED VIEW OF SYSTEM ARCHITECTURE

### 3.5 Use Case Diagrams

Use case diagrams are used to demonstrate the different ways by which the user interacts with the System. It defines the interaction that takes place between a user of a system (an actor) and the System itself. It contains the actors and use cases and shows relations between them. Each use case

defines the functional requirements for a system. Use case diagrams are usually referred to as behavior diagrams used to describe a set of use cases that some system or systems should or can perform in collaboration with one or more external users of the System. Each use case should provide some observable and valuable results to the actors or other stakeholders of the System.

Use case diagrams are used to specify:

- External needs on a subject, mandatory usages of a system - to capture what a system under construction is meant to do
- The functionality offered by a subject – what the System can do
- The requirements, which the specified subject poses on its environment - by defining how the environment should interact with the subject so that it will be able to perform its services

### 3.5.1 Elements of the Use Case Diagrams

1. **Actor:** An actor is an entity that has a certain role to perform in a given system. It is, in a use case diagram interacts with a use case. So the actors of our proposed System are the data warehouse manager, company manager, and data analyst.
2. **Use case:** A use case, in a use case diagram, is a visual depiction of distinct data functionality in a system. A system must have or have an ability to perform such functionalities.
3. **System boundary:** we limit the System's functionalities with the system boundary. A system boundary defines what will be the scope of the System. A system cannot have infinite functionality as it defines the limits of the System.
4. **Relationships:** A relationship between two use cases is a dependency between the two use cases. The types of relationships are Include, Extend, Generalization, etc.

### 3.5.2 User Classes and Characteristics

1. **The data warehouse manager** uses the prediction to perform the database functions dynamically. It involves the extraction of data, correction of data, and integration of data and loads it to the data warehouse. So the data warehouse manager is also one of the users of this prediction.
2. **Data Analyst** is a key user and most important class for this prediction. He is using the System to perform multiple tasks. He is the person who is responsible for the run queries on data, loads it to the application, generates reports, and does analysis on the data by creating dashboards. a data analyst has access to the whole critical data of a company. Based on the geographic area, he can find which area contains more users of a particular telecom company. Based on the individual-based prediction, the company's internet packages are highly used in all telecom companies
3. **The company owner/manager** is also a very important user class of this product. Using DA he can know the In the age of data-driven solution, the customer demographic attributes, such as gender and age, play a core role that may enable companies to enhance the offers of their services and target the right customer at the right time and place. This work proposes a method that predicts users' gender and age based on how many adults and middle-age people use which telecommunication company sim. Or which company. He will find out which have company more users.



Fig 3.3 Use case diagram

### 3.6 Activity Diagram

Activity diagrams, which are related to programming flow plans (flowcharts), are used to illustrate activities. In the external view, we use activity diagrams for the description of those processes that describe the functionality of the data analysis system. Contrary to use case diagrams, in activity diagrams it is obvious whether actors can perform analysis use cases together or independently from one another. Activity diagrams allow you to think functionally. Purists of the object-oriented approach probably dislike this fact. We, on the other hand, regard this fact as a great advantage, since users of object-oriented methods, as well as users of functional thinking patterns, find a standard and familiar display format, which is a significant aid for data analysis process modeling.

The following are the notations used for activity diagrams:

- **Activity:** The rounded rectangles represent activities that occur.
- **Initial Node:** The filled-in circles is the starting point of the diagram.
- **Final Note:** The filled circle with a border is the ending point.
- **Fork:** A black bar with one flow going into it and several leaving it.
- **Join** A black bar with several flows entering it and one leaving it.
- **Decision:** A diamond with one flow entering and several leaving.
- **Merge** A diamond with several flows entering and one leaving.
- **Flow Final:** The circle with the X through it. This indicates that the process stops at this point.



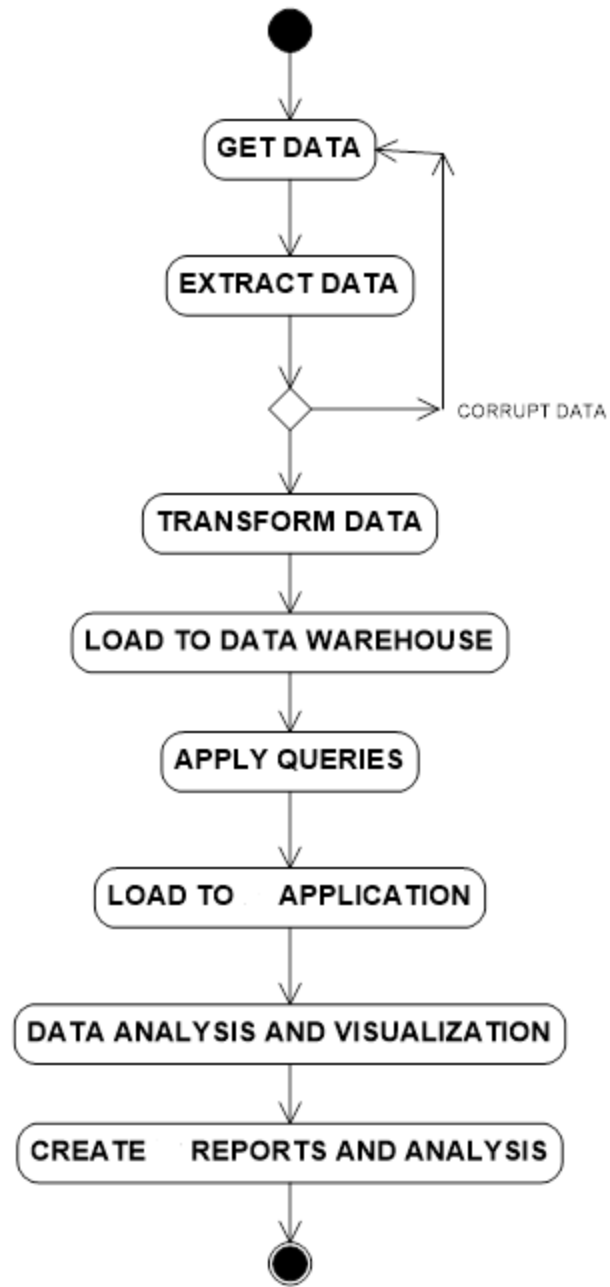


FIG 3.4 ACTIVITY DIAGRAM OF SYSTEM

### 3.7 Class Diagram

A class diagram in the Unified Modeling Language is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations, and the relationships among objects.

4 Different classes are used in this diagram and their operations which are used in previous diagrams.

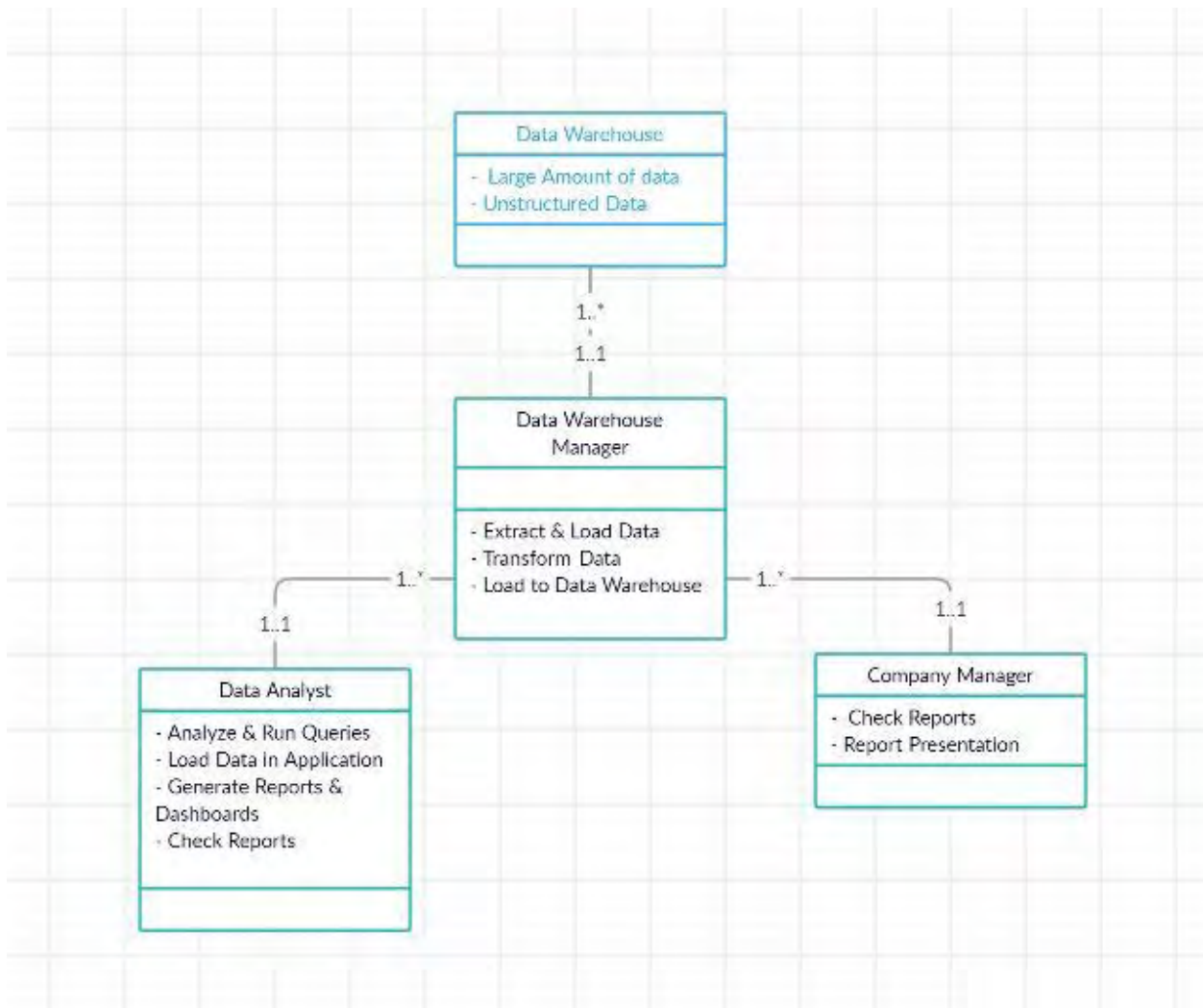


Fig.3.4 Class Diagram

### 3.8 ER Diagram

An entity-relationship model describes interrelated things of interest in a specific domain of knowledge. A basic ER model is composed of entity types and specifies relationships that can exist between entities.

#### Entities Used in this Diagram:

- Data Warehouse
- Data Warehouse Manager
- Data Analyst
- Company Manager

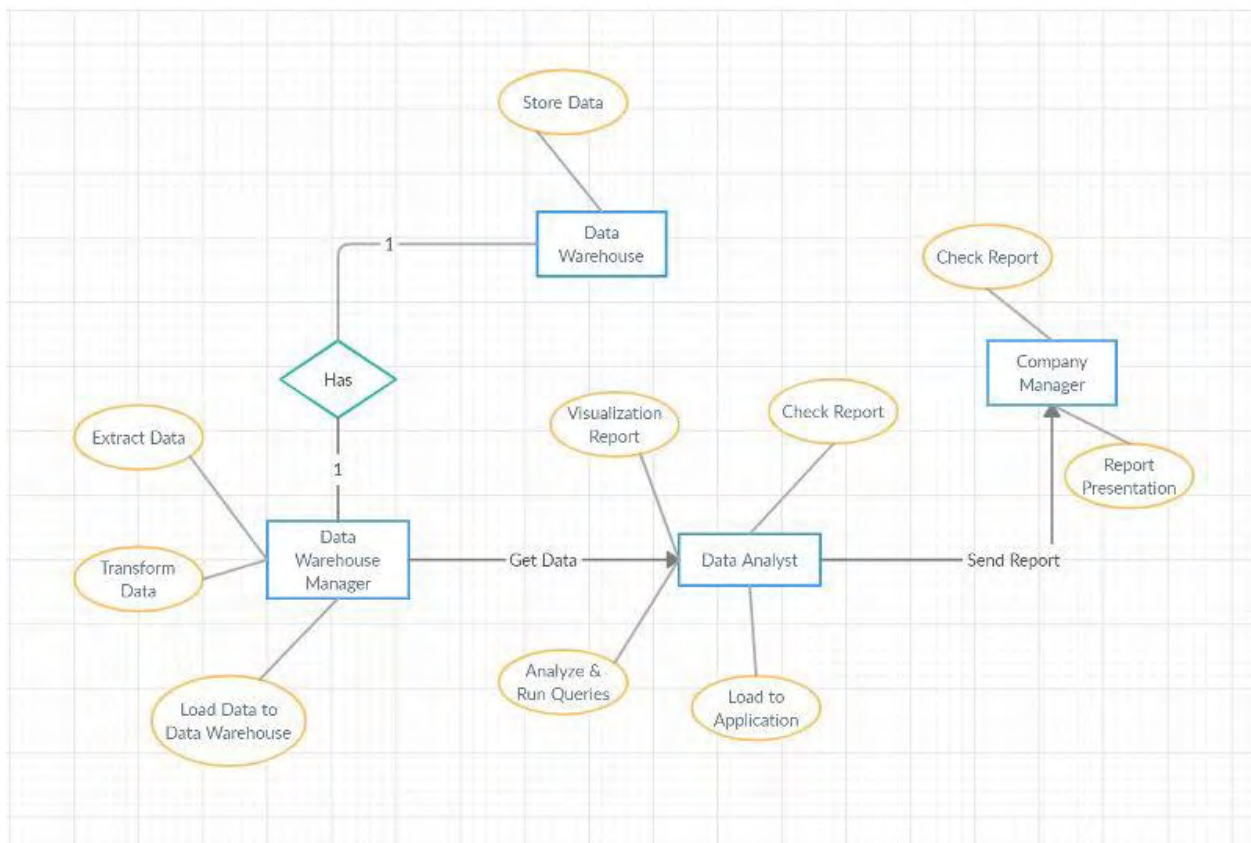


Fig.3.5 ER Diagram

### 3.9 Sequence Diagram

A sequence diagram shows object interactions arranged in time sequence. It depicts the objects and classes involved in the scenario and the sequence of messages exchanged between the objects needed to carry out the functionality of the scenario.

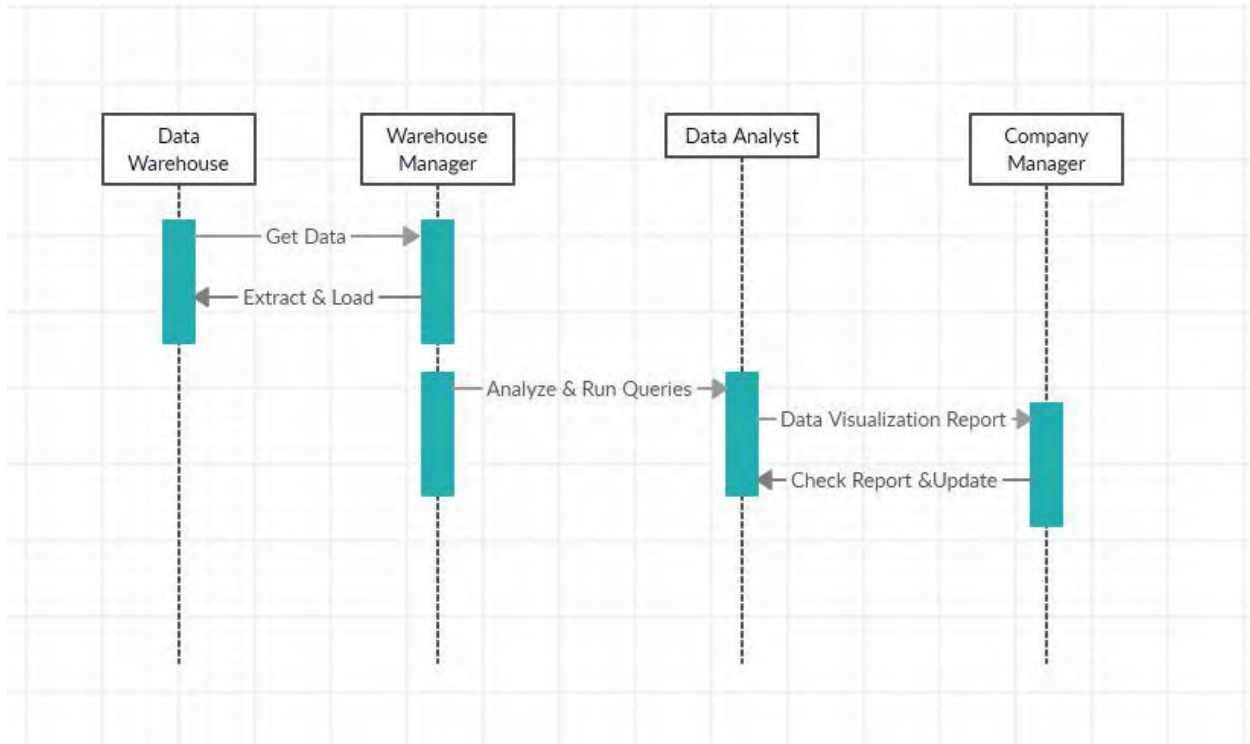


Fig.3.6 Sequence Diagram

# CHAPTER 4

## Tools and Languages

The following are the tools and technology which are being used for the development of the System.

- SQL Server Management Studio
- MS Visio
- Microsoft Word
- SQL
- Erwin Data Modeler
- Tableau

### 4.1 SQL Server Management Studio (SSMS)

**SQL Server Management Studio (SSMS)** is a software application first launched with Microsoft SQL Server 2005 that is used for configuring, managing, and administering all components within Microsoft SQL Server. It's the successor to the **Enterprise Manager** in SQL 2000 or before. The tool includes both script editors and graphical tools which work with objects and features of the server.

A central feature of SSMS is the Object Explorer, which allows the user to browse, select, and act upon any of the objects within the server. It also shipped a separate Express edition that could be freely downloaded, however recent versions of SSMS are fully capable of connecting to and manage any SQL Server Express instance.

The following functions can be performed:

- Create, Modify, and Drop Databases, Users, Roles, Profiles, and User-Defined Types.
- Create Tables
- Grant or Revoke access and system rights
- Copy Table, View or Macro definitions to another database, or another system
- Drop or Rename Tables, Views or Macros

- Move space from one database to another
- Run an SQL query
- Display information about a Database or Users
- Display information about a Table, View or Macro
- Set up the rules for Query and Access Logging

Teradata Administrator keeps a record of all the actions that are taken and can optionally save this record to a file. This record contains a timestamp together with the SQL that is executed, and other information such as the statement's success or failure.

#### **4.1.1 SSMS System Requirements**

SSMS supports the following 64-bit platforms :

##### **Supported Operating Systems:**

- Windows 10 (64-bit) version 1607 (10.0.14393) or later
- Windows 8.1 (64-bit)
- Windows Server 2019 (64-bit)
- Windows Server 2016 (64-bit)
- Windows Server 2012 R2 (64-bit)
- Windows Server 2012 (64-bit)
- Windows Server 2008 R2 (64-bit)

##### **Supported hardware:**

- 1.8 GHz or faster processor. Dual-core or better recommended
- 2 GB of RAM; 4 GB of RAM recommended (2.5 GB minimum if running on a virtual machine)
- Hard disk space: Minimum of 2 GB up to 10 GB of available space

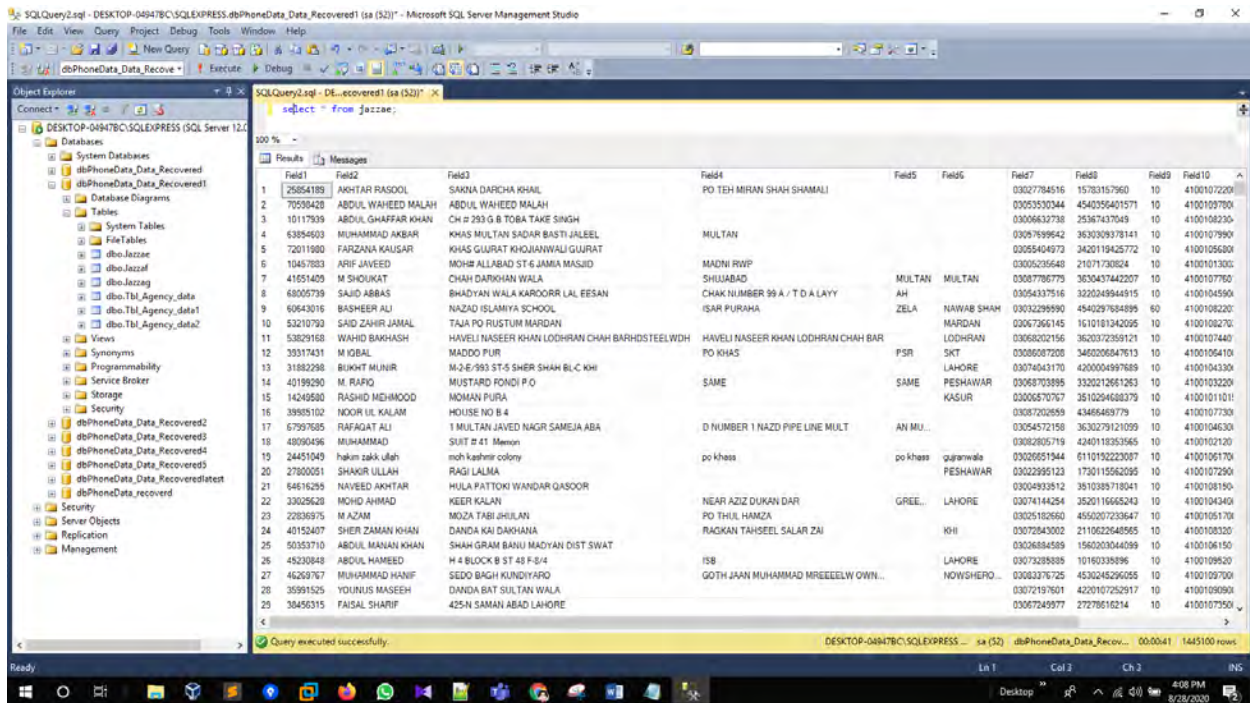


Fig 4.1 SQL server management studio

## 4.2 MS Visio

Visio is a Microsoft Office program that functions like a visualization tool to show data in an easily understandable way. Most often used for data that fits well with diagrams and charts, Visio takes standard images and allows flowcharts and decision diagrams to utilize those images to explain data. One way to think of Visio is that it's all about associating data with shapes on a diagram. You link data to these shapes and can apply graphic options including text, data bars and icons to display numbers in a visually comprehensible way. An awesome feature of Visio is that its data-linked diagrams are dynamic. This means that the data graphics will change in real-time when data is modified at the source (typically, the "source" is an Excel spreadsheet).

### 4.2.1 System requirements for MS Visio

- **Computer and processor:** 1 gigahertz (GHz) or faster x86-bit or x64-bit processor with SSE2 instruction set
- **Memory:** 2 GB RAM
- **Hard disk:** 3.0 GB available disk space

- **Display:** 1280 x 800 screen resolution

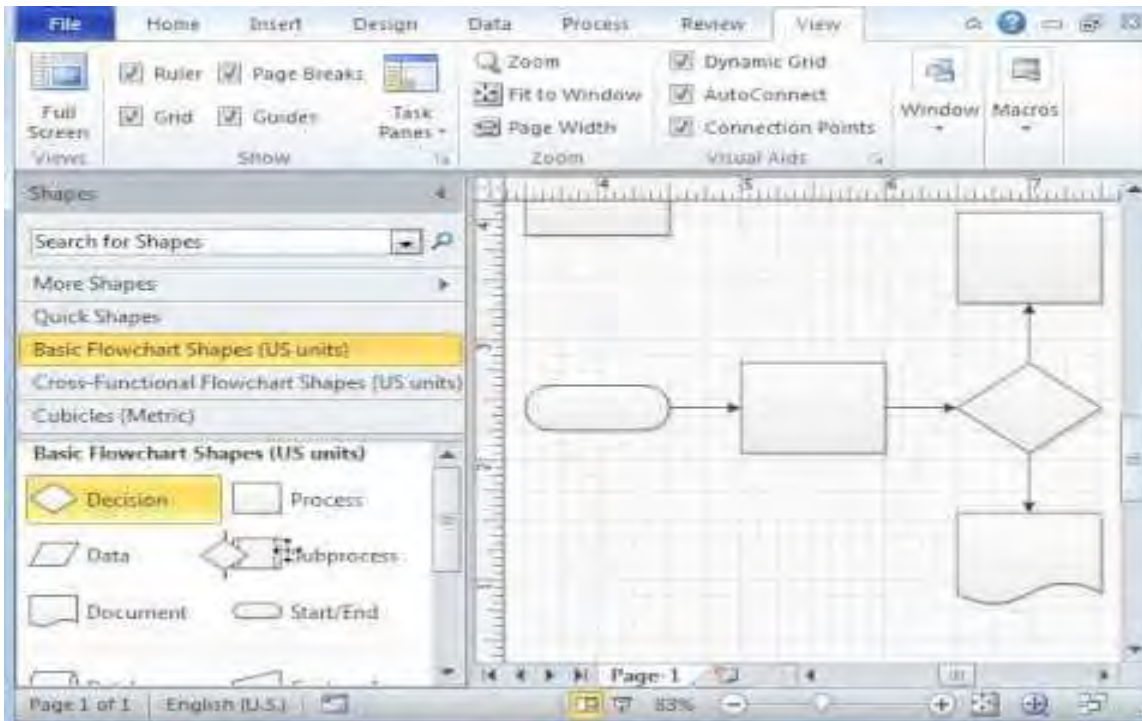


Fig 4.2 MS Visio Home Page

## 4.3 Microsoft Word

Microsoft Word is a word processor developed by Microsoft. Word contains rudimentary desktop publishing capabilities and is the most widely used word processing program on the market.

Microsoft Word offers several features to ease document creation and editing, including:

- **WYSIWYG (what-you-see-is-what-you-get) display:** It ensures that everything displayed on the screen appears the same way when printed or moved to another format or program.
- **Spell check:** Word features a built-in dictionary for spell checking; misspelled words are marked with a red squiggly underline. Sometimes, Word auto-corrects a misspelled word or phrase.
- **Text-level features** such as bold, underline, italic and strike-through



- Page-level features such as indentation, paragraphing and justification
- External support: Word is compatible with many other programs, the most common being the other members of the Office suite.

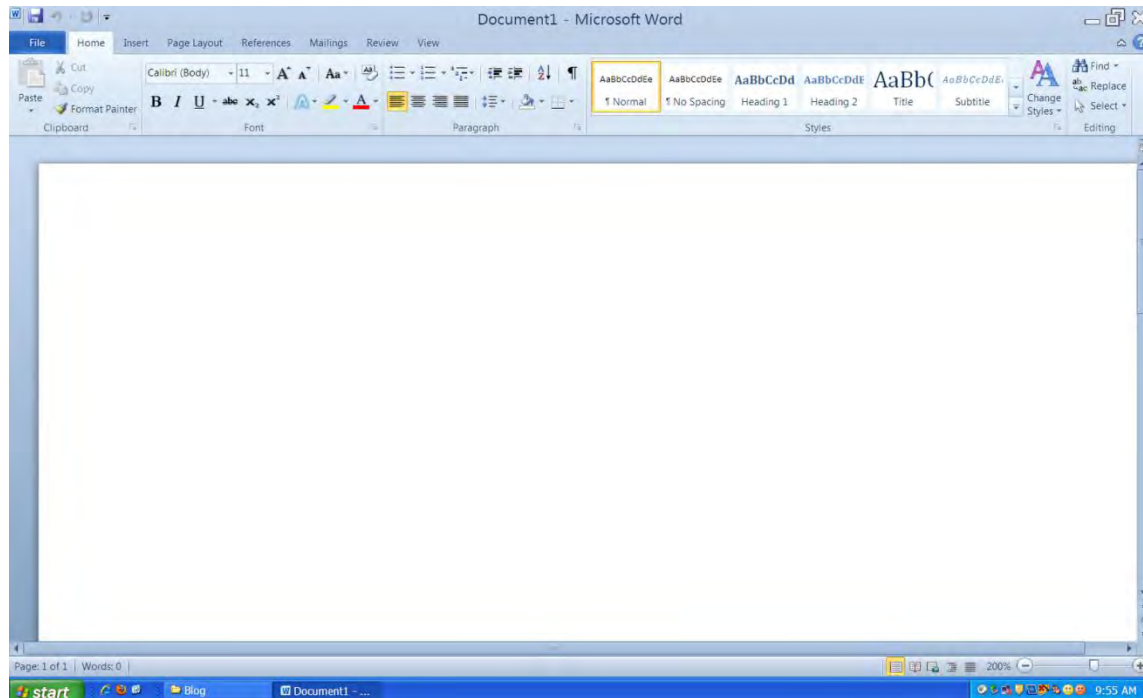
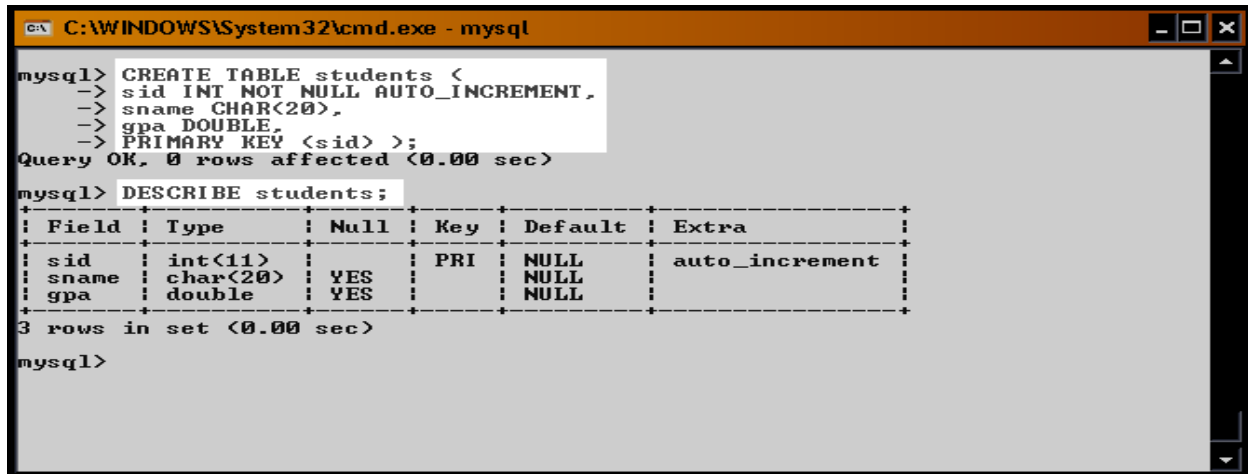


Fig 4.3 MS Word Home Page

All our project documentation is carried out in Microsoft Word.

## 4.4 SQL (Structured query language)

Structured Query Language is a domain-specific language used in programming and designed for managing data held in a relational database management system (RDBMS), or for stream processing in a relational data stream management system (RDSMS). SQL offers two main advantages: first, it introduced the concept of accessing many records with one single command, and second, it eliminates the need to specify how to reach a record, e.g., with or without an index. The scope of SQL includes data query, data manipulation (insert, update and delete), data definition (schema creation and modification), and data access control. Although SQL is often described as, and to a great extent is, a declarative language (4GL), it also includes procedural elements.



```
mysql> CREATE TABLE students (<br>-> sid INT NOT NULL AUTO_INCREMENT,<br>-> sname CHAR(20),<br>-> gpa DOUBLE,<br>-> PRIMARY KEY (sid) >;<br>Query OK, 0 rows affected (0.00 sec)<br>mysql> DESCRIBE students;<br>+-----+-----+-----+-----+-----+-----+<br>| Field | Type | Null | Key | Default | Extra |<br>+-----+-----+-----+-----+-----+-----+<br>| sid   | int(11) | YES | PRI | NULL | auto_increment |<br>| sname | char(20) | YES |     | NULL |                |<br>| gpa   | double | YES |     | NULL |                |<br>+-----+-----+-----+-----+-----+-----+<br>3 rows in set (0.00 sec)<br>mysql>
```

Fig 4.4 SQL Queries

4.5 VMware workstation pro

VMware Workstation is a hosted hypervisor that runs on x64 versions of Windows and Linux operating systems (an x86version of earlier releases was available) it enables users to set up virtual machines (VMs) on a single physical device, and use them simultaneously along with the actual machine. Each virtual machine can execute its operating System, including versions of Microsoft Windows, Linux, BSD, and MS-DOS. VMware Workstation is developed and sold by VMware, Inc., a division of Dell Technologies. There is a free-of-charge version, VMware Workstation Player, for non-commercial use. An operating systems license is needed to use proprietary ones such as Windows. Ready-made Linux VMs set up for different purposes are available from several sources VMware Workstation supports bridging existing host network adapters and sharing physical disk drives and USB devices with a virtual machine. It can simulate disk drives; an ISO image file can be mounted as a virtual optical disc drive, and virtual hard disk drives are implemented as .vmdk files. VMware Workstation Pro can save the state of a virtual machine (a "snapshot") at any instant. These snapshots can later be restored, effectively returning the virtual machine to the saved state, as it was and free from any post-snapshot damage to the VM. VMware Workstation includes the ability to group multiple virtual machines in an inventory folder. The machines in such a folder can then be powered on and powered off as a single object, useful for testing complex client-server environments.

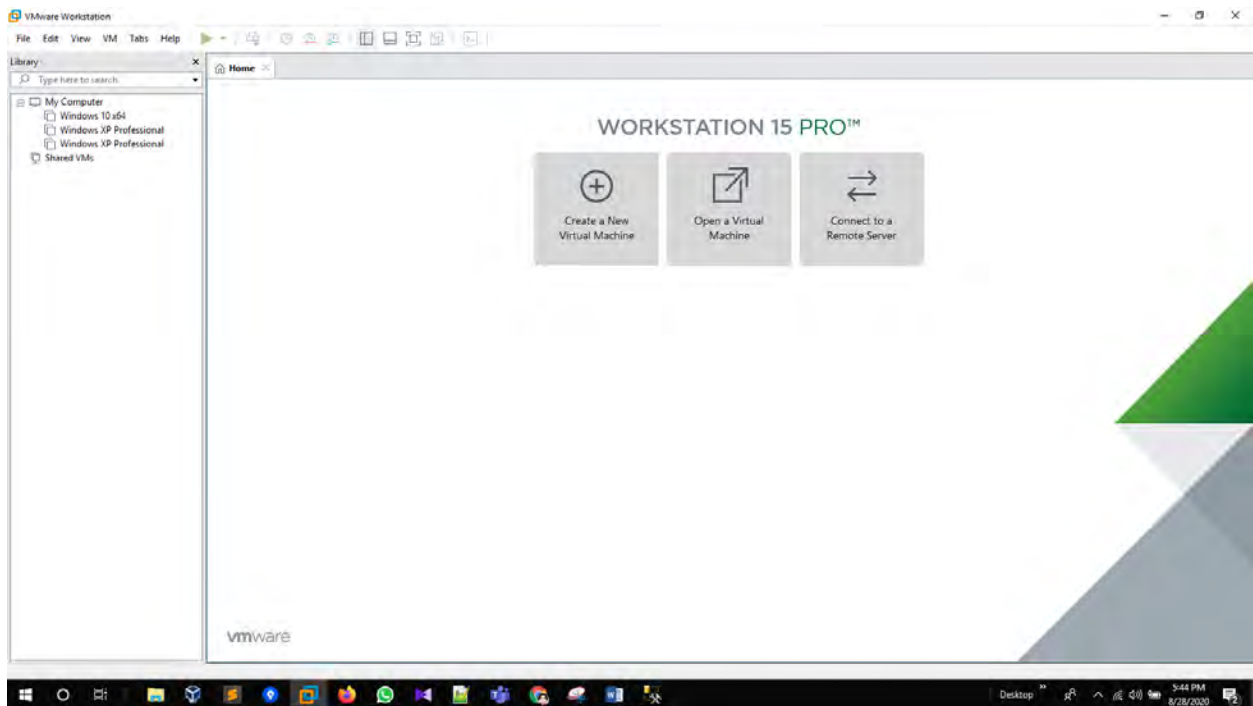


Fig 4.5 VMware Workstation Pro Home Page

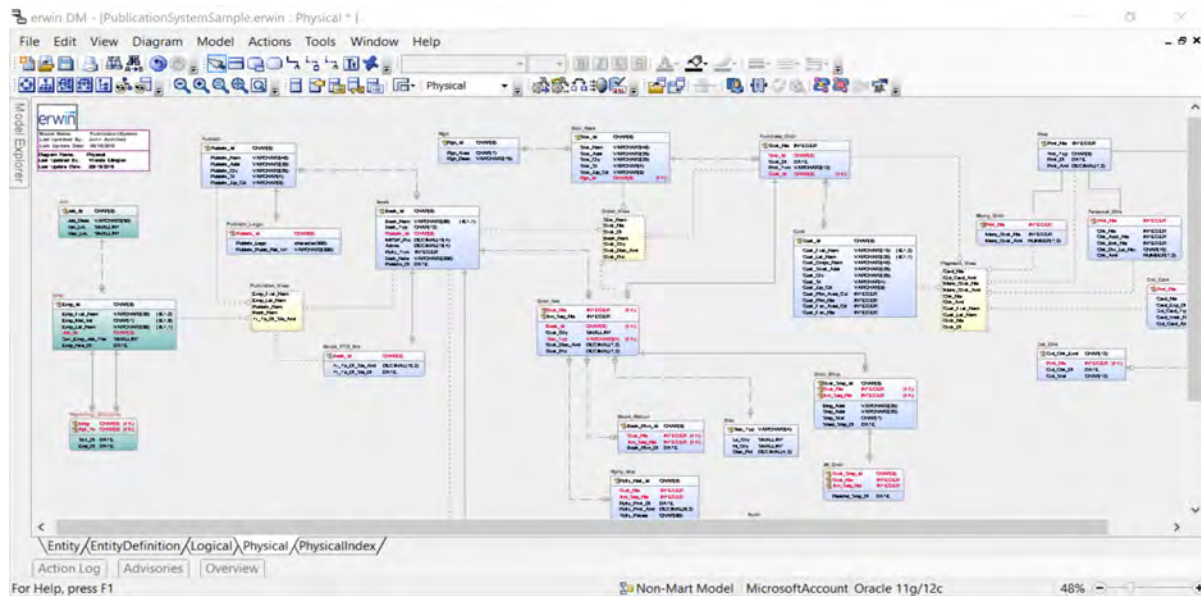
## 4.6 Erwin Data Modeler

Erwin Data Modeler is a computer software for data modeling. Initially developed by Logic Works, Erwin has since been acquired by a series of companies, before being spun-off by the private equity firm Parallax Capital Partners, which acquired and incorporated it as a separate entity, Erwin, Inc., managed by CEO Adam Famularo. A data model is a visual representation of data elements and the relationships between them. Data models help business, and technical resources collaborate in the design of information systems and the databases that power them. They show what data is required and how it needs to be structured to support various business processes. Data modeling is the process of creating a data model to communicate data requirements, documenting data structures, and entity types. It serves as a visual guide in designing and deploying databases with high-quality data sources as part of application development.

There are three basic types of data models, each with a specific purpose:

- **Conceptual Data Models:** High-level, static business structures and concepts

- **Logical Data Models:** Entity types, data attributes, and relationships between entities
- **Physical Data Models:** The internal schema database design.



**Fig 4.6 Erwin Data Modeler**

## 4.7 Tableau Software

**Tableau Software** is an interactive data visualization software company founded in January 2003 by Christian Chabot, Pat Hanrahan, and Chris Stolte, in Mountain View, California.

Chabot, Hanrahan, and Stolte were researchers at the Department of Computer Science at Stanford University who specialized in visualization techniques for exploring and analyzing relational databases and data cubes. The company was started as a commercial outlet for research produced at Stanford between 1999-2002.

Tableau products query relational databases, online analytical processing cubes, cloud databases, and spreadsheets to generate graph-type data visualizations. The products can also extract, store, and retrieve data from an in-memory data engine. Tableau has a mapping functionality and can plot latitude and longitude coordinates and connect to spatial files like Esri Shapefiles, KML, and GeoJSON to display custom geography. The built-in geocoding allows for administrative

places (country, state/province, county/district), postal codes, US Congressional Districts, US CBSA/MSA, Area Codes, Airports, and European Union statistical areas (NUTS codes) to be mapped automatically. You can group geographies to create custom territories or use custom geocoding to extend existing geographic roles in the product

### 4.7.1 Requirements for Tableau Desktop

The following list provides minimum requirements for running Tableau Desktop are:

- Microsoft Windows 7 or newer (32-bit and 64-bit)
- Microsoft Server 2008 R2 or newer
- Intel Pentium 4 or AMD Opteron processor or newer
- 2 GB memory
- 1.5 GB minimum free disk space
- 1366 x 768 screen resolution or higher

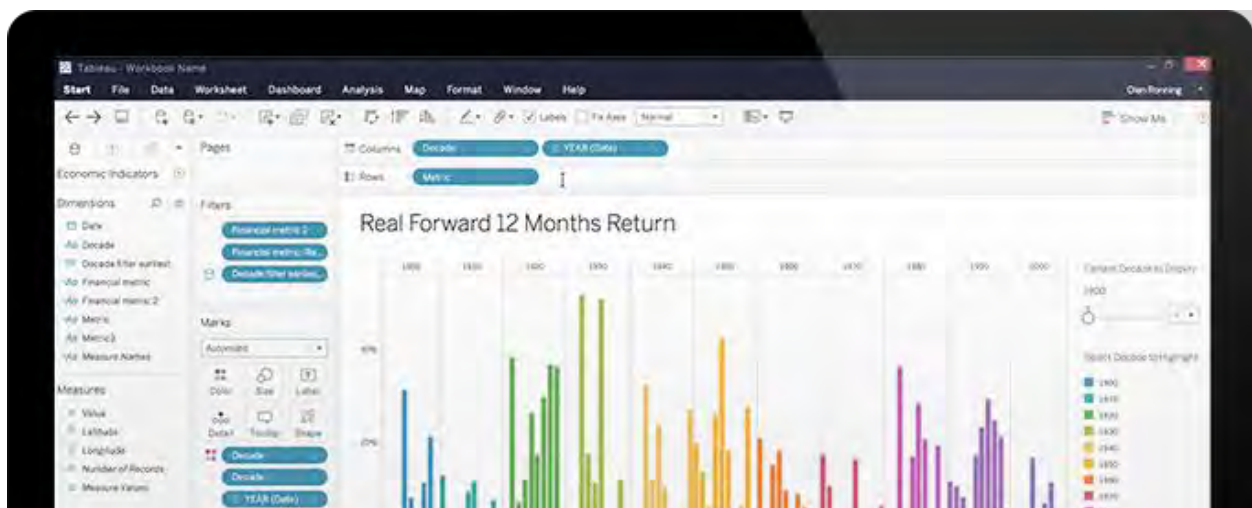


Fig 4.7 Tableau Desktop

# CHAPTER 5

## System Implementation

### 5.1 Implementation Phase

During the implementation phase, the project plan is put into motion, and the work of the project is performed, or we can say the automated system/application or other IT solution is moved from development status to production status. It is essential to maintain control and communicate as needed during implementation. Progress is continuously monitored, and appropriate adjustments are made and recorded as variances from the original plan. This is the logical conclusion, after evaluating, deciding, visioning, planning, applying for funds, and finding the financial resources of a project. The implementation stage is the construction period of an application. The process of implementation is dependent on the characteristics of the project and the IT solution, and thus may be synonymous with installation, deployment, rollout, or go-live. If necessary, data conversion phased implementation, and training for using, operating, and maintaining the System are accomplished during the Implementation Phase. From a system security perspective, the final design must be certified and accredited for use in the production environment during the Implementation Phase. The Implementation Phase ends with a formal decision to release the final IT solution into the Operations and Maintenance Phase. Most of the time is spent at this stage.

### 5.2 Objectives of Implementation Phase

The objectives of the implementation phase can be summarized as follow:

- Putting the action plan into operation
- Achieving tangible change and improvements
- Ensuring that new infrastructure, new institutions, and new resources are sustainable in every aspect
- Ensuring that any unforeseen conflicts that might arise during this stage are resolved
- Ensuring transparency about finances
- Ensuring that potential benefits are not captured by elites at the expenses of lower social groups

## 5.3 Project Objectives and Goals

Following are primary objectives of our project

- Getting Telecommunication data
- Making a data warehouse (as telecom companies generate a large amount of data)
- Loading the data into the data warehouse
- Analysis of the data
- Data visualization, Reports, and dashboards

## 5.4 Getting the data

Companies are doing analytics to make decisions based on reports, dashboards, or charts (Visualization) generated by different tools. These reports are generated from the data provided to these tools. So, getting data is the most crucial concern in these projects. Data can have different sources; it may be coming from operational databases, servers, storage groups, flat files, excel sheets, etc. For our project, we have telecom companies' data. This data is taken from a source with some restrictions on it. Data is in CSV and fixed-width files format.

Our project data contains:

- Customer information containing customer Name, CNIC, Location
- Mobile Number, Date of Activation, Status, Status Date
- Other Network Number if any, IMSI Number

We are going to carry out an analysis of this data. This will consist of dashboards and data visualizations which will enable them to view the details of data graphically and diagrammatically and make better decisions.



serialnumber	MOB_NO	CNIC	NAME	ACTIVATION_DATE	ADDRESS	Service_Provider	GENDER
1	3214004163	352011453343-3	ANWAR SAJJAD CHATTA	May 24 2005 9:34AM	SARFARAZ RAFIQUE ROAD CANTT 52 F.....Lahore.....Punjab	Wand	Male
2	3214004171	354032745551-9	MUHAMMAD SHAHID GHANI	May 24 2005 9:45AM	GASI M PURA COLLEGE ROAD 45.04.....Sargodha.....	Wand	Male
3	3214004123	352001453415-7	ABDUL JABBAR KHAN	May 24 2005 10:24AM	WARID TELECOM BUSINESS CENTER 18 SPENCER BUILDIN.....	Wand	Male
4	3214004124	352024720935-1	SYED SAJJAD HAIDER SHAMSEE	May 24 2005 1:20PM	BLOCK 3, SECTOR 1C TOWN SHIP LAHORE,114.....Lahore.....	Wand	Male
5	3214004126	352029031216-5	IMTIYAZ ALI SHAH	Jun 2 2005 9:09PM	WARID TELECOM, EFU HOUSE 9TH FLOOR, JAIL ROAD LAH.....	Wand	Male
6	3214004127	352025052793-5	MUHAMMAD JAVED AFRIDI	Jun 2 2005 8:29PM	G-8LOCK SABAZAR SCHEME MULTAN ROAD 219.....Lah.....	Wand	Male
7	3214004138	27272453919	SHEKH ZULFIQAR ALI	Jun 2 2005 8:59PM	2-A, Mallan Road, Lahore.....Lahore.....Punjab	Wand	Male
8	3214004129	352022979184-5	MUHAMMAD MUZAFFAR	May 26 2005 11:22AM	BOULEVARD GULBERG 2, 4TH FLOOR NEW AURIGA CENTR.....	Wand	Male
9	3212451936	422012947708-7	SAMUEL ZIA DEAN	Sec 8 2005 11:02PM	306,HOUSE 24/7, DRIGH ROAD CANTT. BAZAR.....Korachi.....	Wand	Male
10	3214004134	372011777101-7	CHAUHURY HAQ NAHAZ	May 24 2005 9:05AM	SARWAR ROAD LAHORE CANTT 23,ASKARI VILLA.....Lahore.....	Wand	Male
11	3214004089	352023997436-5	SYED MUSHA RAZA	May 23 2005 7:17AM	.....Lahore.....Punjab	Wand	Male
12	3214004111	352022910417-9	MIRZA SHUAIB ALI	May 23 2005 3:35AM	FIASAL TOWN 575-C.....Lahore.....Punjab	Wand	Male
13	3214001001	352011685655-7	SHAHID SAEED	May 2 2005 9:25AM	SECTOR U - DHA H # 25 ST # 1.....Lahore 540001.....Punjab	Wand	Male
14	3214001007	422010767464-7	SALMAN KHURRAM	May 1 2005 6:00PM	A-45, BLOCK 03, GULSHAN-E-GHAI, KARACHI.....Karachi.....	Wand	Male
15	3214004009	352022938223-7	SYED M RAZA	May 24 2005 3:06AM	JT/15 AUDIT 3RD FLOOR MCB HOUSE 15-MAIN JAIL ROAD GU.....	Wand	Male
16	3214004006	354041624080-5	MUHAMMAD AMIN UL HASSAN	May 31 2005 7:36AM	GAZNI ROAD SAMIA MSJID LAHORE.....Lahore.....Punjab	Wand	Male
17	3214004009	352024564401-8	NAILA AZZA	May 25 2005 10:03AM	HOUSE NUMBER 486, BLOCK ONE FEDERAL B AREA KARAC.....	Wand	Female
18	3214004010	352011585201-7	AMIR ANWAR KHAN	May 24 2005 10:28AM	HOUSE NUMBER 486, BLOCK ONE FEDERAL B AREA KARAC.....	Wand	Male
19	3214004025	11989	ZAHID MANZOOR	May 24 2005 3:33AM	M D H A 248.....Lahore.....Punjab	Wand	Male
20	3214004038	352011662133-6	ZUBAIR UD DIN	May 24 2005 8:53AM	MUHAMMADABAD MUGALPURA 41 A 12.....Lahore.....Punjab	Wand	Male
21	3452786787	421011679035-5	Hafiz Rehman	Jun 26 2005 9:13AM	JWA-955, SECT11-A, NORTH KARACHI, NEAR SALEEM CENTR.....	Wand	Male
22	3214004049	352011646213-9	MIRAN ALI	May 24 2005 9:08AM	A-NEAR LESCO OFFICE MAN AMRUDIN PARK 17.01.....La.....	Wand	Male
23	3214004055	341012523862-9	AMIR ZIA	May 24 2005 7:55AM	TOTTIYAN WALA MEHAR BOTTA.....Sargodha.....	Wand	Male
24	3214004056	352017745660-3	RAIHAN MEHMOOD	May 23 2005 6:23AM	NEW GIBAL PARK LAHORE CANTT LAHORE, E 430.....Lah.....	Wand	Male
25	3214004057	352032471626-1	SAMI AHMAD	May 24 2005 9:20AM	Wand Office Lahore.....Lahore.....Punjab	Wand	Male
26	3214004098	352021344959-9	SYED FARUKH SAJJAD	May 23 2005 10:33AM	7,SARFRAZ NAWAZ STREET MODEL TOWN LINK ROAD LAH.....	Wand	Male
27	3214004074	8825	NADEEM MAHMOOD	Jun 4 2005 10:25AM	THE MALL BEADON RD.05.# 8.....Lahore.....Punjab	Wand	Male
28	3214004076	352022196333-9	SAJJAD AHMED BUTT	May 23 2005 8:24AM	GULBERG III 2 FLOOR EMPIRE CENTER.....Lahore.....Punjab	Wand	Male
29	3214006600	611012020325-1	USMAN MASOOD NOOR	May 24 2005 4:00AM	HOUSE NUMBER 486, BLOCK ONE FEDERAL B AREA KARAC.....	Wand	Male

Fig 5.1 Data Sheet

## 5.5 Data Warehouse

A data warehouse is a central repository of information that can be analyzed to make better-informed decisions. Data flows into a data warehouse from transactional systems, relational databases, and other sources, typically on a regular cadence. Business analysts, data scientists, and decision-makers access the data through tools, SQL clients, and other analytics applications.

Businesses use reports, dashboards, and analytics tools to extract insights from their data, monitor, and support decision making. These reports, dashboards, and analytics tools are powered by data warehouses, which store data efficiently to minimize I/O and deliver query results at blazing speeds to hundreds and thousands of users concurrently.

A data warehouse is

- Subject oriented
- Integrated
- Time-varying
- Nonvolatile

The following name also knows the data warehouse system:



- Decision Support System (DSS)
- Executive Information System
- Management Information System
- Business Intelligence Solution
- Analytic Application
- Data Warehouse

### 5.5.1 Data Warehouse Architecture

A data warehouse architecture consists of three tiers.

- **Bottom tier:** The bottom level of the architecture is the database server, where data is loaded and stored.
- **Middle tier:** The intermediate level consists of the analytics engine that is used to access and analyze the data.
- **Top tier:** The top level is the front end client that presents results through reporting, analysis, and business intelligence tools.

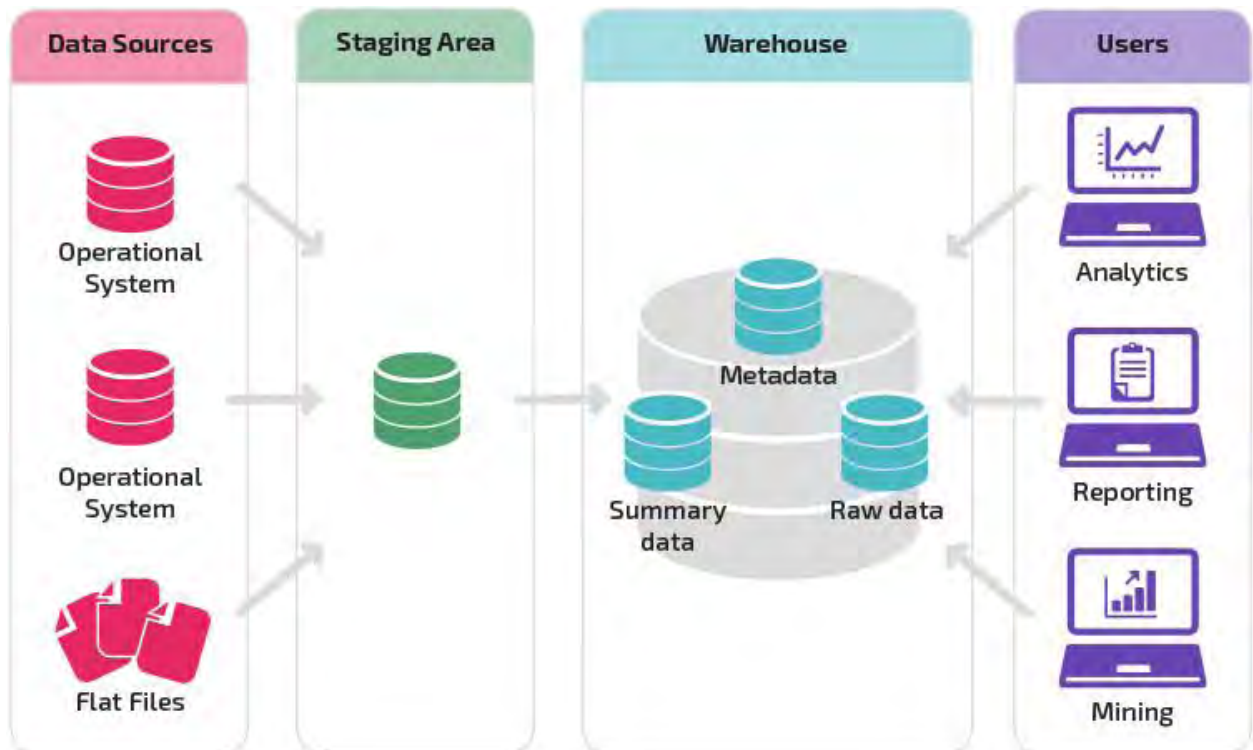


Fig 5.2 Data warehouse architecture

### **5.5.2 How does a Data Warehouse Work**

A data warehouse works by organizing data into a schema that describes the layout and type of data, such as integer, data field, or string. When data is ingested, it is stored in various tables described by the schema. Query tools use the schema to determine which data tables to access and analyze.

Data may be:

- Structured
- Semi-structured
- Unstructured

The data is processed, transformed, and ingested so that users can access the processed data in the Data Warehouse through tools, SQL clients, and spreadsheets. A data warehouse merges information coming from different sources into one comprehensive database.

By merging all of this information in one place, an organization can analyze its customers more holistically. This helps to ensure that it has considered all the information available. Data warehousing makes data mining possible. Data mining is looking for patterns in the data that may lead to higher sales and profits.

### **5.5.3 Benefits of Data Warehouse**

The benefits of a data warehouse are:

- Better decision making
- Consolidates data from many sources
- Data quality, consistency, and accuracy
- Historical intelligence
- Separates analytics processing from transactional databases, improving the performance of both systems

## **5.6 ETL**

(Extract, Transform and Load) is a process in data warehousing responsible for pulling data out of the source systems and placing it into a data warehouse. During this process, data is taken (extracted) from a source system, converted (transformed) into a format that can be analyzed, and stored (loaded) into a data warehouse, or another system ETL involves the following tasks:

### **5.6.1 Extracting the data**

From source systems ,data from different source systems is converted into one consolidated data warehouse format, which is ready for a transformation.

### **5.6.2 Transforming the data**

It may involve the following tasks:

- applying business rules (so-called derivations, e.g., calculating new measures and dimensions),
- Cleaning (e.g., mapping NULL to 0 or "Male" to "M" and "Female" to "F" etc.),
- Filtering (e.g., selecting only specific columns to load),
- Splitting a column into multiple columns and vice versa,
- Joining together data from various sources (e.g., lookup, merge),
- Transposing rows and columns,
- Applying any simple or complex data validation (e.g., if the first three columns in a row are empty then reject the row from processing)

### **5.6.3 Loading the data.**

Load the data into a data warehouse or data repository.

### 5.6.4 How ETL Works

Data from one or more sources is extracted and then copied to the data warehouse. When dealing with large volumes of data and multiple source systems, the data is consolidated. ETL is used to migrate data from one database to another, and is often the specific process required to load data to and from data marts and data warehouses, but is a process that is also used to massive convert (transform) databases from one format or type to another



Fig 5.3 ETL

## 5.7 Data Analysis

Data analysis is a primary component of data mining and is key to gaining the insight that drives business decisions. Organizations and enterprises analyze data from a multitude of sources using Data management solutions and customer experience management solutions that utilize data analysis to transform data into actionable insights. This is a process of inspecting, cleansing, transforming, and modeling data to discover useful information, informing conclusions, and supporting decision-making. Data analysis is a process for obtaining raw data and converting it into valuable information for decision-making by users. Data is collected and analyzed to answer questions, test hypotheses or disprove theories

Data analysis involves asking questions about what happened, what is happening, Data analysis is a proven way for organizations and enterprises to gain the information they need to make better decisions, serve their customers, and increase productivity and revenue. The benefits of data analysis are almost too numerous to count, and some of the most rewarding benefits include getting

the right information for your business, getting more value out of IT departments, creating more effective marketing campaigns, gaining a better understanding of customers, and so on.

But, there is so much data available today that data analysis is a challenge. Namely, handling and presenting all of the data are two of the most challenging aspects of data analysis. Traditional architectures and infrastructures are not able to handle the sheer amount of data that is being generated today, and decision-makers find it takes longer than anticipated to get actionable insight from the data.

Fortunately, data management solutions and customer experience management solutions give enterprises the ability to listen to customer interactions, learn from behavior and contextual information, create more effective actionable insights, and execute more intelligently on insights to optimize and engage targets and improve business practices.

## **5.8 Data Visualization**

Data visualization is a general term that describes any effort to help people understand the significance of data by placing it in a visual context. Patterns, trends, and correlations that might go undetected in text-based data can be exposed and recognized more comfortably with data visualization software. Data visualization refers to the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization is an accessible way to see and understand trends, outliers, and patterns in data. It enables decision-makers to see analytics presented visually, so they can grasp difficult concepts or identify new patterns. With interactive visualization, you can take the idea a step further by using technology to drill down into charts and graphs for more detail, interactively changing what data you see and how it's processed. Today's data visualization tools go beyond the standard charts and graphs used in Microsoft Excel spreadsheets, displaying data in more sophisticated ways such as infographics, dials and gauges, geographic maps, sparklines, heat maps, and detailed bar, pie, and fever charts. The images may include interactive capabilities, enabling users to manipulate them or drill into the data for querying and analysis. Indicators designed to alert users when data has been updated, or predefined conditions occur can also be included.

### **5.8.1 How Data Visualization Works**

Most of today's data visualization tools come with connectors to popular data sources, including the most common relational databases, data warehouses, and a variety of cloud storage platforms. The visualization software pulls in data from these sources and applies a graphic type to the data. Data visualization software allows the user to select the best way of presenting the data, but, increasingly, the software automates this step. Some tools automatically interpret the shape of the data and detect correlations between individual variables and then place these discoveries into the chart type that the software determines is optimal.

Typically, data visualization software has a dashboard component that allows users to pull multiple visualizations of analyses into a single interface, generally a web portal.

### **5.8.2 Benefits of Data Visualization**

Our eyes are drawn to colors and patterns. We can quickly identify red from blue, square from a circle. Our culture is visual, including everything from art and advertisements to TV and movies. Data visualization is another form of visual art that grabs our interest and keeps our eyes on the message. When we see a chart, we quickly see trends and outliers. If we can see something, we internalize it quickly. It's storytelling with a purpose. If you've ever stared at a massive spreadsheet of data and couldn't see a trend, you know how much more effective a visualization can be.

#### **Common general types of data visualization:**

- Charts
- Tables
- Graphs
- Maps
- Infographics
- Dashboards

- **More specific examples of methods to visualize data:**

- Area Chart
- Bar Chart
- Box-and-whisker Plots
- Bubble Cloud
- Bullet Graph
- Cartogram
- Circle View
- Dot Distribution Map
- Gantt Chart
- Heat Map
- Highlight Table
- Histogram
- Matrix
- Network
- Polar Area

- Radial Tree
- Scatter Plot (2D or 3D)
- Stream graph
- Text Tables
- Timeline
- Tree map
- Wedge Stack Graph
- Word Cloud
- And any mix-and-match combination in a dashboard!

### 5.8.3 Tool Used for Visualization

**Tableau Desktop** is a data visualization software that is used for data science and business intelligence. Tableau can create a wide range of different visualization to present the data and showcase insights interactively. It comes with tools that allow us to drill down data and see the impact in a visual format that can be easily understood by any individual. Tableau also comes with real-time data analytics capabilities and cloud support.



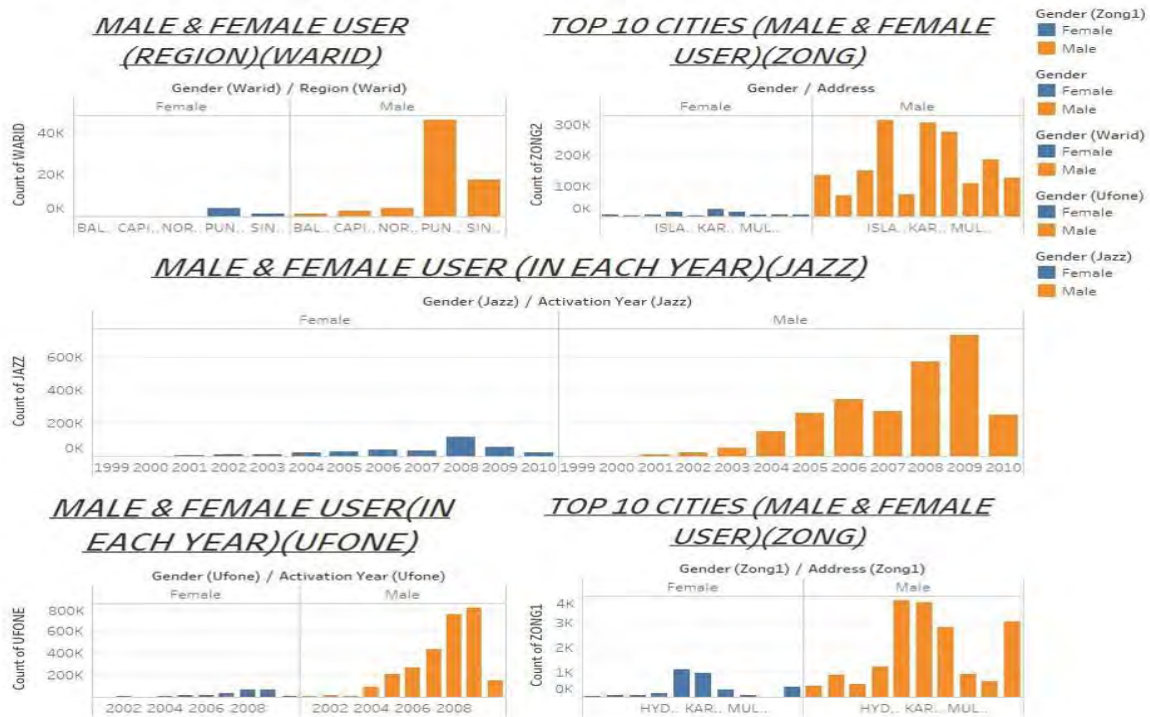


Fig 5.4 Tableau Dashboard

# CHAPTER 6

## Testing

Testing of Data Warehouse applications is a little different than testing traditional transactional applications as it requires a data-centric testing approach.

### 6.1 The Importance of Data Warehouse Testing

With data driving critical business decisions, testing the data warehouse data integration process is essential. Data comes from numerous sources. The data source affects data quality, so data profiling and data cleaning must be ongoing. Source data history, business rules, or audit information may no longer be available.

Additionally, in the ETL process, data flows through a pipeline before reaching the data warehouse. You must test the entire ETL pipeline to ensure each type of data is transformed or copied as expected. Most importantly, the data warehouse is a strategic enterprise resource. Testing is required.

### 6.2 ETL Testing

ETL stands for Extract-Transform-Load, and it is a process of how data is loaded from the source system to the data warehouse. ETL testing is done to ensure that the data that has been loaded from a source to the destination after business transformation is accurate. It also involves the verification of data at various middle stages that are being used between source and destination. ETL stands for Extract-Transform-Load.

#### 6.2.1 ETL Testing Process

ETL testing includes multiple phases, and testing should be executed throughout the lifecycle of the data warehouse implementation, not just at the end. Similar to other Testing Process, ETL also goes through different phases. The different phases of the ETL testing process are as follows

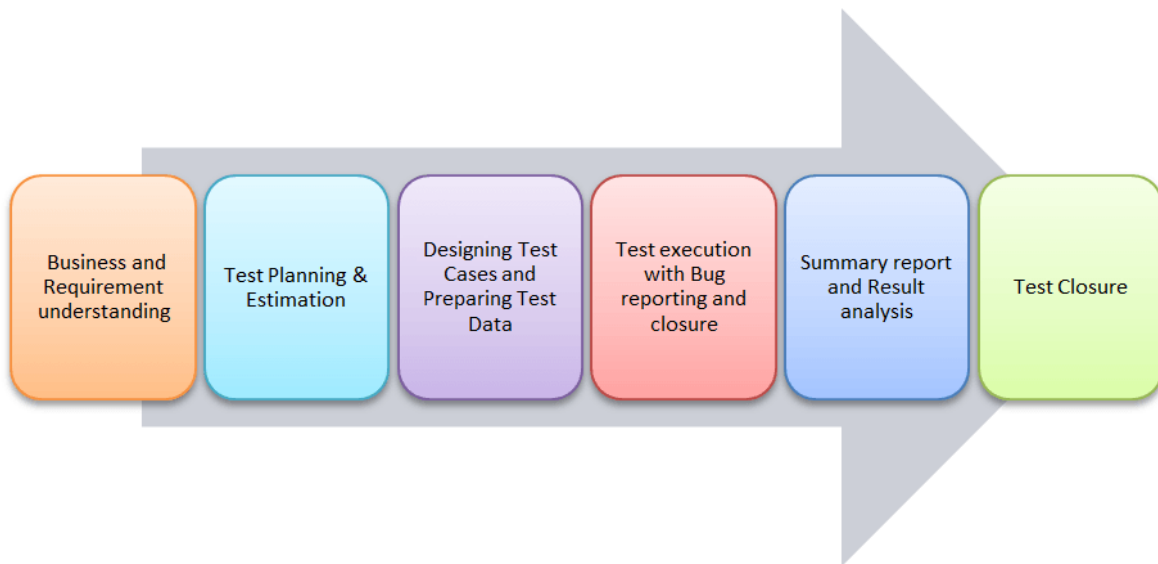


Fig 6.1 ETL Testing Stages

**ETL testing is performed in five stages**

1. Identifying data sources and requirements
2. Data acquisition
3. Implement business logic and dimensional Modelling
4. Build and populate data
5. Build Reports

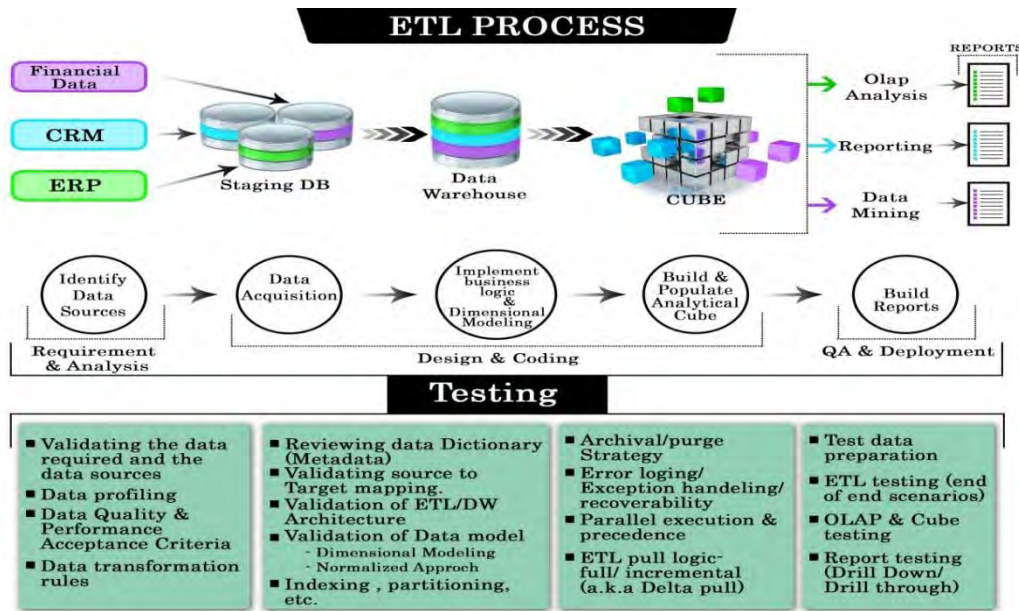


Fig 6.2 ETL Testing

## 6.2.2 Types of ETL Testing

- **Metadata Testing:**

Metadata testing confirms that the table definitions conform to the data model and application design specifications. This test should include data type check, data length check, and index/constraint check.

- **Data Completeness Testing:**

Data Completeness testing validates that all the expected source data has been successfully loaded to the target. Tests include: Compare and Validate counts, aggregates (min, max, sum, avg), and actual data between the source and destination.

- **Data Quality Testing:**

Data Quality tests validate the accuracy of the data. Data profiling is used to identify data quality issues, and the ETL is designed to fix or handle these issues. Automating the data quality checks between the source and target system can help to mitigate problems post-implementation.

- **Data Transformation Testing:**

Data Transformation comes in two flavors: white box testing and black-box testing. White box data transformation testing examines the program structure and develops test data from the program logic/code. Testers review the transformation logic from the mapping design document and the ETL code to create test cases. Black-box testing examines the functionality of an application without looking at internal structures for transformation testing; this involves reviewing the transformation logic from the mapping design document creating the appropriate test data.

- **ETL Regression Testing:**

ETL Regression testing validates that the ETL produces the same output for a specific input before and after the change. Incremental ETL testing verifies that updates on the sources are getting loaded into the target system correctly.

- **ETL Integration Testing:**

ETL integration testing is end-to-end testing of the data in the ETL process and the target application.

### 6.2.3 ETL Test Scenarios and Test Cases

Test Scenario	Test Cases
Mapping doc validation	Verify mapping doc whether corresponding ETL information is provided or not. The changelog should maintain in every mapping doc.
Validation	1. Validate the source and target table structure against the corresponding mapping doc.

	<ol style="list-style-type: none"> <li>2. Source data type and target data type should be the same</li> <li>3. Length of data types in both source and target should be equal</li> <li>4. Verify that data field types and formats are specified</li> <li>5. Source data type length should not less than the target data type length</li> <li>6. Validate the name of columns in the table against the mapping doc.</li> </ol>
Constraint Validation	Ensure the constraints are defined for a specific table as expected
Data consistency issues	<ol style="list-style-type: none"> <li>1. The data type and length for a particular attribute may vary in files or tables through the semantic definition is the same.</li> <li>2. Misuse of integrity constraints</li> </ol>
Completeness Issues	<ol style="list-style-type: none"> <li>1. Ensure that all expected data is loaded into the target table.</li> <li>2. Compare record counts between source and target.</li> <li>3. Check for any rejected records</li> <li>4. Check data should not be truncated in the column of target tables</li> <li>5. Check boundary value analysis</li> </ol>

	6. Compares unique values of critical fields between data loaded to WH and source data
Correctness Issues	<ol style="list-style-type: none"> <li>1. Data that is mi" spelled or" inaccurately recorded</li> <li>2. Null, non-unique or out of range data</li> </ol>
Transformation	Transformation
Data Quality	<ol style="list-style-type: none"> <li>1. Number check: Need to number check and validate it</li> <li>2. Date Check: They have to follow date format, and it should be same across all records</li> <li>3. Precision Check</li> <li>4. Data check</li> <li>5. Null check</li> </ol>
Null Validate	Verify the null values, where "Not Null" is specified for a specific column.
Duplicate Check	<ol style="list-style-type: none"> <li>1. Needs to validate the unique key, primary key and any other column should be unique as per the business requirements are having any duplicate rows</li> <li>2. Check if any duplicate values exist in any column which is extracting from multiple columns in source and combining into one column</li> </ol>

	<ol style="list-style-type: none"> <li>3. As per the client requirements, needs to ensure that no duplicates in a combination of multiple columns within target only</li> </ol>
Date Validation	<p>Date values are using many areas in ETL development for</p> <ol style="list-style-type: none"> <li>1. To know the row creation date</li> <li>2. Identify active records as per the ETL development perspective</li> <li>3. Identify active records as per the business requirements perspective</li> <li>4. Sometimes based on the date values, the updates and inserts are generated.</li> </ol>
Complete Data Validation	<ol style="list-style-type: none"> <li>1. To validate the complete data set in source and target table minus a query in the best solution</li> <li>2. We need to source minus target and target minus source</li> <li>3. If the minus query returns any value those should be considered as mismatching rows</li> <li>4. Needs to matching rows among source and target using intersect statement</li> <li>5. The count returned by intersecting should match with individual counts of source and target tables</li> <li>6. If minus query returns of rows and count intersect less than source count or target</li> </ol>



	table then we can consider as duplicate rows have existed.
Data Cleanness	Unnecessary columns should be deleted before loading into the staging area.

### 6.2.4 Types of ETL Bugs

Type of Bugs	Description
User interface bugs/cosmetic bugs	<ul style="list-style-type: none"> <li>Related to GUI of application</li> <li>Font style, font size, colours, alignment, spelling mistakes, navigation and so on</li> </ul>
Boundary Value Analysis (BVA) related bug	<ul style="list-style-type: none"> <li>Minimum and maximum values</li> </ul>
Equivalence Class Partitioning (ECP) related bug	<ul style="list-style-type: none"> <li>Valid and invalid type</li> </ul>
Input/output bugs	<ul style="list-style-type: none"> <li>Valid values not accepted</li> <li>Invalid values accepted</li> </ul>
Calculation bugs	<ul style="list-style-type: none"> <li>Mathematical errors</li> <li>The final output is wrong</li> </ul>

### 6.3 Test Scenarios and Test Cases

The following are generic test cases that need to be validated for any Analytical Testing Project.

Test Scenarios	Test Cases
----------------	------------

<b>ETL verification</b>	<ul style="list-style-type: none"> <li>• Verify data is mapped correctly from source to target system</li> </ul>
	<ul style="list-style-type: none"> <li>• Verify all tables and their fields are copied from source to target</li> </ul>
	<ul style="list-style-type: none"> <li>• Verify keys configured to be auto-generated correctly created in the target system</li> </ul>
	<ul style="list-style-type: none"> <li>• Verify that null fields are not populated</li> </ul>
	<ul style="list-style-type: none"> <li>• Verify data is neither garbled nor truncated</li> </ul>
	<ul style="list-style-type: none"> <li>• Verify data type and format in the target system is as expected</li> </ul>
	<ul style="list-style-type: none"> <li>• Verify there is no duplicity of data in the target system</li> </ul>
	<ul style="list-style-type: none"> <li>• Verify transformations are applied correctly</li> </ul>
	<ul style="list-style-type: none"> <li>• Verify that the precision of data in numeric fields is accurate</li> </ul>
	<ul style="list-style-type: none"> <li>• Verify exception handling is robust</li> </ul>
<b>Staging data</b>	<ul style="list-style-type: none"> <li>• Reconciliation check- record count between the STG (staging) tables and target tables are same after applying filter rules</li> </ul>
	<ul style="list-style-type: none"> <li>• Insert a record which is not loaded into the target table for the given key combination</li> </ul>

	<ul style="list-style-type: none"> <li>Copy records, sending same records that are already loaded into target tables-should not be loaded</li> </ul>
	<ul style="list-style-type: none"> <li>Update a record for a key when value columns changed on day_02 loads</li> </ul>
	<ul style="list-style-type: none"> <li>Delete the records logically in the target tables</li> </ul>
	<ul style="list-style-type: none"> <li>Values loaded by process tables</li> </ul>
	<ul style="list-style-type: none"> <li>Values loaded by reference tables</li> </ul>
<b>Data Loading in Tool</b>	<ul style="list-style-type: none"> <li>Check if the target and source databases are connected well, and there are no access issues.</li> </ul>
	<ul style="list-style-type: none"> <li>For a full load, check the truncate option and ensure its working fine.</li> </ul>
	<ul style="list-style-type: none"> <li>While loading the data, check for the performance of the session</li> </ul>
	<ul style="list-style-type: none"> <li>Check for non-fatal errors.</li> </ul>
	<ul style="list-style-type: none"> <li>Verify you can fail the calling parent task if the child task fails.</li> </ul>
	<ul style="list-style-type: none"> <li>Verify that the logs are updated</li> </ul>
	<ul style="list-style-type: none"> <li>Verify mapping and workflow parameters are configured accurately</li> </ul>
	<ul style="list-style-type: none"> <li>Verify the number of tables in source and target systems is the same</li> </ul>

	<ul style="list-style-type: none"> <li>Compare the attributes from stage tables to that of the target tables. They should be matched.</li> </ul>
<b>Tool Reports</b>	<ul style="list-style-type: none"> <li>Display date and time</li> </ul>
	<ul style="list-style-type: none"> <li>Decimal precision for key figures</li> </ul>
	<ul style="list-style-type: none"> <li>In a given page display the number of rows and columns</li> </ul>
	<ul style="list-style-type: none"> <li>Free characteristics in the report</li> </ul>
	<ul style="list-style-type: none"> <li>How are blank values/data displayed for both characteristics and key figures in the report</li> </ul>
	<ul style="list-style-type: none"> <li>Whether the search for characteristics is based on key or key &amp; text as Applicable</li> </ul>
	<ul style="list-style-type: none"> <li>Does search option on text is case sensitive- Upper, Lower or both</li> </ul>

# CHAPTER 7

## Conclusive Discussion and Future Work

### 7.1 Conclusion

Nowadays, modern companies are overwhelmed with data. Data collection and exploitation is an evident trend on which companies bet to ensure their presence in the market of the future. Telecom providers are the ones among many. In this thesis work, we started analyzing the data collected by the telecom companies. Then we decided to model and understand the relation of the data as we know that there must be an operational database system installed in all organizations, which is responsible for the daily transactions. A company or an organization must have to maintain these active databases 24/7. No doubt, these database applications tell us about the total trades on a daily or weekly basis. If a company's employee wants to perform the analysis on the business, they must have to move on a data warehouse.

So the purpose of the project is to provide end-users, mostly the company's manager, business analyst, a way of making better business decisions to improve their business. Intelligent reports which are developed from the company's data is stored in a data warehouse, can help end user to analyze the weak and strong areas. Telecom data analysis & Reporting was a small effort in a bid to accomplish the purpose. This project comprised two significant parts, and the first part is based upon designing a data warehouse. There comes a step of ETL, and In doing so, we have to extract the data on which we can perform analysis. After extraction, The other part of the project includes the reporting, in which we connect the semantic layer of a data warehouse with the visualization tool like Tableau.

### 7.2 Future Work

We are living in a world that is day by day evolving technology-wise. With the advancement in almost every practical field, the expectations and needs of people are also increasing. E.g., when the computer was invented, the expectation from it was to solve mathematical problems and calculations quickly to save time and increase efficiency. After some time and enhancement in

technology, the needs of a human being were improved and became advance. The same case applies here and in all IT projects.

Feature enhancement has its importance because with time system evolves, and there is a need for more features and functionalities, so the System should allow future enhancement. The prototype of the System has to develop in such a flexible way that it will enable the developer to add or remove more functionalities in it. The System should fulfill the constraints such as maintainability, portability, e.tc. Purposed System is developed by keeping this in mind.

This project can be improved in the following ways:

- Instead of building a single data warehouse on the source data, the Federated Data warehouse is made with the specific subject of analysis.
- Instead of carrying out simple analysis, python & R be used for the data analysis
- Instead of carrying data visualization with Tableau, visualization will be carried out with Power BI
- Instead of using Erwin Data Modeler, use Toad Data Modeler
- Web/mobile applications are developed to publish reports on them instead of making dashboards manually.