



T.C.

ALTINBAŞ UNIVERSITY

Graduation of Science and Engineering

Information Technology

**ANALYSIS OF BIG DATA USING CLOUD  
COMPUTING**

Asmida Ali R. ALASFAR

Master Thesis

Supervisor

Prof. Dr. Oğuz BAYAT

Istanbul, 2019

# **ANALYSIS OF BIG DATA USING CLOUD COMPUTING**

by

**Asmida Ali R. Alasfar**

Information Technologies

Submitted to the Graduate School of Science and Engineering

in partial fulfillment of the requirements for the degree of

Master of Science

ALTINBAŞ UNIVERSITY

2019

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of "Master of Science".

---

Prof. Dr. Oguz BAYAT

Supervisor

Examining Committee Members (first name belongs to the chairperson of the jury and the second name belongs to supervisor)

Prof. Dr. Oguz BAYAT

School of Engineering and  
Natural Sciences,  
Altinbas University

---

Asst. Prof. Dr. Abdullahi Abdu  
IBRAHIM

School of Engineering and  
Natural Sciences,  
Altinbas University

---

Asst. Prof. Dr. Adil Deniz  
DURU

Faculty of Sport Sciences,  
Marmara University

---

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science.

---

Asst. Prof. Dr. Oguz ATA

Head of Department

Approval Date of Graduate School of  
Science and Engineering: \_\_\_\_/\_\_\_\_/\_\_\_\_

---

Prof. Dr. Oguz BAYAT

Director

I, hereby, declare that all the information presented in this document has been obtained and presented according to the academic rules and the ethical code of conduct. I also declare that, following the mentioned rules and code of conduct, I have fully cited and gave references of all the resources, materials and results, which are not original to this work.

Asmida Ali R. ALASFAR



## **DEDICATION**

I dedicate this work to my mother, who is my first teacher, a great supporter and role model, and also to my father, who has remarkable contributions to my life, education and career. Without my parents, my dream of accomplishing higher education could never have come true. I am deeply indebted to the persistence of my wife, who has supported me during my master's degree, and specifically when I was accomplishing this thesis.



## ACKNOWLEDGEMENTS

In the beginning, I praise Allah, the Lord of the Worlds, who is benevolent to the people. I thank my father and mother for what they have done for me, and I acknowledge their contribution to my education, career, and life. I appreciate and deeply acknowledge the kind support, cooperation, and guidance of honorable supervisor Prof. Dr. Oğuz BAYAT, who helped me when I was completing this work. I also acknowledge the help and support of the university administration, people working in the concerned department, and my dear friends who extended great cooperation and encouragement during my studies.



## ABSTRACT

### ANALYSIS OF BIG DATA USING CLOUD COMPUTING

Alasfar, Asmida Ali R.

M.Sc., Information Technologies, Altınbaş University,

Supervisor: Prof. Dr. Oğuz BAYAT

Date: November, 2019

Pages: 73

The primary importance of the data lies in the fact that it provides information to take correct decisions. The data becomes unhelpful if we are unable to extract information out of it. The process of analyzing data means the ability to extract information from a specific data set that a user requires to make the decision-making process easier, and besides, it is helpful to make accurate decisions. There are many obstacles in Relational Database Management Systems (RDBMS) because of their old design. It is very difficult to meet the ever-changing and continuously increasing needs of the information technology systems. For example, these systems find it difficult to handle data in exabytes or zettabytes. Such large and unstructured data is termed as "Big Data." This thesis explains how to process magnanimous and difficult-to-handle unstructured data in order to control and analyze it. This problem has been highlighted when different solutions were introduced, which include the emergence of Cloud Computing (CC) that provides a useful alternative to the traditional systems, also known as massive Data Analytics. A large number of students, engineers, and staff members of Altinbas University participated in a survey as a part of this research. Some Libyan participants also participated in the survey. The participants were asked to fill in a questionnaire that had 30 questions focused on how Cloud Computing services are helpful to handle and analyze big data. We found almost of them participants had no knowledge of cloud computing services; so, it became a challenge; however, this thesis throws new light on the subjects like big data and cloud computing. A big data analysis project shows how to analyze one of the largest datasets in a quick and easy way. By the end of the thesis, we aim to provide a complete definition with a clear example that encourages users to use Cloud Computing Platforms instead of traditional memory.

**Keywords:** AWS, Big data, Cloud Computing, GCP, RDBMS, MS Azure.

## ÖZET

### BÜYÜK VERİ ANALİZİ VIA BULUT BİLİŞİM

Alasfar, Asmida Ali R.

Yüksek Lisans, Elektrik ve Bilgisayar Muhendisliği, Altınbaş Üniversitesi

Danışman: Prof. Dr. Oğuz BAYAT

Tarih: Kasım, 2019

Sayfa: 73

Verilerin öncelikli önemi, doğru kararlar almak için bilgi sağlamasıdır. Verilerden bilgi çıkarmazsak veriler işe yaramaz hale gelir. Verilerin analiz süreci, bir kullanıcının karar verme sürecini kolaylaştırmak için ihtiyaç duyduğu belirli bir veri kümesinden bilgi çıkarma yeteneği anlamına gelir ve ayrıca doğru kararların alınmasında yardımcı olur. İlişkisel Veri Tabanı Yönetim Sistemlerinde (RDBMS) eski tasarımlarından dolayı birçok engel var. Bilgi teknolojisi sistemlerinin sürekli değişen ve sürekli artan ihtiyaçlarını karşılamak çok zor. Örneğin, bu sistemler exabyte veya zettabaytlardaki verileri işlemeyi zor buluyor. Bu kadar büyük ve yapılandırılmamış veriler "Büyük Veri" olarak adlandırılır. Bu tez, kontrol etmek ve analiz etmek için büyük ve ele alınması zor yapılandırılmamış verilerin nasıl işleneceğini açıklamaktadır. Bu problem, muazzam Veri Analitiği olarak da bilinen geleneksel sistemlere yararlı bir alternatif sunan Cloud Computing'in (CC) ortaya çıkmasını içeren farklı çözümler ortaya çıktığında vurgulanmıştır. Bu araştırmanın bir parçası olarak Altınbaş Üniversitesi'nden çok sayıda öğrenci, mühendis ve personel ankete katıldı. Bazı Libyalı katılımcılar da ankete katıldı. Katılımcılardan, Cloud Computing hizmetlerinin büyük verilerin işlenmesi ve analiz edilmesi için nasıl yardımcı olduğuna odaklanan 30 soruluk bir anket doldurmaları istendi. Neredeyse katılımcıların bulut bilişim hizmetleri hakkında hiçbir bilgisi olmadığını gördük; Böylece, bir meydan okuma oldu; Ancak, bu tez büyük veri ve bulut bilişim gibi konularda yeni bir ışık tutuyor. Büyük bir veri analizi projesi, en büyük veri kümelerinden birinin hızlı ve kolay bir şekilde nasıl analiz edileceğini gösterir. Tezin sonunda, kullanıcıları geleneksel bellek yerine Cloud Computing Platformları kullanmaya teşvik eden açık bir örnekle tam bir tanım sağlamayı hedefliyoruz.

**Kelimeler:** AWS, Büyük veri, Bulut Bilişim, GCP, RDBMS, MS Azure.



# TABLE OF CONTENTS

	Pages
<b>ABSTRACT</b> .....	vii
<b>ÖZET</b> .....	viii
<b>LIST OF TABLES</b> .....	xii
<b>LIST OF FIGURES</b> .....	xiv
<b>LIST OF ABBREVIATIONS</b> .....	xvi
<b>1. INTRODUCTION</b> .....	<b>1</b>
1.1. LITERATURE REVIEW .....	2
1.2. PROBLEM STATEMENT .....	6
1.3. OBJECTIVES .....	8
1.4. THESIS OUT LINE.....	8
<b>2. CLOUD COMPUTING</b> .....	<b>9</b>
2.1. CLOUD COMPUTING ARCHITECTURE .....	10
2.1.1. The Conceptual Reference Model .....	10
2.1.2. Cloud Consumer .....	10
2.1.3. Cloud Provider .....	11
2.1.4. Cloud Auditor .....	11
2.1.5. Cloud Broker.....	11
2.1.6. Cloud Carrier .....	12
2.2. CLOUD SERVICE MODELS.....	12
2.2.1. Infrastructure-as-a-Service (IaaS).....	12
2.2.2. Platform-as-a-Service (PaaS).....	12
2.2.3. Software-as-a-Service (SaaS) .....	12
2.3. CLOUD COMPUTING BENEFITS .....	13
2.3.1. The Cost.....	13

2.3.2. The Speed.....	13
2.3.3. The Global .....	14
2.3.4. Increased Productivity .....	14
2.3.5. Better Performance .....	14
2.3.6. More Security.....	14
<b>2.4. DISADVANTAGES OF CLOUD COMPUTING.....</b>	<b>14</b>
2.4.1. Network Connectivity.....	15
2.4.2. Security Concerns .....	15
2.4.3. Prone to Attack .....	15
<b>2.5. DIFFERENT TYPES OF CLOUD COMPUTING .....</b>	<b>15</b>
2.5.1. The Public Cloud .....	16
2.5.2. The Private Cloud .....	16
2.5.3. The Hybrid Cloud .....	17
2.5.4. Community Cloud.....	18
<b>2.6. THE CLOUD DATA CENTER .....</b>	<b>19</b>
2.6.1. Data Center Locations.....	19
<b>3. CLOUD COMPUTING PROVIDERS .....</b>	<b>21</b>
3.1. TOP TREE FAMOUS PROVIDERS .....	22
3.1.1 Amazon Web Services (AWS) .....	22
3.1.2. Microsoft Azure .....	23
3.1.3. Google Cloud Platform (GCP).....	24
<b>4. BIG DATA .....</b>	<b>25</b>
4.1. BIG DATA TECHNOLOGIES .....	25
4.2. BIG DATA BROPERTIES.....	26
4.3. BIG DATA ANALYTICS .....	27
<b>5. METHODOLOGIES.....</b>	<b>29</b>
5.1 INTRODUCTION .....	29

5.2. AWS BIG DATA AND ANALYTICT PRODUCTS .....	29
5.3. MICROSOFT AZURE BIG DATA AND ANALYTICS PRODUCTS .....	31
5.4. GOOGLE CLOUD PLATFORM'S BIG DATA AND ANALYTICS PRODUCTS .....	35
5.5. GETTING STARTED WITH GOOGLE CLOUDPLATFORM .....	37
5.5.1. Signing up for GCP.....	37
5.5.2. Steps OF Creating a free Trail Account in GCP.....	38
5.5.3. Exploring the Console.....	40
5.5.4. Installing SDK.....	41
5.6. INTERACTING WITH GCP.....	41
5.6.1. In the browser: the Cloud Console.....	41
5.7. BIG QUERY.....	43
5.7.1. What Is BigQuery? .....	43
5.7.2. Why BigQuery?.....	43
5.8. CASE STUDY .....	44
5.8.1. Survey .....	44
5.8.1.1. Data survey collection.....	44
5.8.1.2. Data survey analysis using ibm spss.....	45
5.8.2. Querying data.....	55
5.8.2.1. BigQuery’s public datasets .....	55
5.8.2.2. BigQuery’s uploading dataset .....	56
5.8.3. Real-Life Case Study .....	57
5.8.3.1. Real-life case study on a public dataset .....	57
5.8.3.2. Real-life case study by uploading the dataset .....	62
5.9. RESULTS OF THE CASE STUDY.....	67
<b>6. CONCLUSION .....</b>	<b>68</b>
<b>REFERENCES.....</b>	<b>70</b>

## LIST OF TABLES

	<b><u>Pages</u></b>
Table 5.1: AWS Big data and Analytics product list.....	30
Table 5.2: Microsoft Azure Big Data and Analytics Product.....	32
Table 5.3: GCP Big Data and Analytics Product.....	35
Table 5.4: Demographic Data.....	46
Table 5.5: The result of analysis for section I.....	47
Table 5.6: The result of analysis for section II.....	50
Table 5.7: The result of analysis for section III.....	52

## LIST OF FIGURES

	<u>Pages</u>
Figure 2.1: Cloud computing .....	9
Figure 2.2 : Cloud Architecture .....	10
Figure 2.3: Cloud Provider - Major Activities .....	11
Figure 2.4: Cloud Computing services .....	13
Figure 2.5: Public Cloud .....	16
Figure 2.6: Private Cloud .....	17
Figure 2.7: Hybrid Cloud .....	18
Figure 2.8: Community Cloud .....	18
Figure 2.9: Google Data Center .....	19
Figure 2.10: Google Data Locations Center .....	20
Figure 3.1: Most Used Cloud Computing Services in the Enterprise Market .....	21
Figure 4.1: shown Different properties of Big data .....	27
Figure 4.2: the benefits of big data analytics .....	28
Figure 5.1: Google Cloud Platform .....	37
Figure 5.2: Google Cloud Platform free trial.....	38
Figure 5.3: Select The Country And Agree To The Policy Of GCP.....	39
Figure 5.4: Enter The Personal Details.....	39
Figure 5.5: Enter The Payment Details .....	40
Figure 5.6: Submit And Start Your Trial Version.....	40
Figure 5.7: Google Cloud Console.....	41
Figure 5.8: Google Cloud Console, where you can create a new virtual machine.....	42
Figure 5.9: Form where you define your virtual machine.....	42

Figure 5.10: Showing The Demographic Data.....	46
Figure 5.11: Showing The Results of Section I.....	49
Figure 5.12: Showing The Results of Section II.....	52
Figure 5.13: Showing The Results of Section III .....	55
Figure 5.14: BigQuery’s public datasets .....	56
Figure 5.15: Sample CSV file .....	56
Figure 5.16: The yellow taxi trips schema .....	57
Figure 5.17: The yellow taxi trips table .....	58
Figure 5.18: BigQuery results of the most expensive trip .....	59
Figure 5.19: Results of querying with grouping by pickup time .....	60
Figure 5.20: Results showing the day and hour with most pickups .....	61
Figure 5.21: Finding the TED dataset on Kaggle website .....	62
Figure 5.22: Process to create new datasets .....	63
Figure 5.23: Creating a data set in BigQuery .....	64
Figure 5.24: Uploading file to BigQuery Datawarehouse .....	65
Figure 5.25: Adding table name .....	65
Figure 5.26: Querying table on BigQuery Datawarehouse on Created Datasets .....	66
Figure 5.27: Query output .....	67

## LIST OF ABBREVIATIONS

API	: Application Program Interface
AWS	: Amazon Web Services
BD	: Big Data
BDaaS	: Big Data as a Service
BI	: Business Intelligence
Capex	: Capital Expenditure
CC	: Cloud Computing
CPU	: Central Processing Unit
CSV	: Comma Separated Values
DR	: Disaster Recovery
EC2	: Elastic Compute Cloud
EMR	: Elastic Map Reduce
ETL	: Extract, Transform, Load
GCP	: Google Cloud Platform
GPU	: Graphics Processing Unit
HDD	: Hard Disk Drive
IaaS	: Infrastructure as a Service
LB	: Load Balance
MS	: Azure Microsoft Azure
NIST	: National Institute of Standards and Technology
Opex	: Operational Expenditure
PaaS	: Platform as a Service
RAM	: Random Access Memory
RDBMS	: Relational Database Management Systems
S3	: Simple Storage Service
SaaS	: Software as a Service
TED	: Technology Entertainment and Design
VM	: Virtual Machine

# 1. INTRODUCTION

The modern computers have enough memory to process and store small data because the memory comprises sorting algorithms; so, it is easy for this memory to process this data, for instance, up to 10GB. What if there is a need to process 100 GB or One Terabyte (TB) of data? Some high-end configuration servers are available to hold this much data in the memory but they are quite expensive; therefore, it is better to select disk-based systems, but in this case, algorithms such as mergesort can be used for sorting the data. But what if there is a need to sort 50TB, 100TB or even more data? This is only possible with multiple parallel disk systems, but in this case, a different algorithm such as bitonic sort should be used. These scenarios clearly indicate that different data sizes require different solutions [1][2].

Since data are enormously growing around the world, so it has become extremely difficult to process and analyze them. This situation is not surprising or strange, because if we look at the beginning, when huge amounts of data was created in the year 2000 and since then, the data has been piling up; so, it has become difficult to handle. This data overstock is termed as big data, and it is subdivided into structured data (for example RDBMS) or unstructured data (data on social media sites or organizational data). The process of analyzing huge data is known as big data analytics. Cloud computing is a network of internet-connected servers, which are used to manage, process, and store data on internet network rather than on local servers. Also, through these connected servers, various analyses and applications can be used, and even more can be accomplished.

On the other hand, this thesis educates the new data users, who are unaware of some of the data processing solutions, which are available. In the nutshell, this thesis provides full knowledge of the available solutions to the big data issue. The thesis also explores and examines the cloud computing fundamentals mentioning the major benefits of cloud computing for big data analytical projects. Several cloud service providers offer data analytics solutions, for example, three most famous cloud service providers are operating, which include Microsoft Azure (MS AZURE), Amazon Web Services (AWS), and Google Cloud Platform (GCP). We also conducted a survey of a wide range of internet users' views and computing devices to demonstrate the importance cloud computing research. This is done by answering many questions. The answers to the



questionnaire and the results have been compiled to understand cloud computing more clearly and comprehensively. A small demonstration has been given to demonstrate how easy it is to conduct big data analytics in a cloud computing environment.

### **1.1. LITERATURE REVIEW**

Several data analysis software designers believe that data is the “new oil,” and the system analysts are mining into the available data to extract information. In the past, the enterprise/commercial data was not as interesting for investors as it is now. The reason is that the data can be reused for real results; so, many countries have opened regional data centers believing that these data are important for future use. Data analysis means the use of preliminary information to get a final value to improve the effectiveness of the enterprise. In fact, the economic value of data will be the same as the economic value of petroleum in the near future. The volume of data created over the last three years is greater than the data created since the beginning of history. According to an Intel company report, the data generated since the beginning of history is almost 5 exabytes. This number increased 500 times in 2012; so, it is expected that the size of the data will be about 44 zettabytes by 2020, which means that it will be 4.4 trillion gigabytes. This has happened because of the following reasons:

- Every minute, 300 hours of videos are uploaded on YouTube.
- Every minute, a million emails are sent.
- Every minute, a million likes are placed on Facebook.
- Google produces 40,000 searches every second.

Dealing with these data has become imperative, especially when we know that until now, only 1% of the data has been managed. This clearly indicates the total data worldwide, expanse of data and its magnitude. The objective behind the emergence of big data technologies is to manage and use this huge amount through new technology. One of the best technologies to deal with large data is called Hadoop. It is important because it helps many countries, institutions and governments create new jobs in the fields of data collection and analysis. In the nutshell, we must prepare ourselves for a future based primarily on managing, controlling, and analyzing vast data to build realistic results that we can use in our daily lives. This thesis disseminates information on how to analyze big data and contributes by providing comprehensive and clear data management and analysis concepts for researchers. Big data have very quickly demonstrated that

they are a great example of the impact of information technology on business, also it has the same impact in the decision making at the enterprise level this was according to what wrote Constantiou and Kallinikos ( 2015), This brings us to the fact that big data will be a clear impact on institutions[41]. They transform the way that companies relate to both their customers and employees and the way they enact and perform business operations, Wamba et al (2015)[48]. Through the appropriate implementation of big data initiatives, companies have the potential of renewed business value creation along with increased productivity and innovativeness, This was by Maglio & Lim (2016) [49]. according to Constantiou & Kallinikos, (2015), The importance of big data and business analytics is evident throughout the literature and as they evolve, they have various applications creating multiple emerging research areas. A heated discussion over the past years has been on the opportunities and challenges that big data bring to organizations, communities, and individuals[41]. and according to McAfee et al (2012) and Varian (2014), the Past studies have highlighted the importance of big data and business analytics in various areas such as customer relationship management, new business models and short-term economic predictions[42]. however, the value of big data has also been noted in terms of combining information from various sources, creating innovative services. This suggestion, Makes it easy to handle big data as required and the ability to analyze this data that offers the opportunity to excel in the field, and this is done of combining multiple sources of data in order to derive value, has been advocated by multiple scholars following recent paradigms presented in the business world. indicates Constantiou & Kallinikos, (2015) in his article The application of big data in driving organizational decision making has attracted much attention over the past few years. A growing number of firms are focusing their investments on big data analytics (BDA) with the aim of deriving important insights that can ultimately provide them with a competitive edge[41]. also, McAfee et al, (2012) say the same talk, The need to leverage the full potential of the rapidly expanding data volume, velocity, and variety has seen a significant evolution of techniques and technologies for data storage, analysis, and visualization. However, there has been considerably less research attention on how organizations need to change in order to embrace these technological innovations, as well as on the business shifts they entail. he continues in the talk, Despite the hype surrounding big data, the issue of examining whether, and under what conditions, big data investments produce business value, remains underexplored, severely hampering their business and strategic potential[42]. This talk gives a good impression and a real

motivation for more effort and time for further studies and researches about big data analytics(DBA). While Gupta & George, (2016) says Most studies to date have primarily focused on infrastructure, intelligence, and analytics tools, while other related resources, such as human skills and knowledge, have been largely disregarded. Furthermore, orchestration of these resources, the socio-technological developments that they precipitate, as well as how they should be incorporated into strategy and operations thinking, remains an underdeveloped area of research[43]. In fact, it is up to us at this time how to integrate these new technologies into our work. According to his opinion the Laudon & Laudon, (2014), IT infrastructure models themselves have also evolved, most recently from Enterprise to Cloud Computing after The development of the Internet. in fact, There are five stages in this evolution representing different infrastructure elements; namely, Stage 1: General-Purpose Mainframe and Mini-computer Era, Stage 2: Personal Computer Era, Stage 3: Client/Server Era, Stage 4: Enterprise Computing Era and Stage 5: Cloud and Mobile Computing Era (Laudon & Laudon, 2014, p. 197–200). Cloud Computing started in the 2000s as the last stage of IT infrastructure evolution which refers to a computing model where organizations or individuals obtain computing power and software solutions over the Internet or other networks. These phases of development were a rapid response to the tremendous development in communications technology[45]. This research has a great deal of interest in cloud computing and therefore, we must consider and throw some literature reviews of cloud computing. It could start with a definition as according to Foster et al, (2008), Cloud computing can be regarded to a certain degree, as the evolution of grid computing. Such a close relationship has caused confusion. The grid framework is originally driven by scientific purposes and aimed at coordinating resources that are not subject to centralized control under standard, open, general-purpose protocols and interfaces. Cloud computing is born for commercial purposes and naturally, service-oriented. It is based on centralized data centers. The protocols and interfaces used may not be the same across cloud providers. Is characterized by Cloud computing has a completely distinct business model. It offers clear SLAs (Service Level Agreements) and is based on a “pay per use” pricing model, it was according to Weinhardt et al., (2009). Therefore it is promised that with nothing but a credit card, one can get on-demand access to 100,000+ processors from the clouds. Grid computing, on the other hand, is based on a sharing system, that is, one needs to join the grid and contribute computing power to gain access to the computing

power of other members. In this co-operative model, SLAs are not required or enforceable[46]. while the Vaquero et al., (2009) said These resources of cloud computing are typically built in centralized data centers and are dynamically re-configured to achieve optimum utilization. Clouds are provided by a pay-as-you-go model in which guarantees are offered by the providers by means of customized SLAs (Service Level Agreements). This turns computing power into a public utility that will bring a profound “paradigm shift” to the IT industry and even to society as a whole[47]. as a supplement, the Fox et al, (2009), said the Cloud computing has promised many technological and sociological benefits. The computing power is generated from highly centralized and standardized data centers that contain up to millions of servers, with a considerable economy of scale. From an enterprise standpoint, cloud computing can deliver on-demand computing power at a very low if not any cost of the upfront infrastructure and ongoing maintenance. Cloud computing also promises to provide better performance, reliability, and scalability. There is some evidence that these are being delivered. From an environmental standpoint, owing to the advanced electrical and cooling systems used by its centralized data centers, cloud computing has promised to bring low environmental cost and high energy efficiency, compared to the traditional scattered enterprise data centers. All in all, these seductive promises have drawn drastically increasing the attention on a worldwide scale[50]. It is very difficult for traditional databases to handle big data that is why, it is very difficult and complex. Due to this difficulty, there is a need for large programs to deal with huge data. Big data is product of several modern technologies that help large companies take decisions. This assistance allows them to gain a practical vision that gives them priority in reaching very important results in a short time [5]. To delve into and deal with big data, especially with respect to the analysis according to Fan, Han, and Liu (2014), the knowledge of big data deserves attention, especially in the statistical and accounting aspects, because it is followed by several challenges. These challenges are very important features of the big data, and they increase complexity. Generally, challenges emerge when there is complex, large, dependent, or irregular data that has noise. All these challenges must be encountered. This necessitates continuously finding solutions to the big data issue, understand all of its positive aspects, and harness all the possibilities to understand the challenges. According to Gupta and Khan (2015), the use of cloud computing has become the global focus of attention in various fields. It has become very important since all data must be preserved from theft and malicious attempts. Massive data analysis and cloud computing have great futuristic possibilities [6]. It is

so because large-sized data cannot be handled through traditional memory as it assures desired results, especially the data of the range in petabytes and zettabytes etc. It is necessary to move forward to identify the alternative and solve the problem of large data. Gupta et al. (2012) stressed the need for big data processing using existing cloud computing resources, but he did not hide the concerns of some companies about moving to cloud services, need to increase the cloud security, and effectiveness to access the data [7]. According the Yang and Zhao (2017), it is possible to add important applications to cloud computing such as biomedical applications, which can work with the data storage in the online data cloud without causing any serious problem [8]. When data processing is accomplished through cloud computing, it needs additional software for processing biomedical data. The views of various experts and researchers show that there is a need to continuously study big data and its analysis to get the best advantage, especially after the availability of the necessary technology and service providers. This is a good direction to move forward, and of course, this gives us the incentive to understand more about the big data and the available analyzing options. Through this review, it will be clarified that cloud computing is the core of big data. There is no doubt that traditional data storage and processing methods are going to disappear in the near future.

## **1.2. PROBLEM STATEMENT**

The small size data can be absorbed by traditional memory but everyday data changes and increase in size have been observed all over the world. This change in data size has become a fact because in the last 18 years, this increased data has become a real problem, and even until now, the data is still continuously generated; so, this humongous data is now a big problem to handle and process. Big data or BD was introduced as a terminology to represent large-scale data, be it structured (like RDBMS) or unstructured (like social media or organizational data etc.). The analyzing process of such large-scale/big data is, as mentioned earlier, is called as big data analytics. Cloud computing is a series of computing services including networks, servers, large-scale storage devices, databases, analytic applications, and internet devices. The problem is not only limited to how to handle and process humongous data. Many people do not know that these techniques and services are provided through the web. This leads us to the fact that the problem addressed in this research can be described in two important questions [3] [4].

- How can we solve the problem of big data and analyze it?
- How do we move users to learn about cloud computing services and stop using traditional storage devices?

### **Why Big Data is a current need?**

The two questions mentioned above are in the minds of many computing professionals. The big data allows access to large data volumes, which are useful for critical insights in the repeated/unique data patterns. It is a machine-managed learning process that has very limited human intervention that makes analysis error-free and simple. If we look at the reason, as to why it is important to use it in the current era, we'll find out that now the platforms, infrastructure and data processing frameworks such as Hadoop and NoSQL are commercially available at significantly low cost and high accessibility as compared to traditional systems and resources. The processing architecture of new platforms has its limitations in terms of working in cooperation with the conventional data processing technologies and existing algorithms and methodologies. It is important to understand that Big Data has an always present data part, which is manually used in processes that need substantial human processing and analytical refinements before using it for decision-making. Big Data has created buzz in the industry because it leads to automated data processing at a very high pace with flexible processing options.

Every organization has its own data needs; so here, we have mentioned some examples of data:

- **Contract data:** Many organizational contracts comprise this form of data because organizations do several contracts every year, which are important because they have multiple liabilities.
- **Weather data:** Large weather data is reported when the governmental agencies collect and report weather changes for the benefit of scientific organizations, farmers and consumers. This data is presented on TV, internet and radio in a final form in terms of key performance indicators (KPIs) such as weather forecast and temperature.
- **Labor data:** Different types of workforce are almost always linked with some kind of issue that an organization has to solve; therefore, labor data is important to maintain.
- **Financial reports/data:** It includes tax documents, annual stock listings and corporate performance reports.
- **Maintenance data:** It includes data pertaining to maintenance of machines, automobiles, facilities, and non-computer systems.

- Clinical trials' data: Research organizations and pharmaceutical companies maintain this kind of data for minimizing the time required to process clinical trials' data; so it can be an opportunity for Big Data.
- Compliance data—financial, healthcare, life sciences, hospitals, and many other agencies that file compliance data on the behalf of their client corporations.
- Doctors' notes, diagnoses and treatments: It is another potential area that has valuable insights and it results in proactive diagnosis.

### **1.3. OBJECTIVES**

This research aims to shed light on the most important modern techniques in the field of big data analytics. The goals can be summarized at several points :

1. Search for new ways to accommodate big data.
2. Solving the problem of big data by knowing how to analyze it and finding analysis methods to control this data by providing the real cases study and apply the big query.
3. access to the most important and the latest techniques of big data analytics and know-how the analysis data.
4. Provide a questionnaire targeting large samples of engineers and students participating in order to know and evaluate their opinions about cloud computing services and big data analytics.

### **1.4. THESIS OUT LINE**

This thesis has several sections. The first section mentions the purpose of this thesis in addition to its scope. The second and third sections describe cloud computing principles and description of cloud computing service providers. The fourth section presents big data fundamentals, The fifth section includes research methodologies adapted for this thesis while the final section, the conclusions of the thesis will be presented.

## 2. CLOUD COMPUTING

During the last two decades, cloud computing showed amazing development in the science of information technology, which combined with the great development in the field of telecommunications as well as high-speed networks. It has led to the emergence of new possibilities in IT. Cloud computing is a summation of several online computing facilities, such as data processing, configuration and high-speed access to applications on the internet [9]. Cloud computing gives us many services like online storage, servers, databases, business intelligence options, analytics, and software etc.

Cloud computing infrastructure provides right environments for storage. Cloud computing is capable of handling millions of operations in seconds distributing instructions on several interconnected servers over the internet. The operations of users on a desktop or portable computer require only one installation that allows them to access the cloud, which in turn provides all the users' needs to host programs or applications and web services [10]. This feature reduces the load on local hardware and also decreases the cost of using the resources that are needed to run applications. Moreover, the access to cloud computing services requires only the internet access. For example, access to Google Drive services requires only a user name and a secret number.

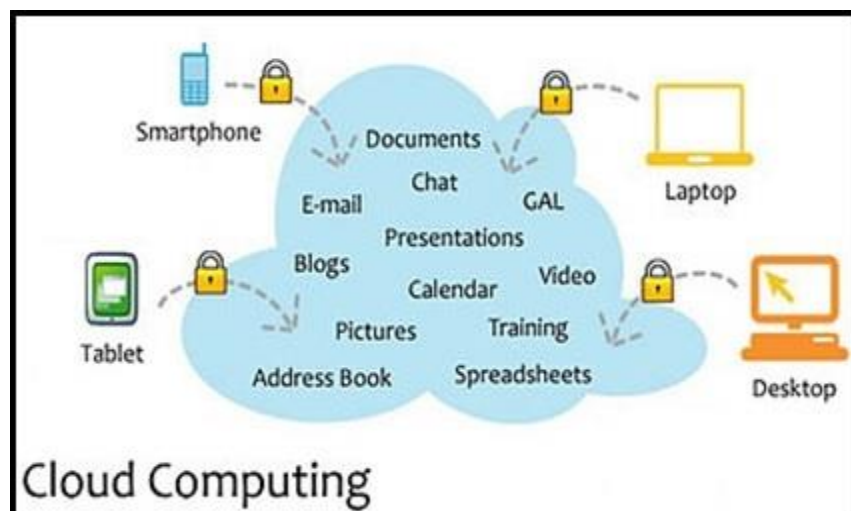


Figure 2.1: Cloud Computing [9]



## 2.1. CLOUD COMPUTING ARCHITECTURE

### 2.1.1. The Conceptual Reference Model

A model has been presented at the National Institute of Standards and Technology (NIST) that serves as reference architecture to understand cloud computing. It has been divided into five components, which represent cloud computing architecture. The following diagram facilitates the process of understanding the cloud computing needs. The simplified form for each of the components is given below [11].

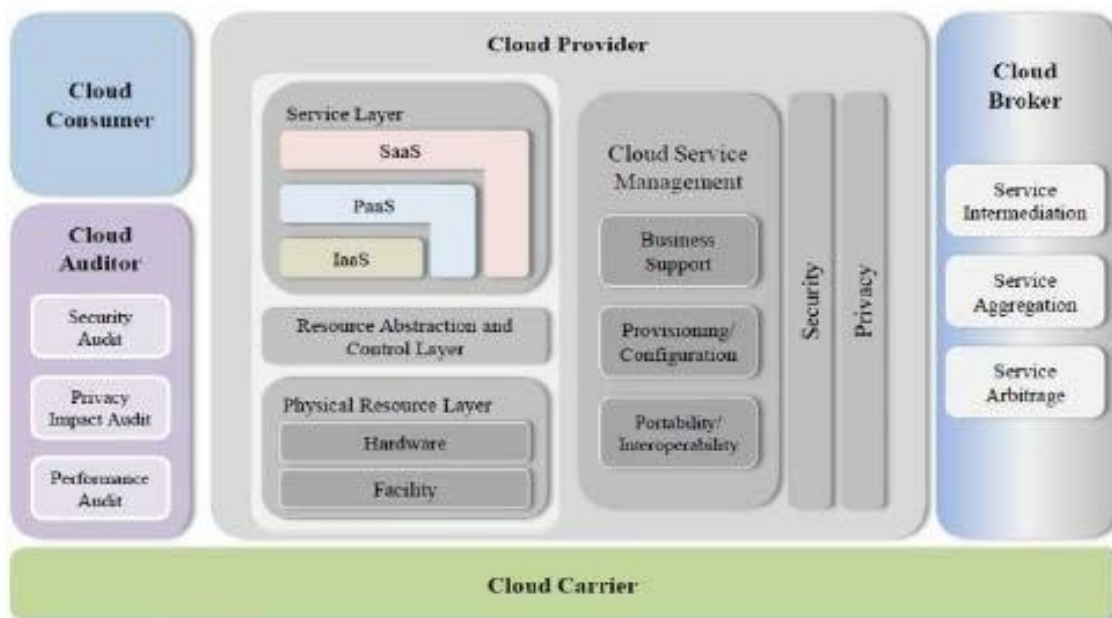


Figure: 2.2: Cloud Architecture [11]

### 2.1.2. Cloud Consumer

Cloud consumers request the cloud computing organizations to provide them with online data processing and storage facilities. The service consumers usually review different options that a cloud service provides, and chooses the most appropriate services they wish to receive. The service is availed after signing contracts between the service providers and consumers for agreed fees. Service consumers and service providers need to enter into SLAs that are mutually agreed by the type of service, terms, conditions, and restrictions [12].

### 2.1.3. Cloud Provider

Cloud service providers are responsible for providing cloud computing services to other parties according to a prior agreement. One of the most important tasks of service providers is to purchase and manage the required computing infrastructure as well as manage the cloud program that arranges storage space and provides the cloud service. Cloud service providers have activities in five key areas, as Figure 2.3 shows [12].

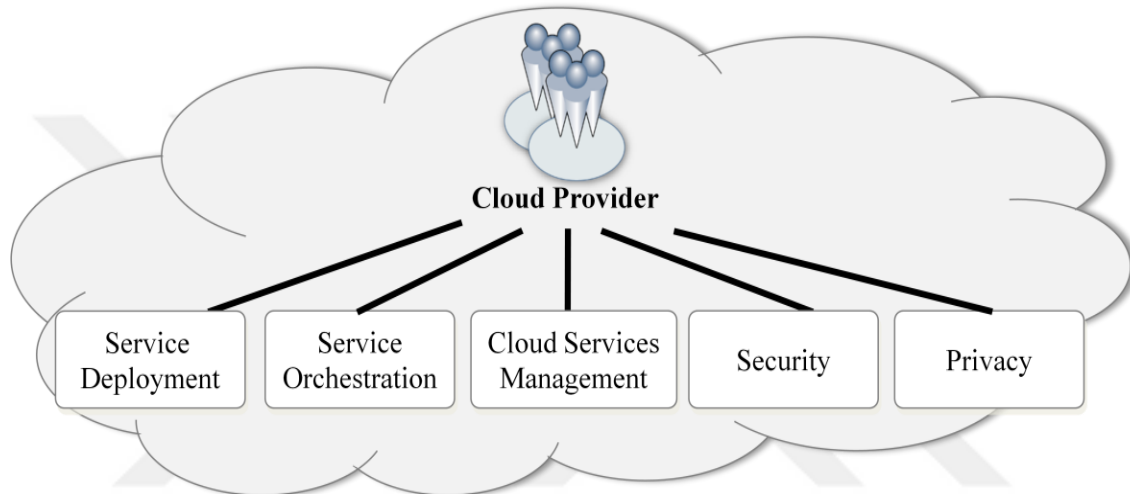


Figure 2.3: Cloud Provider - Major Activities [11]

### 2.1.4. Cloud Auditor

Generally, it is a third party that checks the controls of the cloud service by evaluating the services. These measures work to ensure users' privacy and maintain a proper image of a cloud computing resource [12].

### 2.1.5. Cloud Broker

A broker acts as a link between the cloud providers and the consumers because it manages and uses the cloud services and helps the two parties negotiate with each other. This service has enormous development in the field of cloud computing services, which has added to the complexity of this service. Cloud brokers primarily provide the following three services:

- **Service Intermediation:** Cloud brokers enhance the computing service by adding a particular capability to the systems, and sometimes, they act as value-added service sellers to the consumers. Such improvements may be improved access, managing identities, generating

performance reports, or improved security [12].

- **Service Aggregation:** Cloud brokers integrate and combine many services to form a new service. Data integration assures the secure transfer of data between multiple cloud providers and the cloud consumers [12].
- **Service Arbitrage:** It is just like service aggregation but in this case, aggregated services aren't fixed. In this case, the broker chooses variety of services from different cloud vendors, for instance, a cloud broker might provide a credit-scoring service or choose the best scoring agency [12].

### **2.1.6. Cloud Carrier**

Cloud carriers are intermediaries, which assure connectivity as well as cloud service provision to cloud consumers after making a deal with cloud service providers. Generally, cloud operators assure consumer access through telecom services, networks, and data access devices. Cloud consumers then connect with the cloud using network access through laptops, computers, handheld/mobile devices or mobile internet devices (MIDs) [12].

## **2.2. CLOUD SERVICE MODELS**

Several cloud computing services are offered, which can be mainly subdivided as follows:

### **2.2.1. Infrastructure-as-a-Service (IaaS)**

IaaS services are offered to the client via applications loaded on the substations, which are mainly application software interfaces that allow access, the operation of their virtual servers and their storage units, and, if necessary, some virtual services depending on the users' needs [13].

### **2.2.2. Platform-as-a-Service (PaaS)**

PaaS developers build or create software and applications on the server platform that is operated through internet. These applications allow users to access several service providing tools [13].

### **2.2.3. Software-as-a-Service (SaaS)**

Many companies compete with other companies to provide cloud computing service, although this technique is new. In structure, SaaS vendors provide users with the necessary infrastructure

such as hardware and software [14]. Figure 2.4 represents three layers of cloud service models. These three layers are further described below:

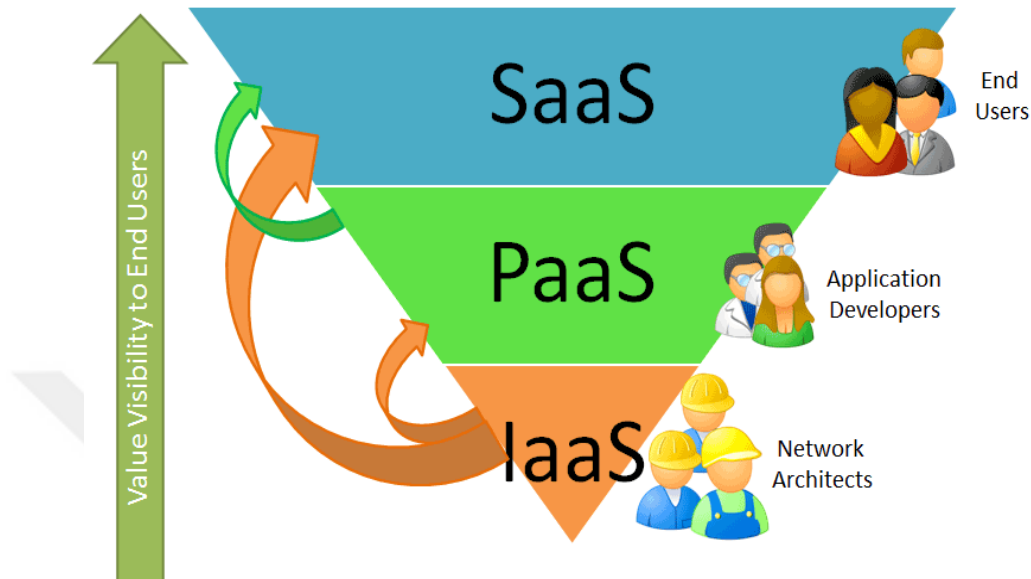


Figure 2.4: Cloud Computing Services [11]

## 2.3. CLOUD COMPUTING BENEFITS

Cloud technology has changed the concept of doing business for a lot of companies, where companies tend to request cloud computing technologies because they provide a lot of benefits to all the users. Major advantages of using cloud technology are presented below:

### 2.3.1. The Cost

Cloud computing technologies have reduced the cost for companies, especially startups because they do not need to spend large money on data. Such expenses are paid when a company pays for purchasing software and hardware, and besides, huge data management center for managing their data. This entails the need for the presence of specialized teams of IT engineers and technicians. Also, costs are linked with operations processes such as electricity, building, and rental insurance, etc. [37].

### 2.3.2. The Speed

Cloud services provide a fast response to any request, as well as great flexibility to provide cloud computing resources to the user whenever needed, and besides, it reduces the need for capacity

### **2.3.3. The Global**

One of the biggest benefits of cloud computing is the global scalability. Companies can deliver the exact amount of required resources, be it power, bandwidth, or storage wherever and whenever needed, and from the most suitable geographic location [37].

### **2.3.4. Increased Productivity**

When an on-site datacenter is established, a lot of work is needed to set up hardware, manage it, and modify the software as well as many other IT jobs that take time. With cloud computing, most of them are either no longer required or they remotely perform through the cloud service that gives the IT teams more time for achieving all their goals [37].

### **2.3.5. Better Performance**

Some of the largest cloud services are run with the help of huge network of data centers across the world, all of them are secure and updated, and they are regularly upgraded to assure speed and efficiency. Compared to an on-site data center, this provides reduced latency, and far better scaling economies [37].

### **2.3.6. More Security**

Several cloud computing services provide the technologies, policies, and controls needed to ensure full security. This maintains data, applications and infrastructure much safer from potential threats. The benefits of cloud computing substantially outweigh any potential downside, which is why; so many businesses have opted for it [37].

## **2.4. DISADVANTAGES OF CLOUD COMPUTING**

Although cloud computing has several benefits, the "Cloud Security Alliance" found out many barriers, which are stopping the companies to adapt it. According to a survey, 73% company executives expressed concerns over data security in cloud-based options. About 38% executives expressed concerns over regulatory compliance, and many of them believed that cloud will make companies lose their control over their own IT processes. Approximately 34% respondents had experience of both managing businesses and hands-on experience on IT projects. Since organizations need to deal with their compliance and security issues through appropriate corporate policies and invest to overcome gaps so as to take full advantage of cloud services [38].

### **2.4.1. Network Connectivity**

Users are always dependent on internet connectivity to access cloud services. Different services have different requirements for internet connections in terms of internet speed and network latency etc. [38].

### **2.4.2. Security Concerns**

One of the main issues pertaining to cloud is the security issue. A company needs to decide before adopting this technology whether it is willing to give sensitive information to a cloud service or not. It can expose a company to a potential threat/risk; therefore, a reliable service provider is needed to assure information security [38].

### **2.4.3. Prone to Attacks**

Cloud information storage might make a company vulnerable to hacking attempts as well as threats to its information [38].

Cloud service providers are consistently targeted, and it is a top priority for cloud service providers to remain protected as they have many organizations' data in their data centers. Some common cloud attacks include:

- Distributed denial of service attacks: Traditionally in DDoS, many systems at once overload a target server, causing it to either be less effective or cease its operations. In 2016, Dyn attack demonstrated that large websites such as Amazon and Twitter's accesses can be made unavailable to customers [38].
- Man-in-the-cloud Attack: It is a recent attack type, which targets a cloud user's synchronization token. A synchronization token is either a file stored in cloud, users' machine in a directory, registry or Windows Credential Manager. The victim (user) is hit with malware through emails or websites, to gains access to the victim's local files.

When the cloud synchronization token is replaced with the attacker's cloud account, and the original token is placed in the selected files, which are synchronized; consequently, the victim uploads his/her original token, which is accessible to the attacker. It can be used to access the victim's cloud data [38].

## **2.5. DIFFERENT TYPES OF CLOUD COMPUTING**

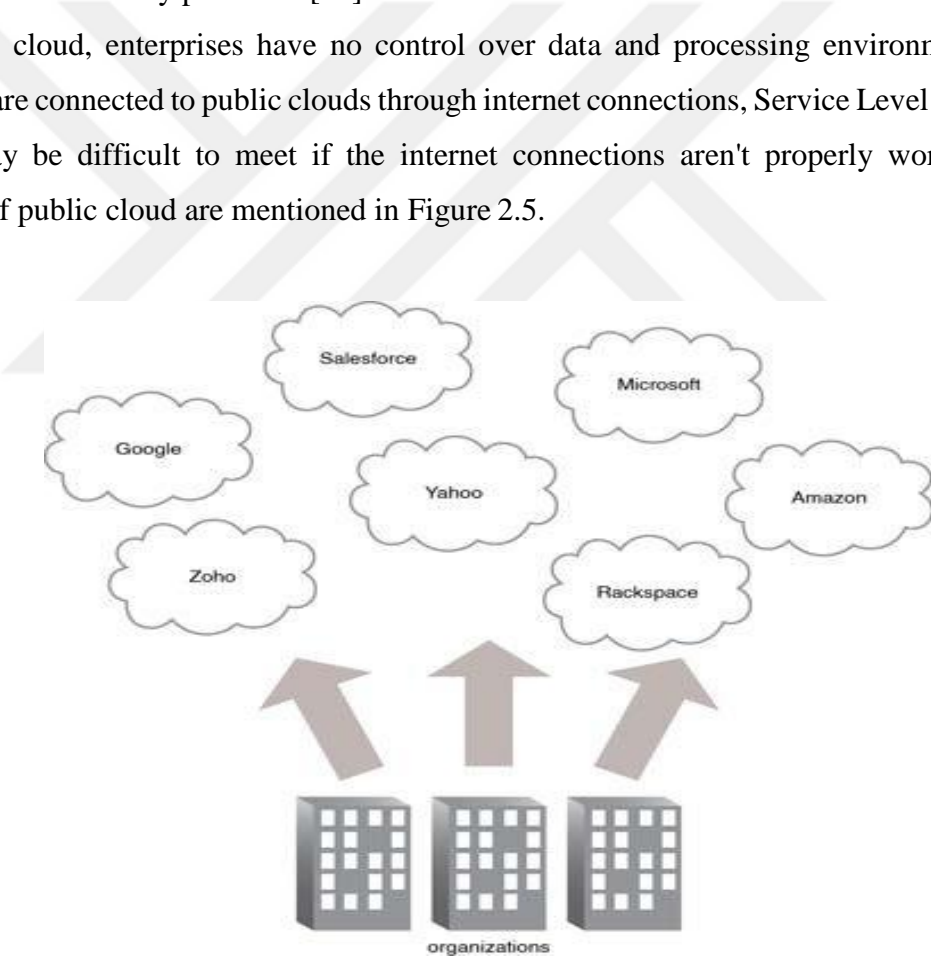
Cloud computing services are numerous and vary from one user to another over time as cloud

computing evolved. Currently, there are many models available; so, there are three main cloud deployment service types.

**2.5.1. The Public Cloud**

Some public cloud services (see Figure 3.5) follow the standard Cloud computing model that bounds a service provider to assure availability of virtual machines (VMs), storage or applications to public through internet. The public cloud is a service offered to the general public; therefore, businesses, educational institutions or governments own, manage, and operate them. These cloud infrastructures are hosted on cloud providers' huge scale data centers spread across the globe solving disaster recovery problems [15].

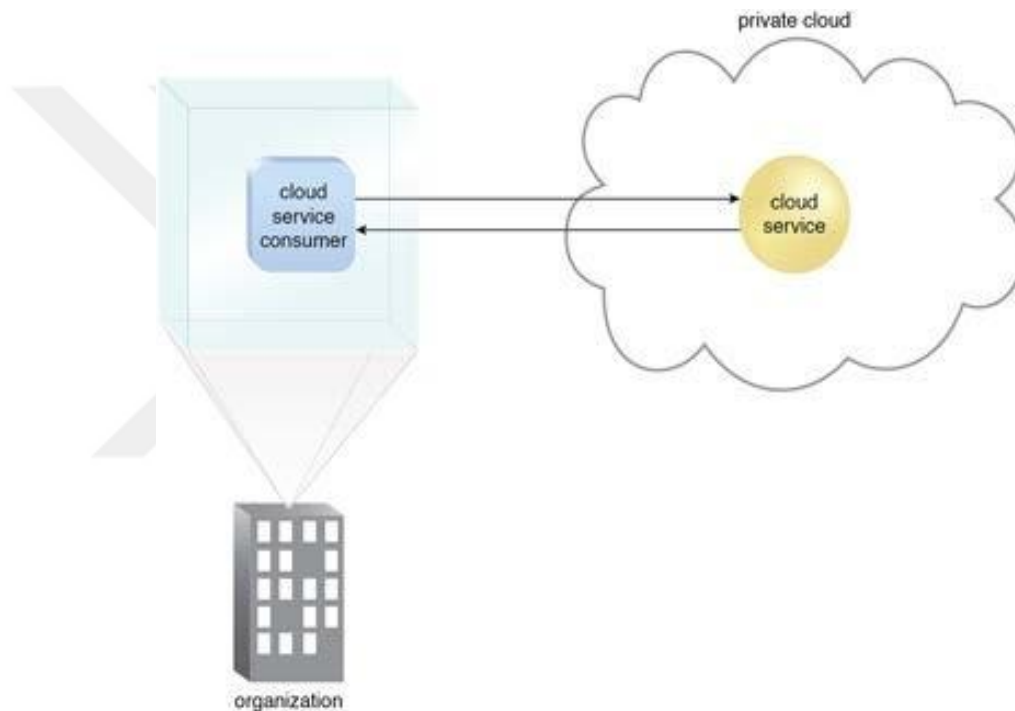
In a public cloud, enterprises have no control over data and processing environments. Since customers are connected to public clouds through internet connections, Service Level Agreements (SLAs) may be difficult to meet if the internet connections aren't properly working. Some examples of public cloud are mentioned in Figure 2.5.



**Figure 2.5:** Public Cloud [15]

### 2.5.2. The Private Cloud

Private cloud services (see Figure 2.6) exclusively serve a single organization that has multiple clients, for example, business units, third parties, or vendors accessing the shared resources. Private clouds are mostly operated by a single organization, a partnership, or after agreement with a third party. It can be seen in Figure 3.6 that the infrastructure can be either on or off premises. In the last decade, majority of large organizations established their own data centers, and transformed them into private cloud-based solution centers through virtual technologies [15].



**Figure 2.6:** Private Cloud [15]

### 2.5.3. The Hybrid Cloud

This cloud type (see Figure 2.7) consists of more than one cloud providers, who serve the same consumer. As an example, a banking customer may want to keep its business related financial data in a private cloud and it might be willing to operate marketing related tasks on a different cloud. Hybrid cloud is quite popular among enterprises. This is due to their large existing infrastructure, which is not so straightforward to move onto the public cloud; so, the enterprise either moves a part of their existing resources into a cloud or directly start new projects in the cloud [15].



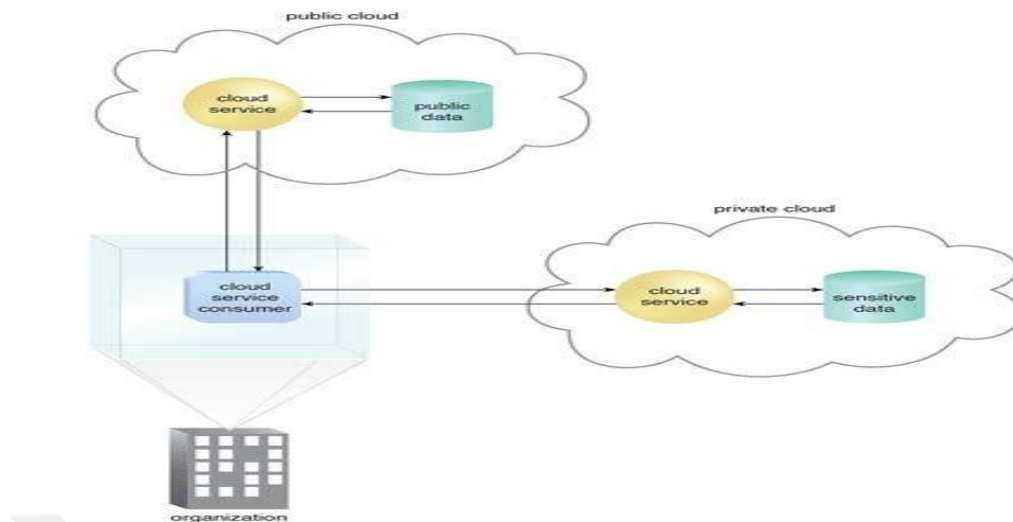


Figure 2.7: Hybrid Cloud [15]

#### 2.5.4. Community Cloud

As Figure 2.8 indicates, this cloud type is provisioned for a specific group of users or organizations. In this case, the infrastructure is provided by a specific consumers' community with common interests (security, mission policy, and compliance); therefore, one or more organizations own, manage, and operate it. Sometimes, a partnership or third party arrangement may also take responsibility of such an initiative. It works almost like a public cloud but it only serves a specific community. Community members not only share the responsibility of laying the foundation of the infrastructure and its evolution but also ensure that only allowed parties have access to it unless otherwise agreed [15].

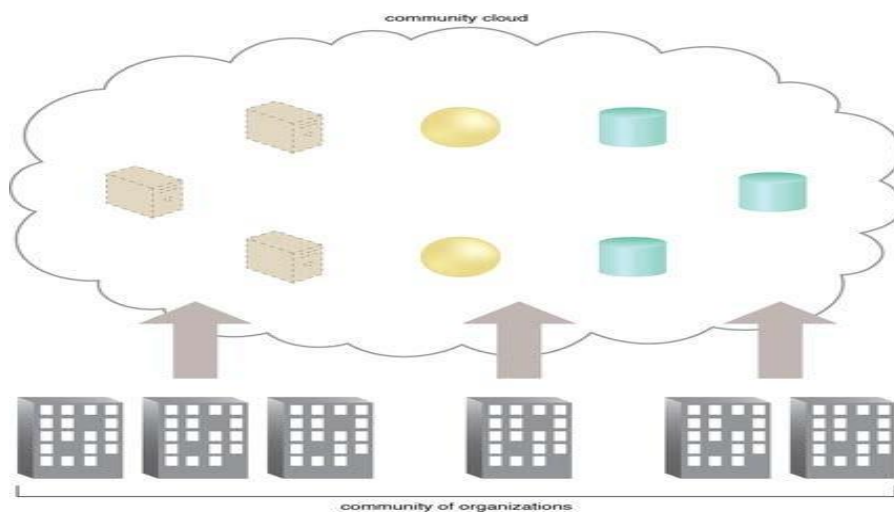


Figure 2.8: Community Cloud [15]

## 2.6. THE CLOUD DATA CENTER

It works like a web hosting service because the data center has all the data just like conventional web hosting; therefore, a virtual system is used for uploading files to the cloud. In this case, the resources remain in the data center. Google has impressive data centers Figure 2.9 [33].

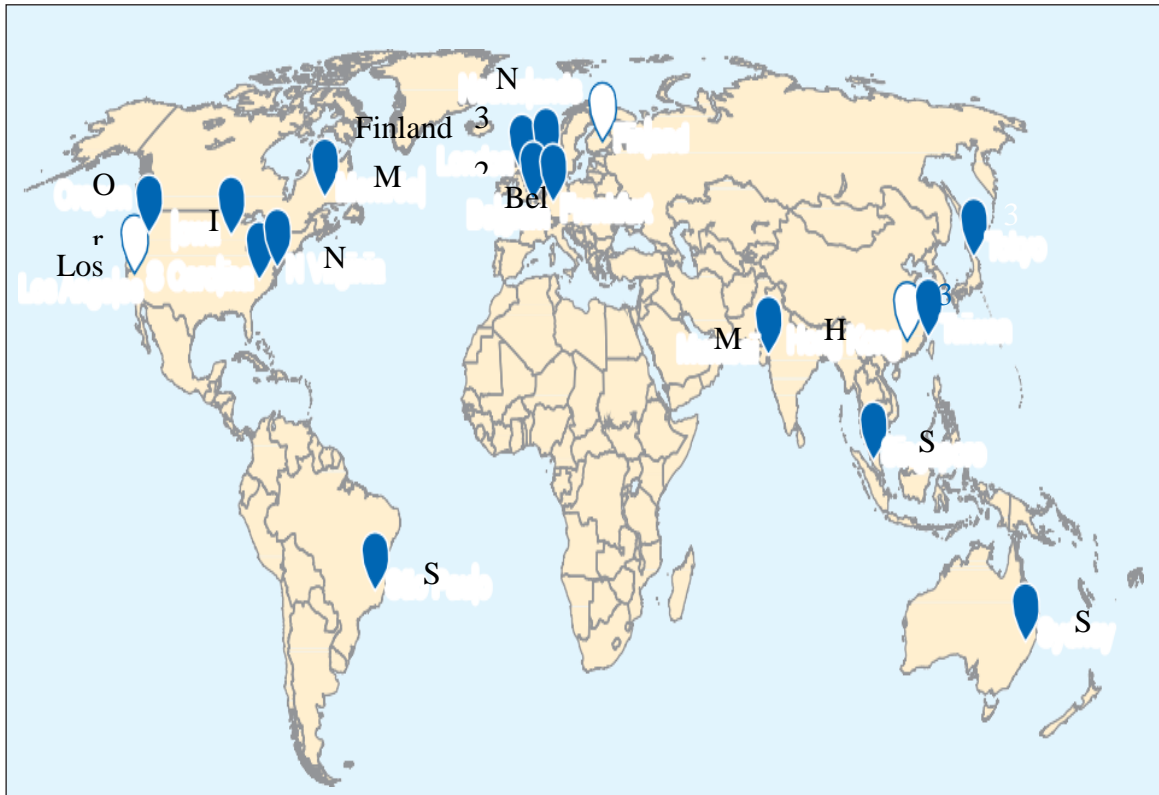


**Figure 2.9:** Google Data Center [33]

### 2.6.1. Data Center Locations

Location is very critical for data centers because a data center has a series of servers, which have all the resources in the distributed form rather saving everything in a single place. Some services offer saving the data in multiple locations; however, some services, for example, Compute Engine place the data in a single location, which means that the data center should be closer to the customers.

For choosing the right place, location choices should be clearly mentioned. Google Cloud has been operating data centers in 15 locations around the globe including Brazil, the United States, India, Western Europe, Australia, and East Asia [34]. See Figure 2.10.

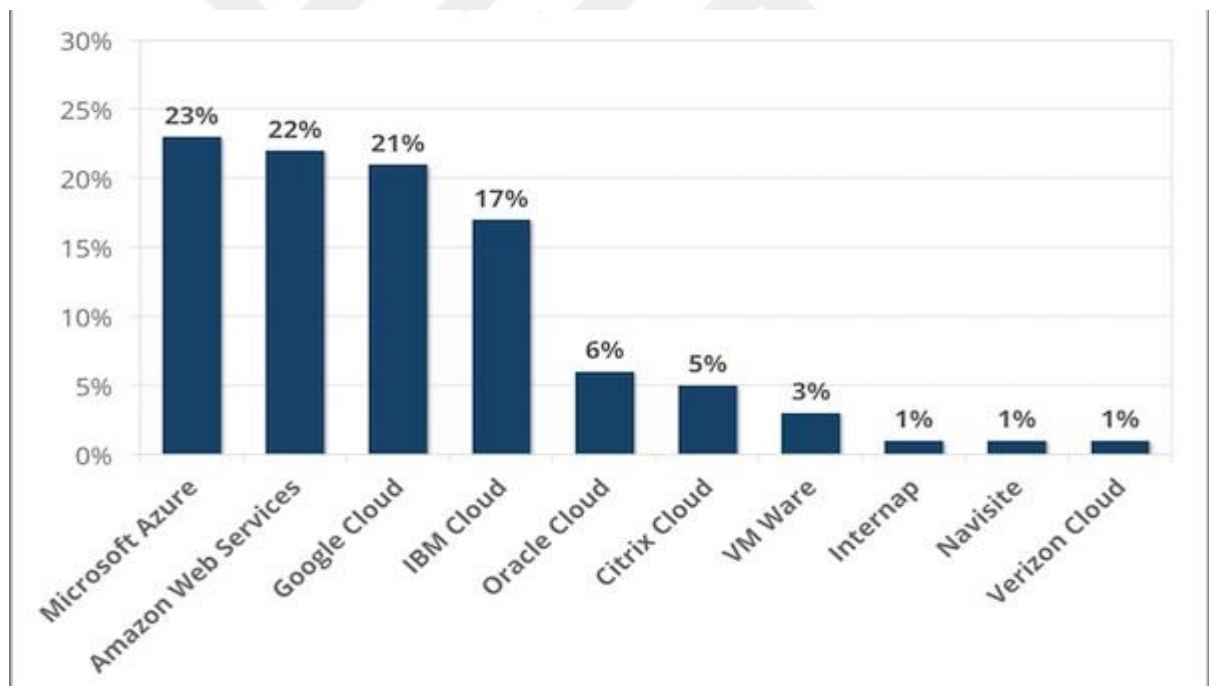


**Figure 2.10:** Google Data Locations Center [34]

### 3. CLOUD COMPUTING PROVIDERS

In year 2006, Amazon started AWS with virtual server EC2 or Elastic Compute Cloud as well as S3 or simple storage service [17]. Their success led to birth of many other cloud providers and cloud computing has evolved into new business models. According to the Gartner report published in June 2017, three main cloud providers included Microsoft (MS) Azure, Amazon Web Services (AWS), and Google Cloud Platform. Besides these, many other cloud providers have entered into this business such as Alibaba, IBM, Oracle, Rackspace, and Fujitsu are few to mention which have evolved into cloud computing businesses.

It can be seen in Figure 3.1 that Microsoft and AWS were the leaders while Google is the visionary.



**Figure 3.1:** Most Used Cloud Computing Services in the Enterprise Market. [17]

### **3.1. TOP TREE FAMOUS PROVIDERS**

In this section top three leading cloud computing providers are discussed.

#### **3.1.1 Amazon Web Services (AWS)**

In Dec 2017, AWS enjoyed the position of industry leader among cloud computing service providers. Amazon's AWS has a wide range of products offering more than any other cloud provider. AWS also has more customers than any other. According to reports, Amazon Web Service offers a scalable, reliable, and economical infrastructure that has powered several hundred thousand organizations in 190 countries [18]. Amazon's AWS claimed that their data centers were spread across 18 regions in the U.S., Europe, Japan, Singapore, Brazil, and Australia, ensuring that the customers from all industries get benefit out of cloud computing services. AWS claims to have advantages such as agility, low cost, openness, flexibility, and capability to ensure security of the data with multiple security standards. The AWS Cloud is operational in 49 zones located over 18 global locations, the company has announced to operate in further 12 availability zones located in Sweden, Bahrain, US, and Hong Kong. It is planning to commence another AWS GovCloud Region in the United States [19]. AWS has a broad range of cloud services in many product categories:

- Storage
- Computation
- Networking and content delivery
- Database
- Management tools
- Developer tools,
- Machine learning
- Security identity and compliance
- Analytics
- AR & VR
- Customer engagement
- Application integration
- Business productivity
- Internet of things

- Desktop and streaming apps
- Migration
- Media services
- Mobile services
- Game development
- Software and AWS cost management

### **3.1.2. Microsoft Azure**

It is world's second largest cloud server; however, it is gaining popularity amongst enterprises, as it has been originally a business software company. Azure cloud computing service was started in year 2010 and since then, continuous innovation and growth has lead the service to the leader's category in Gartner's magic quadrant back in June 2017. Microsoft Azure claims that 90 percent of Fortune 500 customers trust it. As per web definition of MS Azure, it is a comprehensive cloud service that IT professionals and developers utilize, build, and manage applications. Some integrated tools such as DevOps and online market support allow making anything from basic mobile applications to internet solutions [20]. Azure provides services in all three categories of SPI, i.e. SaaS, PaaS, and IaaS. At the same time, it supports several programming languages, frameworks, and tools for Microsoft and other organizations. Azure claims to have data centers in 42 locations, which is more than any other cloud provider. Azure has a huge list of products in the following categories:

- Networking
- Compute,
- Storage,
- Containers
- Web and Mobile,
- Databases,
- AI and cognitive services
- Data and Analytics
- Internet of things
- Security + Identity

- Enterprise Integration,
- Monitoring and management
- Developer's tools,

### **3.1.3. Google Cloud Platform (GCP)**

GCP has been built using the same top-class infrastructure, which Google has created and installed for services including Google search that has outperformed as compared to any other search engine in the world. Google is geographically widespread and advanced computer network. Google's main strength is its thousands of miles long fiber-optic cable operating with the help of advanced networking. It has cutting-edge services, which give speedy and consistent performance. It has state of the art fiber-optic cable network underneath the Pacific Ocean [21]. GCP helps software app developers by making them test, build or monitor their apps. Moreover, it helps system administrators to pay attention to software stack, and allows task outsourcing [21]. GCP started performing in 2011 just one year after Microsoft's launching cloud services. As per web definition of GCP, it is a cloud computing services suite, which has shared infrastructure with the remaining Google products, including Google Search and YouTube [21]. GCP also includes several cloud-based services but not as many as AWS and Azure have at the time of writing this thesis:

- Networking,
- Computation,
- Data Transfer,
- Storage and Databases,
- Big data,
- API platform and ecosystem,
- Management tools,
- Internet of things,

## **4. BIG DATA**

In its simplest term, big data has several voluminous data sets that traditional data processing handling systems are incapable to process. The term big data has been used since 1990s when John Mashey introduced the term. The International Telecommunications Union (ITU) has defined big data as “paradigm” that enables storage, collection, analysis, management, and visualization [16].

This term (big data) has not come from vacuum. It is the product of increasing research and exploration that has led to the accumulation of data for years. It is also known as unstructured data, which is difficult for traditional machines and memory to store, sort, process, and analyze. The obvious relationship between the reason for the existence of the mentioned form of data and the cloud computing services is that the cloud is capable of dealing with huge data. Data complexity is the most difficult thing for organizations and service providers because data sources are different and very difficult to control, which makes controlling this data a great challenge.

### **4.1. BIG DATA TECHNOLOGIES**

Cloud providers have their own reference architecture for analytical products, but in general, common approaches to majority of analytics projects are: Extract, Transform and Load (ETL) and Map Reduce. Extract process in ETL is about searching data in a database, which is collected through multiple sources. Transformation process is used to collected data in the extract phase, which is converted to the needed form; so, it can be shifted to some other database. Load process is conducted when the converted data is written in a target database [22]. Teradata corporation was first to analyze 1 TB of stored data on RDBMS in year 1992 when hard disk drives (HDD) had 2.5GB size. In 2017, Teradata claimed to have installed RDBMS of size over 50 Peta bytes. Traditionally, all the RDBMS has been practicing analytics methodology called as ETL. With the advent of big data, the requirement has evolved into “Map Reduce” architecture. Google published a paper in 2004 on the architecture of Map Reduce that has a parallel processing model to handle big data. Hadoop is based on Map Reduce architecture. Map step is a process, through which, analytical queries are separated and distributed in the parallel nodes, where the data are simultaneously processed. Reduce step gathers and delivers the queried data. Map reduce technology is most popular in big data analytics projects. [23]



## 4.2. BIG DATA PROPERTIES

- **Volume**

Volume implies the amount of generated and stored data, for example, on Facebook alone, over 10bn messages are daily sent, “like” button is clicked more than 4.5bn times and more than 350 million pictures are uploaded every day. Processing and storing such scale of humongous data isn’t possible through conventional database. Big data is an answer to such questions because it is possible by horizontally distributing the data sets i.e. in multiple parallel networked computers and processing these data sets through latest algorithms [16].

- **Velocity**

Traditionally, corporations analyzed data using batch process, where a chunk of data is submitted to a server for analysis and results were observed later. This approach is good when data input is slower than the batch processing job interval. To understand velocity, the example of Facebook says a lot. Facebook has approximately 250 billion images as of March 2018. Facebook users upload more than 900 million photos a day [16]. This data generation speed is termed as velocity.

- **Variety**

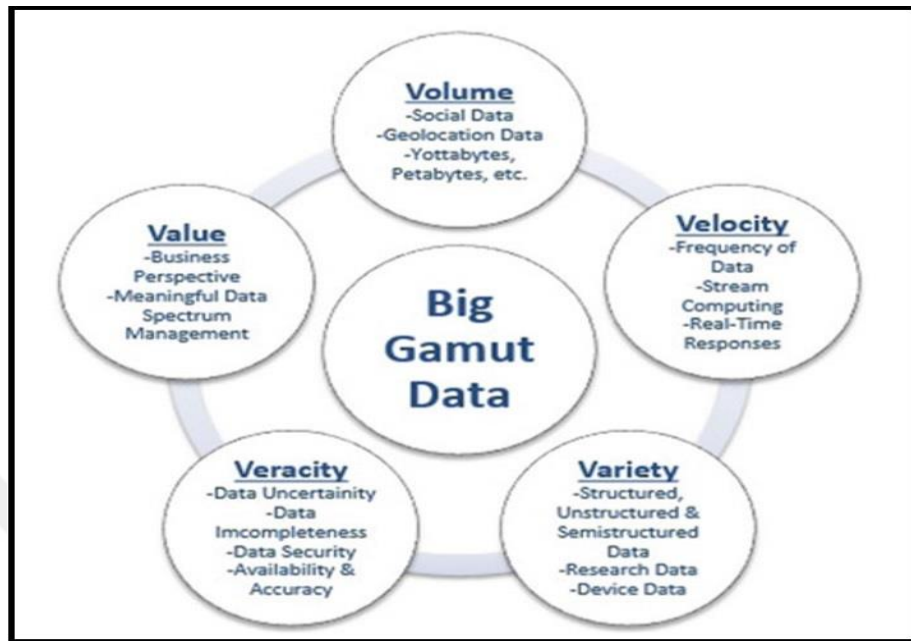
Variety means different data types in data science. Earlier, it used to be only structured data stores in tables with rows and columns. Financial data, supply chain, and ERP systems were all using RDBMS databases; however, using big data technology, it is possible to load and process data types such as images, audios, videos, JavaScript Object Notation (JSON) etc. [16].

- **Veracity**

Veracity refers to data authenticity. When data types are different, it is not easy to control data reliability and quality; however big data technologies make it possible.

- **Value**

Value is a very significant big data feature because data isn’t useful until value is created out of it. All four Vs: veracity, volume and velocity help deriving value or get an insight in the decision making. It is important to identify what value a project is going to achieve with big data analytics [16].



**Figure 4.1:** Different Big Data Properties [25]

### 4.3. BIG DATA ANALYTICS

The huge data gives economic value when it is analyzed. These analyses use a new set of technologies, which help dealing with large unstructured data. One way to discover this data is to develop old exploration methods. Having the ability to analyze this data gives an organization an advanced position and priority to respond to the world, and it helps making decisions when it is needed on demand. The world around us is built on the opportunity to take advantage of the decision, availableness of data and knowing how to analyze; it gives a person the motivation to be at the forefront and the individual to make the right decisions. The use of private cloud computing reduces risk and increases data security. Timely analyzing data helps creating the correct image, and making future predictions, as Figure 4.2 shows [25]. Big data analytics helps experts, users, and data scientists to analyze voluminous data taken from various data sources, and conventional business systems, which is otherwise difficult to handle. Conventional systems do not sufficiently perform because they don't have capacities to analyze several sources of data. Latest software is used to analyze big data; however, the unstructured data cannot be handled through traditional data warehousing techniques. The high processing needs of big data poorly

adjust with conventional data warehousing. Consequently, new and big data environment and handling techniques have evolved such as Hadoop, NoSQL and MapReduce. Such technologies are part of an open-source framework, which processes large data sets in a clustered system [24].



**Figure 4.2:** Advantages of big data analytics [23]

## **5. METHODOLOGIES**

### **5.1 INTRODUCTION**

The research requires the use of one of the cloud computing platforms, which allows data testing; so that we can demonstrate the practical side of this thesis. The data was loaded on one of the cloud computing platforms to experience rapid access to the cloud computing service. The procedure to create accounts on cloud will be explained later. We found that Google's enterprise data warehouse service, known as a BigQuery warehouse, is not free at all. It charges for data queries, storage, and streaming; however, exporting and loading are free services. The demo is based on data analysis of BigQuery datawarehouse, which was accessed through Google cloud platform. It works on petabyte scale. It is one of the fastest data warehousing solutions, which are available for analyzing big data. Our aim is to demonstrate the quick access of the cloud service.

In order to get the most accurate result, we conducted a survey to get the opinions of large numbers of internet users, regular cloud users, common people, and employees of different companies. We included large number of students in the survey to compile a large data, which helps us understand the users' views about using cloud computing techniques, especially in terms of big data analysis, also by this dataset. I have completed this thesis in conjunction with two cases studies mentioned earlier. It makes the thesis helpful to make the readers understand how to deal with big data in an easy way. Before starting the practical study, we studied big data analytics services offered by the three most popular cloud service providers.

### **5.2. AWS BIG DATA AND ANALYTICT PRODUCTS**

AWS is a multi-purpose cloud service. Amazon's Elastic Map Reduce (EMR) has been running Spark and Hadoop whereas Firehose and Kinesis make it possible to stream big data in AWS. Brief information about these product terms is given below:

- Hadoop/Apache Hadoop acts as open source software in order to store big data and to run applications.
- Spark/Apache Spark acts as an-open source cluster computing in a memory computing environment, which is used for running analytical applications.
- Kinesis Streams: Amazon's Kinesis collects, processes, and analyzes streaming data for timely insights and new information.

- Kinesis Firehose: It is the simplest method to stream data and load it in data stores as well as for analysis.
- Redshift: The scale of this service is in petabytes, and it compresses data for reducing costs.
- Amazon Elasticsearch uses the open-source Elasticsearch in AWS to analyze data with log-monitoring and click-through options. Kinesis analytics also supports it in the data analysis process.

AWS offers bigger data storage options and choices in comparison with Google. It is accomplished through gigantic AWS Simple Storage Service as well as DynamoDB, which is NoSQL database; Titan’s DynamoDB provides large storage that helps Titan graph database. Moreover, Apache HBase is also a massive NoSQL database that operates on petabyte level. Graph databases use graph structures that handles semantic queries that has edges, nodes, and properties for storing as well as representing data.

AWS offers a business intelligence (BI) service as well. QuickSight is a tool that is based on parallel in-memory processing for high speed, and it is also equipped with AWS Internet of Things (IoT) and Amazon Machine Learning, which link different devices to a cloud. Practically, it can connect billions of devices and handle trillions of messages. Another AWS service is called Amazon Machine Learning that provides services and tools to support AI applications. IoT is a cloud service, which assures secure and easy connection between devices and cloud applications’ secure interaction with cloud applications [26].

Table 5.1 shows details about various services offered on AWS cloud platform for big data analytical solutions. AWS has broad range of analytics services such as BI, data warehousing, stream and batch processing, data orchestration and machine learning [27].

**Table 5.1: AWS Big Data and Analytics Product List**

Services	Product Types	Descriptions
Athena	Server-less Query services	Easy to analyze Amazon S3 data with the help of standard SQL. Billing is based on just queries.

EMR	Hadoop	It's a managed framework for speedy and economical processing of large data.
Elastic Search	Elastic Search	It offers easy deployment, security, operations, and scaling for log analytics, complete text searches, and monitoring applications.
Kinesis	Streaming Data	It is the easiest method to use data streaming function on AWS.
Quick Sight	Business Analysis	It is a speedy, easy, and cloud-powered model that has very low cost as compared to traditional BI solutions.
Red Shift	Datawarehouse	It is a speedy, petabyte-scale warehousing service with simplicity and cost-effectiveness for analyzing data with the help of business intelligence software.
Glue	ETL	A service, which enables customers to gather and load data very easily before delivering it for analytical purposes
Data pipeline	DataWorkflow Orchestration	Processes and moves data among AWS computing and storage devices, and premise data sources.

### 5.3. MICROSOFT AZURE BIG DATA AND ANALYTICS PRODUCTS

In order to carry out analytics, Azure offers Data Lake Analytics that applies U-SQL language with C++, SQL, and HDInsight, which is a Hadoop service. Azure Stream Analytics service also

helps analyzing streaming data, which is a data catalog service, because it identifies data assets with the help of global metadata as well as data factory service that connects cloud data sources on-premises that manages data pipelines. Azure's big data storage service uses Data Lake Store, which is a Hadoop-based system. Cloud providers have broad general-purpose storage that includes SQL, StorSimple, and NoSQL database and storage services called as blob storage. Microsoft Azure uses the blob storage feature that helps processing unstructured data.

Azure is a powerful BI and machine learning platform, and it is lined up with AWS as well as some IoT Hub features. Azure IoT is a fully functional service offered by Microsoft Azure that provides secure and reliable two-way communication among the solution providing links and millions of IoT devices. Cloud platforms include a search engine that facilitates searches. Cognitive service providers and Microsoft's Cortana Suite provide several advanced intelligence options [26]. Table 5.2 shows Microsoft Azure Big Data and Analytics product list that includes popular analytical services [26].

**Table 5.2:** Microsoft Azure Big Data and Analytics Product List

Services	Product Types	Descriptions
HD Insight	Hadoop	Azure HDInsight is Hadoop-based service that offers Apache Hadoop in the cloud environment. Getting the value out of big data is possible in this case because it provides a platform and manages all sorts of data.
Stream Analytics	Streaming Data	Azure Stream Analytics operates as an event-processing engine, which helps gaining information from sensors, devices, cloud,

		and about data properties.
Azure Bot Service	Server-less Bot Service	This service helps developing rapid intelligence using the power of Azure Functions and Microsoft Bot Framework.
Data Lake Analytics	Analytics	This analytics service is comparatively a new service, which is based on Apache YARN. It has dynamic scales which help a user focusing on business goals.
Data Lake Store	Repository	This store service provides opportunities to capture data of different sizes, types and speeds with changes in applications.
Data Factory	Data Transformation and Movement	It helps a user produce trusted information out of raw data.
Power BI Embedded	Data visualization	Power BI Embedded is a service that provides completely interactive data for customers; so it saves time and cost for customers.
Data catalogue	Enterprise Data Assets	It is a catalog service, which registers the data for system that recovers enterprise data.
Log Analytics	Analytics	This service assists a user to gather, correlate and visualize machine data, including event



		and network logs, and performance data among other data forms.
Text Analytics API	Cognitive Service	It evaluates data and topics according to the needs of the users.
Azure Analysis service	Analytics	It analyzes data and topics to understand according to the users' choices.
Custom Speech service	Cognitive Service	It is equipped with speech recognition mechanism, and handles complex audio data such as speaking style vocabulary and background noise.
Event hubs	Messaging	Azure Event Hubs support elastic telemetry and event ingestion, and offers durable buffering as well as end-to-end latency to connect millions of events and devices.
SQL Data Warehouse	Datawarehouse	It works as an elastic data warehouse that has enterprise-grade features, which work with massive processing architecture.

#### 5.4. GOOGLE CLOUD PLATFORM'S BIG DATA AND ANALYTICS PRODUCTS

Google has a BigQuery data service, which applies SQL-like interface. It is appropriate for many users because it is easy-to-learn. It is equipped to handle petabyte databases, which streams data at 100,000rows/second. BigQuery allows geographic replication; so, users have the choice to select the location of their data.

BigQuery is in fact a “pay-as-you-go” kind of service that doesn’t have any dedicated infrastructure that requires Google to use several processors for maintaining speedy query time when it is integrated with Hadoop, Spark, Hive and Pig. Pig offers high-level platforms to create programs, which work with Hadoop. Since Hive is a data warehouse overtop Hadoop, which reads, writes and manages huge datasets, which are stored using SQL. Enterprises can choose DoubleClick and Google Analytics for advertising mainly to collect statistics in order to feed them to BigQuery. Google Cloud Dataflow helps creating sequences of cloud data services. Google Big Data services include Cloud Datastore, which stores non-relational data. Cloud BigTable acts as a scalable NoSQL type of database while Cloud Machine Learning provides with ancillary tools and machine learning, which helps speech recognition and translation.

Experts believe that Google lacks Graphical Processing Unit (GPU) to handle big data. Drafting GPU codes for data analysis has high-value because GPU offers incredible increase in performance. Google lacks GPU, which is quite strange even when AWS and Azure has been offering this cutting-edge feature for many years [28].

GCP Big data and Analytics product list: Google is third in Gartner’s magic quadrant and its analytics service solutions are captured in the table 5.3 given below [29]:

**Table 5.3:** Google Cloud Platform Big Data and Analytics Product List

Service	Product Type	Description
BigQuery	Datawarehouse	Google's has low-cost and managed analytics datawarehouse “BigQuery” works server-less; so, infrastructure isn’t needed to

		manage storage, capacity or database administrator.
Cloud Data Flow	Stream Analytics, ETL	This programming model serves as a data management and analytics service that processes data and performs ETL, streaming analytics and batch processing.
Cloud Dataproc	Hadoop and Spark	It is a Hadoop and Spark-based service that easily processes big data with the help of open and powerful tools of Apache Systems.
Cloud Datalab	Data Analysis	It works like an interactive notebook that explores, collaborates, analyzes and visualizes the given data.
Data Studio	Visualization	This tool transforms data into reports and dashboards, which are easily read, shared, and customized.
Dataprep	Data Preparation Service	It is a high-profile and intelligent data preparation service that explores, cleans, and prepares both unstructured as well as structured data before it is analyzed.
Pub/Sub	Serverless Messaging service	This is a reliable and large-scale serverless messaging service, which permits sending and

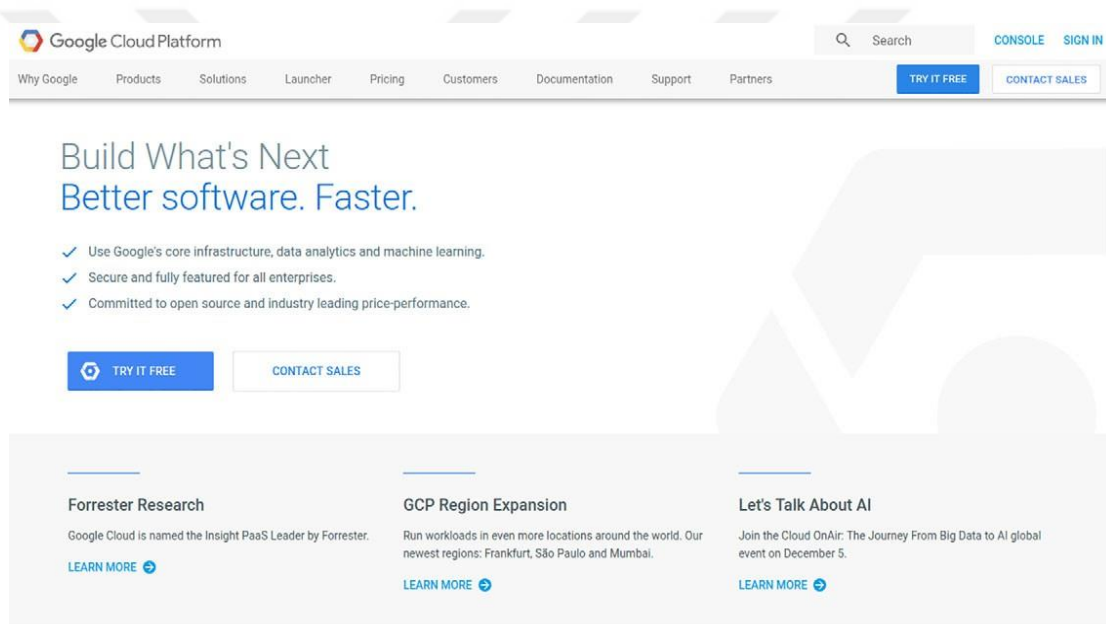
		receiving messages among different applications.
--	--	--

## 5.5. GETTING STARTED WITH GOOGLE CLOUD PLATFORM

After learning about cloud and Google Cloud Platform, we'll review and explore GCP.

### 5.5.1. Signing Up For GCP

Like most of the Google services, a Google Cloud user signs up for an account; however, Google account holders and Gmail users can directly login after subscribing for cloud account [30].

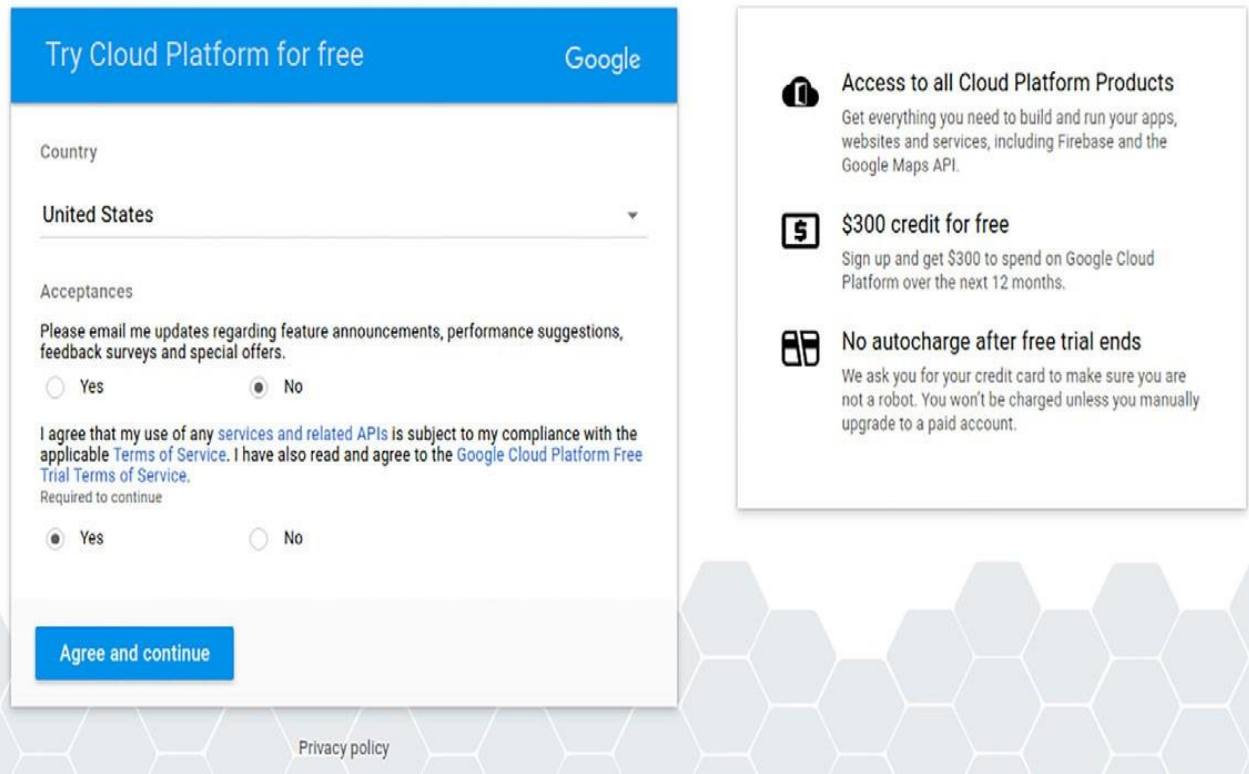


### Why Google Cloud Platform?



**Figure 5.1:** Google Cloud Platform [30]

A user first needs to log on to <https://cloud.google.com>, and subscribes it or try the free service that is possible through familiar Google sign-in procedure. Non-Google users need to create a new account. After approving free trial, a page prompts users to enter their billing information. The free trial worth \$300 (Figure 5.2) allows a Google Cloud user to explore almost everything. Moreover, some Google Cloud products are free usage possibilities. The given exercises remind a user to switch off resources after finishing the exercises [30].



Try Cloud Platform for free Google

Country

United States

Acceptances

Please email me updates regarding feature announcements, performance suggestions, feedback surveys and special offers.

Yes  No

I agree that my use of any [services and related APIs](#) is subject to my compliance with the applicable [Terms of Service](#). I have also read and agree to the [Google Cloud Platform Free Trial Terms of Service](#).  
Required to continue

Yes  No

Agree and continue

Privacy policy

- Access to all Cloud Platform Products**  
Get everything you need to build and run your apps, websites and services, including Firebase and the Google Maps API.
- \$300 credit for free**  
Sign up and get \$300 to spend on Google Cloud Platform over the next 12 months.
- No autocharge after free trial ends**  
We ask you for your credit card to make sure you are not a robot. You won't be charged unless you manually upgrade to a paid account.

**Figure 5.2:** Free Trial of Google Cloud Platform [31]

### 5.5.2. Steps for creating free GCP Trail Account

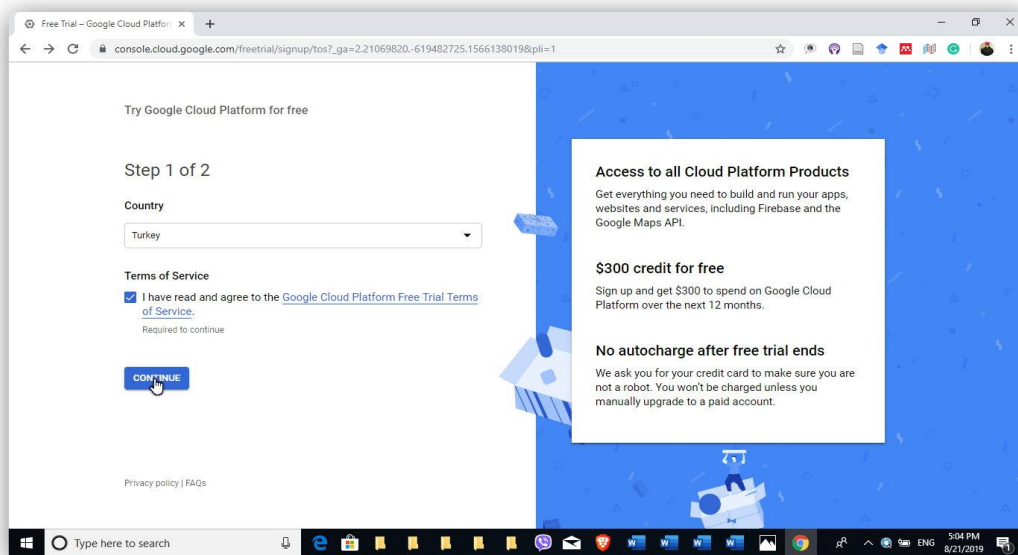
The flow introduces the procedure for creating Google cloud account in BigQuery environment, and how to launch queries on publicly available datasets in BigQuery [28].

**1** - Steps to create Google cloud account: The following steps are performed to sign up for a Google Cloud Account free-of-cost:

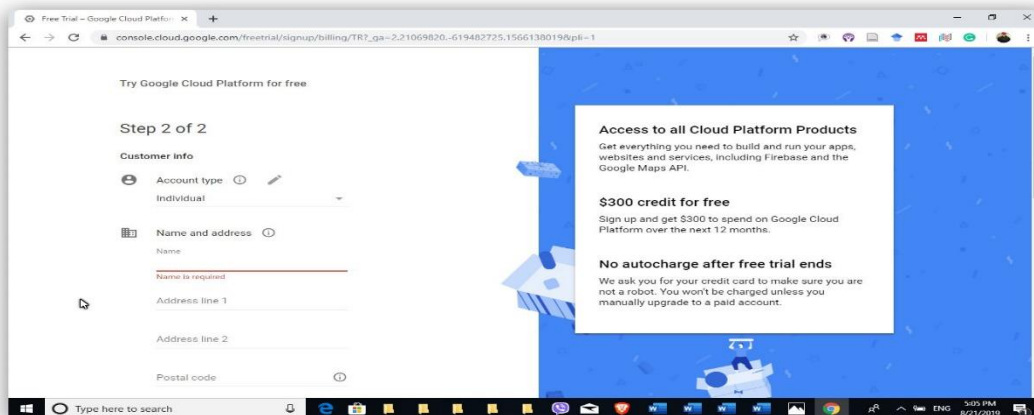
- Log on to <https://cloud.google.com/>
- Find and click the option: Try it free.
- Sign up a Gmail account
- Log in to the Gmail account
- Start with the free trial
- Enter country if not selected by default

- Accept the service terms
- When the customer information page will appear, enter all details including name and address
- Enter payment method, preferably it accepts credit card
- Click Start my free trial

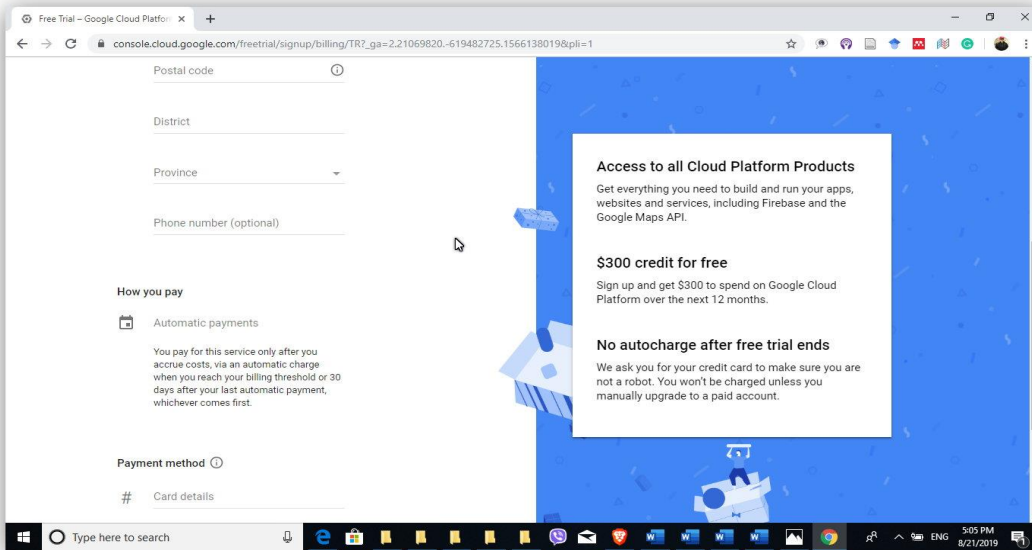
It is obvious in figures below how select the country and agree to GCP policy (Figure 5.3), after which, Figure 5.4 shows how to enter the personal details while Figure 5.5 shows how to enter the payment details. Finally, Figure 5.6 shows how to start your trial version [31].



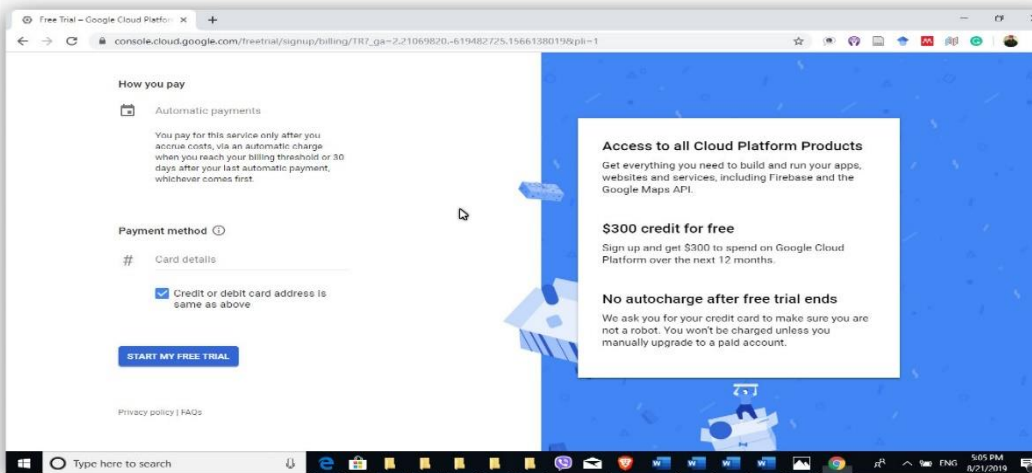
**Figure 5.3:** Select the country and agree to the policy of GCP [31]



**Figure 5.4:** Enter Personal Details [32]



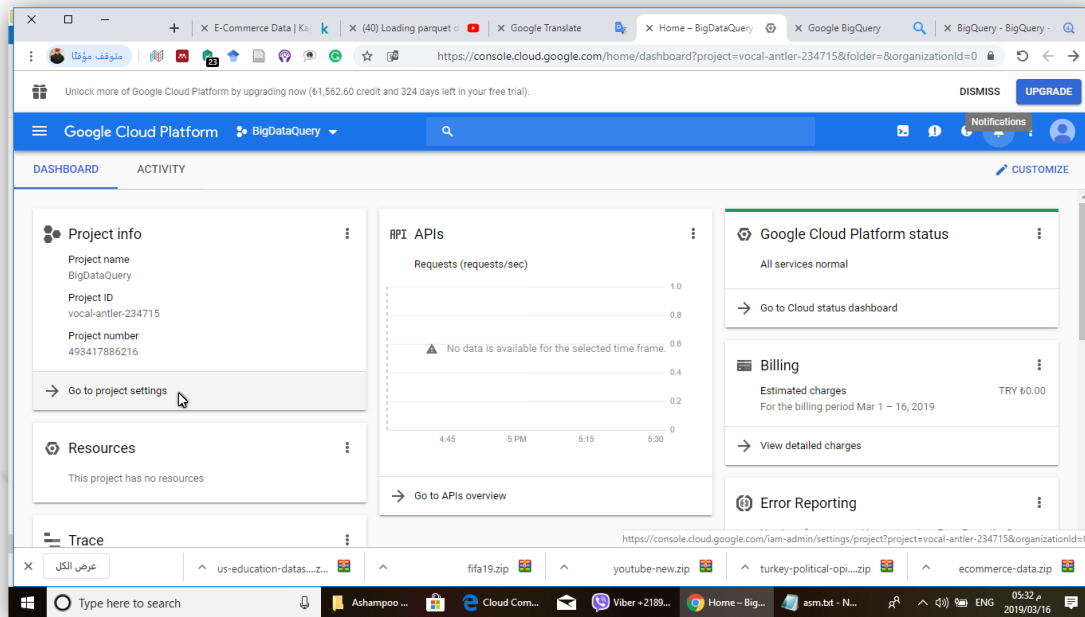
**Figure 5.5:** Enter the Payment Details [31]



**Figure 5.6:** Submit and Start the trial version [31]

### 5.5.3. Exploring the Console

When the sign up process is over, a subscriber automatically goes to Cloud Console (Figure 5.7), which automatically creates a new project that contains a user's work. The resources of a project are clearly sifted apart from the resources of other projects. Categories are mentioned on the left side of the page, which correspond to different services that Google Cloud Platform offers and project-specific configuration sections (such as billing, permissions, and authentication)[29].



**Figure 5.7:** Google Cloud Console [30]

#### 5.5.4. Installing the SDK

When a user becomes familiar with Google Cloud Console, the next step is installation of Google Cloud SDK, which helps building software applied on Google Cloud, and tools to manage production resources. Normally operating the Cloud Console should be done with Cloud SDK/gcloud. For installing SDK, a user should log on to <https://cloud.google.com/sdk/>, and follow instructions [31].

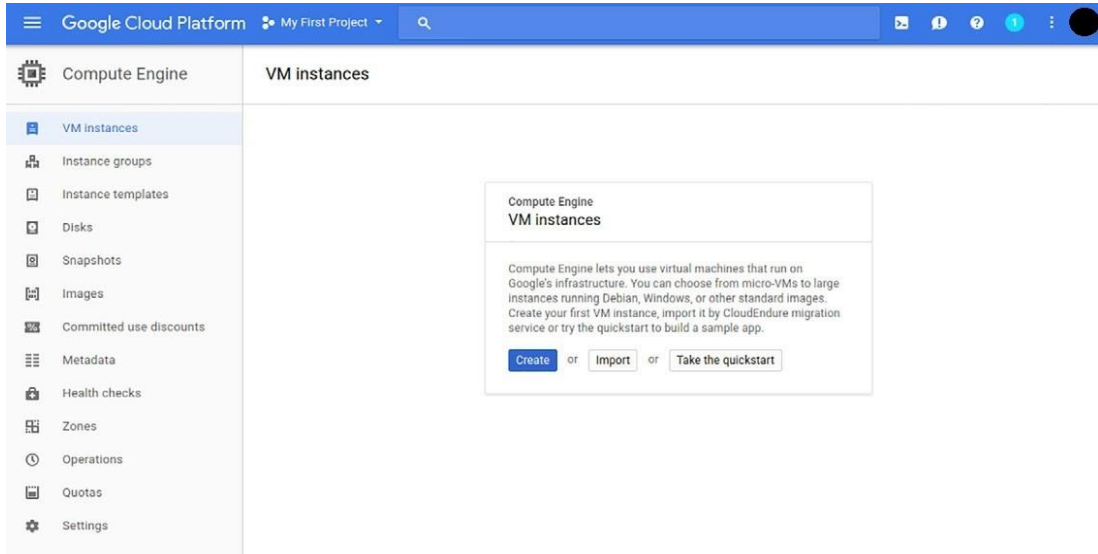
### 5.6. INTERACTING WITH GCP

After signing up and getting familiar with console, it is the right time to accomplish a practice in different for interacting with the GCP. A virtual machine needs to be launched in the cloud, after which, JavaScript is written so as to terminate the virtual machine [30].

#### 5.6.1. In the browser: The Cloud Console

A user should start with Google Console's Compute Engine area by clicking the compute section in order to expand it. After that, click the Compute Engine link, which shows up. After the first click, Google starts Compute Engine. When that is accomplished, Create button will appear that brings a user to a page, which is shown in Figure 5.8. It helps configuring the virtual machine [29].





**Figure 5.8:** Google Cloud Console (create a new virtual machine) [32]

A form, given in Figure 5.9, will appear on the next page that helps configuring the details; so now, we should look at the options [31].

**Figure 5.9:** Defining your virtual machine [32]

## **5.7. BIG QUERY**

If you deal with a lot of data, you probably remember the frustration of sitting around for a few minutes (or hours, or days) waiting for a query to finish running. At some point, a user may have looked at MapReduce (for example, Hadoop) to speed up some of the larger jobs and then feel frustrated again when every little change meant that a user had to change the code, recompile, redeploy, and run the job again. This leads us to BigQuery [35].

### **5.7.1. What is BigQuery?**

BigQuery is a relational-style cloud database that is capable of querying enormous amounts of data in seconds rather than hours. Since BigQuery uses SQL instead of Java or C++ code; so, exploring large data sets is both easy and fast. A user can run a query, tweak it a bit if it's not quite what the user wanted, and run the query over again. It's important to remember the analytical nature of BigQuery. Although BigQuery is capable of running traditional OLTP-style queries (for example, UPDATE table SET name = 'Jimmy' where id = 1), it's the most powerful when a user uses it as an analytical tool for scanning, filtering, and aggregating lots and lots of rows into some meaningful summary data [36].

### **5.7.2. Why BigQuery?**

You may understand what BigQuery is and what it's used for, but a user may be confused about why he/she uses BigQuery instead of some other systems. For example, using MySQL is suitable to explore the data. MySQL can be used in most cases, but for that, the user has to scan more and more data, and besides, MySQL will become overloaded, and performance will degrade. When that happens, it makes sense to start exploring other options.

First, a user might try to tune MySQL's performance-related parameters so that certain queries run faster. Then, read-replicas should be turned on; so that a user runs super-difficult queries on the same database that handles user-facing requests. Next, data warehouse system like Netezza should be used, but the price for those systems can be high (usually millions of dollars), which could be more than a user's willingness to pay.

This is exactly where BigQuery can come in to save the day. Committing to the promise of understanding cloud infrastructure, BigQuery provides some of the power of traditional data warehouse systems only paying for what a user uses. Let's take a quick look at how it works under the hood to understand why BigQuery can handle scenarios where something like MySQL may struggle [36].

## **5.8. CASE STUDY**

This case study has two sections including a survey on cloud computing usage for analyzing data, while the other section shows a demo project on how to conduct data analysis on the platform of the famous service of Google cloud provider.

### **5.8.1. Survey**

#### **5.8.1.1. Data survey collection**

Before starting a demonstration that represents data analysis on a cloud computing platform, an additional step has been taken to support the importance of writing this article, which explains how large data is analyzed. This step is a survey, in which, data were collected from 102 people, most of whom were students, engineers, and computer users. Questions focused on the availability of technological information to people using information technology, especially concerning big data. This was done with thirty questions using the checklist, checkbox and rating scale. The internet-based questionnaire has been adopted as one of the new ways to gather and compile data. This means that all the responses were received and responded to each person via internet through a constantly distributed link, which is:

[https://docs.google.com/forms/d/e/1FAIpQLSe01IeCAqTOsypYYLzpUHgyqTTInwh08Hsp4bR1BsDDZ3g-A/viewform?usp=sf\\_link](https://docs.google.com/forms/d/e/1FAIpQLSe01IeCAqTOsypYYLzpUHgyqTTInwh08Hsp4bR1BsDDZ3g-A/viewform?usp=sf_link) That questionnaire was compiled using Google Drive forms, which provides great and free possibilities in this aspect of the analysis, the extraction of results and the overall analysis. To obtain the best results for analyzing the data set obtained from the questionnaire, the statistical analysis program provided by IBM (SPSS) was used. Data and opinions were entered for 357 respondents and their responses were obtained through the above link via the Internet that was designed and prepared using Google forms on the Web.

Before we start viewing the results, take a quick look at the SPSS statistical analysis program.

- **What is SPSS ?**

SPSS (currently officially “IBM® SPSS® Statistics”) is a commercially distributed software suite for data management and statistical analysis and the name of the company originally developing and distributing the program. Introduced in 1968, it helped revolutionize research practices in the social sciences, enabling researchers to conduct complex statistical analyses on their own. Presently, Windows, Mac and Linux versions of SPSS are available with major version updates released every one to two years. SPSS is a comparably easy-to-handle statistics program providing commonly used procedures. As such, it is widely used in academia including communication studies, although facing increasingly tough competition from more comprehensive and free open-source software[39].

- **Why is SPSS good?**

SPSS is a good program to rely on in analysis, especially digital analysis, because of its many advantages[40].

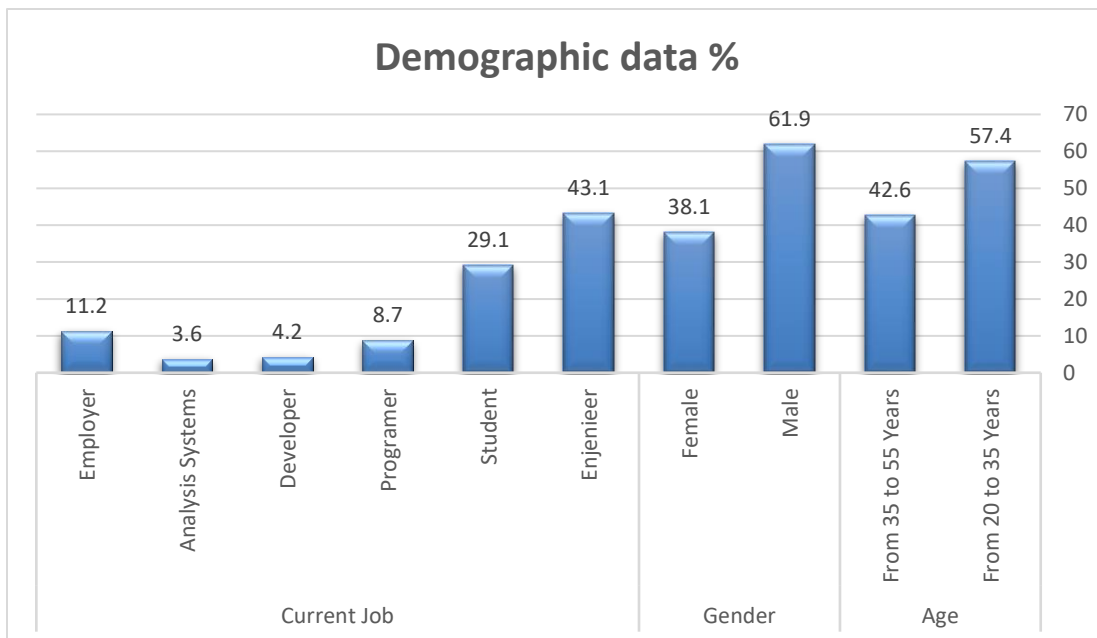
- 1- it can be expanded.
- 2- it can easily and quickly display data tables.
- 3- many complex statistical tests are available in it.
- 4- support a wide variety of charts and graphs.
- 5- it has the ability of data management.
- 6- it has the ability to manage a complex data set.

#### **5.8.1.2. Data survey analysis using ibm spss**

The Survey has been designed to obtain data from the participants who proved their participation by completing the questionnaire and answering all the questions. They answered all the questions according to their knowledge and personal experience. The following tables display answers, which were filled according to the data obtained from the survey, as shown in Table 5.4. The percentage of males and females, who participated in the questionnaire, was about 61.9% male and 38.1% female, and the large percentage of participants' ages ranged between 20 and 35 years. It should also be noted that the two largest groups that participated were engineers and students. This gives a good indicator of the reliance on these results as scientific results with a significant percentage, which is clearly shown in the Table 5.4. The Demographic Data have shown clearly in the figure (5.10)

**Table 5.4: Demographic Data**

Demographic Variable		Count	Percent %
1 - Age	From 20 to 35 Years	205	57.4
	From 35 to 55 Years	152	42.6
2 - Gender	Male	221	61.9
	Female	136	38.1
3 - Current Job	Enjenieer	154	43.1
	Student	104	29.1
	Programer	31	8.7
	Developer	15	4.2
	Analysis Systems	13	3.6
	Employer	40	11.2



**Figure 5.10: Showing The Demographic Data**

To find out the general direction of the respondents' answers to the Survey, the research questions were divided into three sections, the first one contained the 10 questions and the second section included the 9 questions. The last part of the questions was contained 10 questions. The questions of the three sections used a 5-point Likert Scale. The general trend will be determined later after the results are shown after analysis. in table 5.5 is shown the first result of the first section Survey questions followed by the commentary on the results of the analysis.

**Table 5.5:** The result of analysis for section I

Questios		Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree	Mean	Std. Deviation	Rank
Q1: I have a clear and comprehensive concept of data and types	N	6	9	31	224	87	1.94	.762	9
	%	1.7 %	2.5 %	8.7 %	62.7 %	24.4 %			
Q2: Data is the new oil for most system analysts	N	10	22	20	215	90	2.01	.899	8
	%	2.8 %	6.2 %	5.6 %	60.2 %	25.2 %			
Q3: Data is the virtual representation of the human being either text or image or voice or movement	N	11	18	39	234	55	2.15	.850	5
	%	3.1 %	5.0 %	10.9 %	65.5 %	15.4 %			
Q4: Traditional memory has limited ability to analyze data	N	13	22	43	239	40	2.24	.866	1
	%	3.6 %	6.2 %	12.0 %	66.9 %	11.2 %			
Q5: Structured data are easily manipulated and analyzed by Traditional memory	N	10	25	45	230	47	2.22	.863	2
	%	2.8 %	7.2 %	12.6 %	64.4 %	13.2 %			
Q6: UnStructured data is the one that we can not deal with and we need more sophisticated systems for analysis	N	9	22	37	240	49	2.17	.830	4
	%	2.5 %	6.2 %	10.4 %	67.2 %	13.7 %			
Q7: It is worrying that the world produces large amounts of data per day up to 4.4 zettabytes	N	5	17	45	226	64	2.08	.785	7
	%	1.4 %	4.8 %	12.8 %	63.3 %	17.9 %			
Q8: The accumulation of data is due to the wide spread of Social media	N	7	17	50	222	61	2.12	.815	6
	%	2.0 %	4.8 %	14.0 %	62.2 %	17.1 %			
Q9: The real value of data is the ability to analyze it	N	3	20	44	240	50	2.12	.742	6
	%	0.8 %	5.6 %	12.3 %	67.2 %	14.0 %			

Q10: Data collection and analysis are currently important Jobs in the world	<b>N</b>	6	17	58	231	45	2.18	.774	3
	<b>%</b>	1.7 %	4.8 %	16.2 %	64.7 %	12.6 %			
<b>Weighted Mean</b>							<b>2.1108</b>		
<b>Std. Deviation</b>							<b>.50215</b>		

Table 5.5 shows ( What is the relationship between data analysis and its large size? )

From which we find that the highest was awarded to the Q4: ( Traditional memory has limited ability to analyze data ) with mean 2.24 and Std.deviation .866 , followed by Q5: ( Structured data are easily manipulated and analyzed by Traditional memory ) with mean 2.22 and Std.deviation .863 , followed by Q10 ( Data collection and analysis are currently important Jobs in the world ) with mean 2.18 and Std.deviation .774 , followed by Q6 ( UnStructured data is the one that we can not deal with and we need more sophisticated systems for analysis ) with mean 2.17 and Std.deviation .830 , followed by Q3 ( Data is the virtual representation of the human being either text or image or voice or movement ) with mean 2.15 and Std.deviation .850, with Strongly Disagree by percent ( 11.2% , 13.2% , 12.6% , 13.7% , 15.4% , respectively ) and Disagree by percent ( 66.9% , 64.4% , 64.7% , 67.2% , 65.5% , respectively ) .

With the lowest average was awarded to the Q1: ( I have a clear and comprehensive concept of data and types) with mean 1.94 and Std.deviation .762 , followed by Q2: ( Data is the new oil for most system analysts) with mean 2.01 and Std.deviation .899 , followed by Q7 ( It is worrying that the world produces large amounts of data per day up to 4.4 zettabytes) with mean 2.08 and Std.deviation .785 , followed by Q8 ( The accumulation of data is due to the widespread of Social media ) with mean 2.12 and Std.deviation .815 , followed by Q9 ( The real value of data is the ability to analyze it ) with mean 2.12 and Std.deviation .742, with Strongly Disagree by percent ( 24.4% , 25.2% , 17.9% , 14.0% , 12.6% , respectively ) and Disagree by percent ( 62.7% , 60.2% , 63.3% , 62.2% , 68.2% , respectively ) .

The weighted average of section one ( I ) was 2.1108 with Std.deviation .50215 which indicate that the trend of the relationship between data analysis and its large size is ( Disagree ) as a general trend according to 5-point Likert Scale as shown in Table ( 5.5 ) since 2.1108 lie in the interval. [ 1.81 – 2.60 ] .

So the average of the relationship between data analysis and its large size is 2.1108 which consider a low level, since the intervals of level as follow:

Low Level [ 1.181 - 2.60].

Moderate Level [ 2.60 - 3.39].

High Level [ 3.40 - 5].

This is results shown clearly in the below figure ( 5.11 )

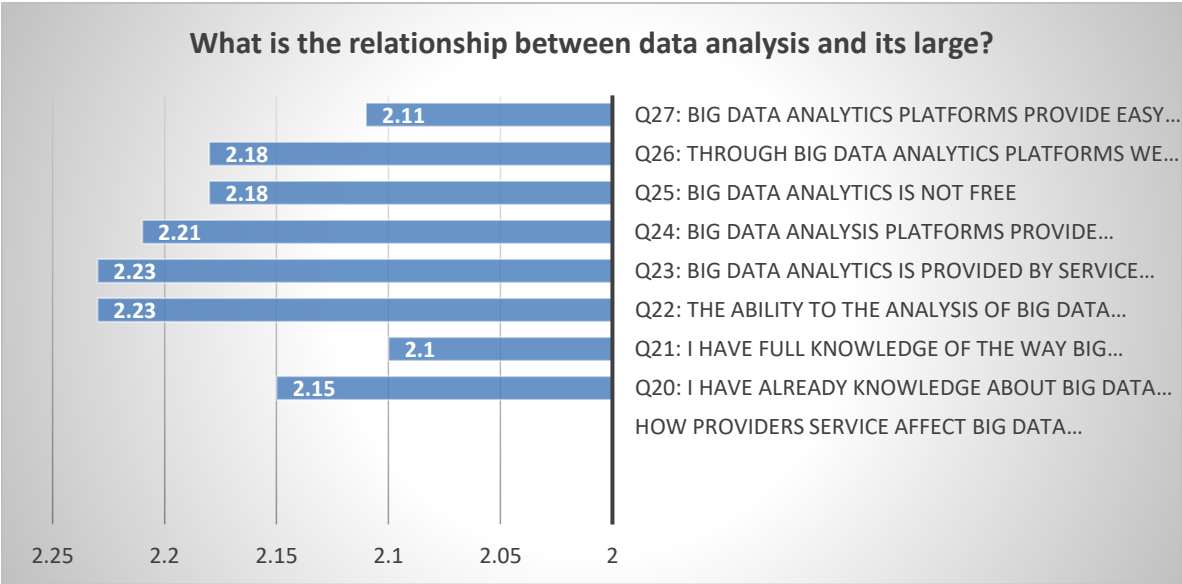


Figure 5.11: Showing The Results of Section I



The following table 5.6 presents the results of the second section of the questions followed by the comment on the results observed in the table with a graph showing the results of the second part of the questions.

**Table 5.6:** The result of analysis for section II

Questios		Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree	Mean	Std. Deviation	Rank
Q11: I have a clear and comprehensive concept of cloud computing	N	2	14	42	254	41	2.13	.724	6
	%	1.7 %	3.9 %	11.8 %	71.1 %	11.5 %			
Q12: I have full knowledge of the way cloud computing works	N	2	23	45	234	50	2.16	.792	3
	%	1.4 %	6.4 %	12.6 %	65.5 %	14.0 %			
Q13: The platforms offered by cloud computing are easy and simple to use	N	1	19	55	209	73	2.06	.771	7
	%	0.3 %	5.3 %	15.4 %	58.5 %	20.4 %			
Q14: The Solutions provided by cloud computing to system analysts and developers are satisfactory	N	3	19	47	247	41	2.15	.717	4
	%	0.8 %	5.3 %	13.2 %	69.2 %	11.5 %			
Q15: Cloud computing services reduced the cost of building applications	N	3	15	54	245	40	2.15	.697	4
	%	0.8 %	4.2 %	15.1 %	68.6 %	11.2 %			
Q16: Cloud computing has provided many of the Operational resources that system developers need to launch their applications	N	3	20	54	228	52	2.14	.760	5
	%	0.8 %	5.6 %	15.1 %	63.9 %	14.6 %			
Q17: Cloud computing provides a secure environment for its users	N	9	24	55	225	44	2.24	.847	1
	%	2.5 %	6.7 %	15.4 %	63.0 %	12.3 %			
Q18: Cloud computing provided platforms for big data analysis	N	5	22	43	231	56	2.13	.797	6
	%	1.4 %	5.3 %	12.0 %	64.7 %	15.7 %			
Q19: Cloud computing will gradually replace all traditional systems in the future	N	9	21	47	230	50	2.18	.838	2
	%	2.5 %	5.9 %	13.2 %	64.4 %	14.0 %			

<b>Weighted Mean</b>	<b>2.1481</b>
<b>Std. Deviation</b>	<b>.48834</b>

Table 5.6 shows ( How does cloud computing affect data analysis? ) From which we find that the highest was awarded to the Q17: ( Cloud computing provides a secure environment for its users ) with mean 2.24 and Std.deviation .847 , followed by Q19: ( Cloud computing will gradually replace all traditional systems in the future ) with mean 2.18 and Std.deviation .838 , followed by Q12 ( I have full knowledge of the way cloud computing works) with mean 2.16 and Std.deviation .792 , followed by Q15 ( Cloud computing services reduced the cost of building applications ) with mean 2.15 and Std.deviation .697 , followed by Q14 ( The Solutions provided by cloud computing to system analysts and developers are satisfactory ) with mean 2.15 and Std.deviation .717, with Strongly Disagree by percent ( 12.3% , 14.0% , 14.0% , 11.2% , 11.5% , respectively ) and Disagree by percent ( 63.0% , 64.4% , 65.5% , 68.6% , 69.2% , respectively ). With the lowest average was awarded to the Q13: (The platforms offered by cloud computing are easy and simple to use) with mean 2.06 and Std.deviation .771 , followed by Q11: ( I have a clear and comprehensive concept of cloud computing ) with mean 2.13 and Std.deviation .724 , followed by Q18 ( Cloud computing provided platforms for big data analysis) with mean 2.13 and Std.deviation .797 , followed by Q16 ( Cloud computing has provided many of the Operational resources that system developers need to launch their applications ) with mean 2.14 and Std.deviation .760 , with Strongly Disagree by percent ( 20.4% , 11.5% , 15.7% , 14.6% , respectively ) and Disagree by percent ( 58.5% , 71.1% , 64.4% , 63.9% , respectively ).

The weighted average of section two ( II ) was 2.1481 with Std.deviation .48834 which indicate that the trend of the cloud computing effect on data analysis is ( Disagree ) as a general trend according to 5-point Likert Scale as shown in Table ( 5.6 ) since 2.1481 lie in the interval. [ 1.81 – 2.60 ] .

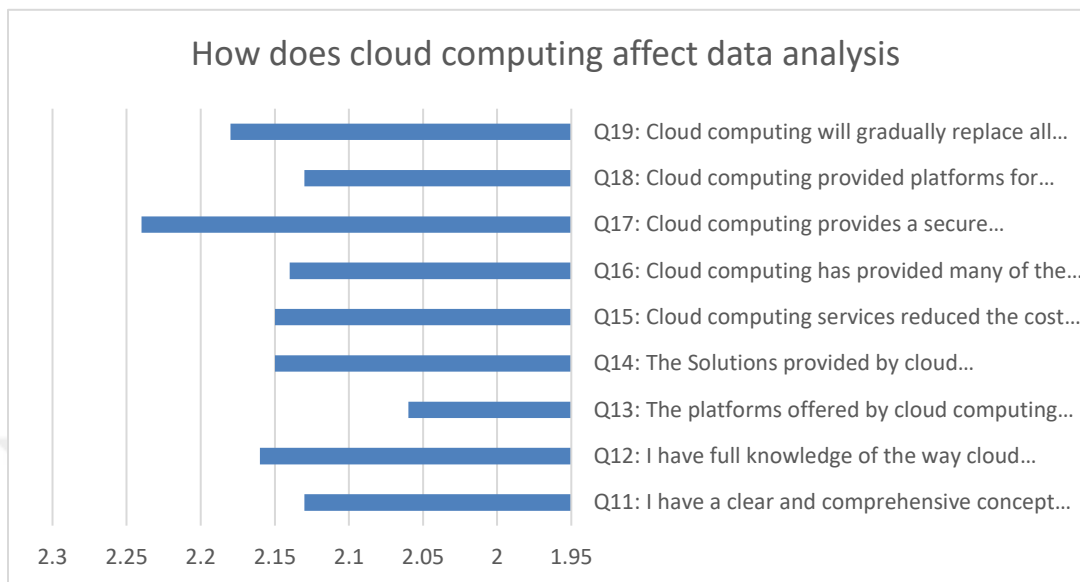
So the average of the relationship between data analysis and its large size is 2.1108 which consider a low level, since the intervals of level as follow:

Low Level [ 1.181 - 2.60].

Moderate Level [ 2.60 - 3.39].

High Level [ 3.40 - 5].

This is results shown clearly in the below figure (5.12)



**Figure 5.12:** Showing The Results of Section II

The following table no 5.7 presents the results of the third section of the questions followed by the comment on the results observed in the table with a graph showing the results of the third part of the questions.

**Table 5.7:** The result of analysis for section III

Questios		Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree	Mean	Std. Deviation	Rank
Q20: I have already knowledge about Big data analytics and I have a complete concept about them	N	12	16	35	244	50	2.15	.837	4
	%	3.4 %	4.5 %	9.8 %	68.3 %	14.0 %			
Q21: I have full knowledge of the way Big data analytics works	N	11	14	30	246	.817	2.10	.817	7
	%	3.1 %	3.9 %	8.4 %	68.9 %	13.7 %			
Q22: The ability to the analysis of Big data gives great economic value	N	16	25	35	229	52	2.23	.937	1
	%	4.5 %	7.0 %	9.8 %	64.1 %	14.6 %			
	N	9	19	49	231	49			

Q23: Big data analytics is provided by service providers only	%	2.5 %	5.3 %	13.7 %	64.7 %	13.7 %	2.18	.826	3
Q24: Big data analysis platforms provide services quickly and effectively	N	10	18	55	227	47	2.21	.836	2
	%	2.8 %	5.0 %	15.4 %	63.6 %	13.2 %			
Q25: Big Data Analytics is not free	N	10	21	47	223	56	2.18	.864	3
	%	2.8 %	5.9 %	13.2 %	62.5 %	15.7 %			
Q26: Through Big data analytics platforms we can analyze any data volume quickly and accurately	N	7	21	43	244	42	2.18	.787	3
	%	2.0 %	5.9 %	12.0 %	68.3 %	11.8 %			
Q27: Big data analytics platforms provide easy access for users	N	2	14	50	245	46	2.11	.683	6
	%	0.6 %	3.9 %	14.0 %	68.6 %	12.9 %			
Q28: Big data analytics platforms help experts analyze data from multiple sources	N	4	17	46	246	44	2.13	.726	5
	%	1.1 %	4.8 %	12.9 %	68.9 %	12.3 %			
Q29: The most important techniques of big data analytics are Hadoop, NoSQL	N	4	15	61	220	57	2.13	.765	5
	%	1.1 %	4.2 %	17.1 %	61.6 %	16.0 %			
<b>Weighted Mean</b>								<b>2.1586</b>	
<b>Std. Deviation</b>								<b>.50327</b>	

Table 5.7 shows ( How Providers Service effect Big Data Analytics? ) From which we find that the highest was awarded to the Q22: ( The ability to the analysis of Big data gives great economic value ) with mean 2.23 and Std.deviation .937 , followed by Q24: ( Big data analysis platforms provide services quickly and effectively) with mean 2.21 and Std.deviation .836 , followed by Q25 ( Big Data Analytics is not free) with mean 2.18 and Std.deviation .864 , followed by Q26 ( Through Big data analytics platforms we can analyze any data volume quickly and accurately ) with mean 2.18 and Std.deviation .787 , followed by Q20 ( I have already knowledge about Big data analytics and I have a complete concept about them ) with mean 2.15 and Std.deviation .837, with Strongly Disagree by percent ( 14.6% , 13.2% , 15.7% , 11.8% , 14.0% , respectively ) and

Disagree by percent ( 64.1% , 63.6% , 62.5% , 68.3% , 68.3% , respectively ).

With the lowest average was awarded to the Q21: ( I have full knowledge of the way Big data analytics works) with mean 2.10 and Std.deviation .817 , followed by Q27: ( Big data analytics platforms provide easy access for users) with mean 2.11 and Std.deviation .683 , followed by Q29 ( The most important techniques of big data analytics are Hadoop, NoSQL) with mean 2.13 and Std.deviation .765 , followed by Q28 ( Big data analytics platforms help experts analyze data from multiple sources ) with mean 2.13 and Std.deviation .726 , followed by Q20 ( I have already knowledge about Big data analytics and I have a complete concept about them) with mean 2.15 and Std.deviation .837, with Strongly Disagree by percent ( 13.7% , 12.9% , 16.0% , 12.3% , 14.0% , respectively ) and Disagree by percent ( 68.9% , 68.6% , 61.6% , 68.9% , 68.3% , respectively ).

The weighted average of section one ( III ) was 2.1586 with Std.deviation .50327 which indicate that the trend of the How Providers Service effect Big Data Analytics? is ( Disagree ) as a general trend according to 5-point Likert Scale as shown in Table ( 5.7 ) since 2.1586 lie in the interval. [ 1.81 – 2.60 ] .

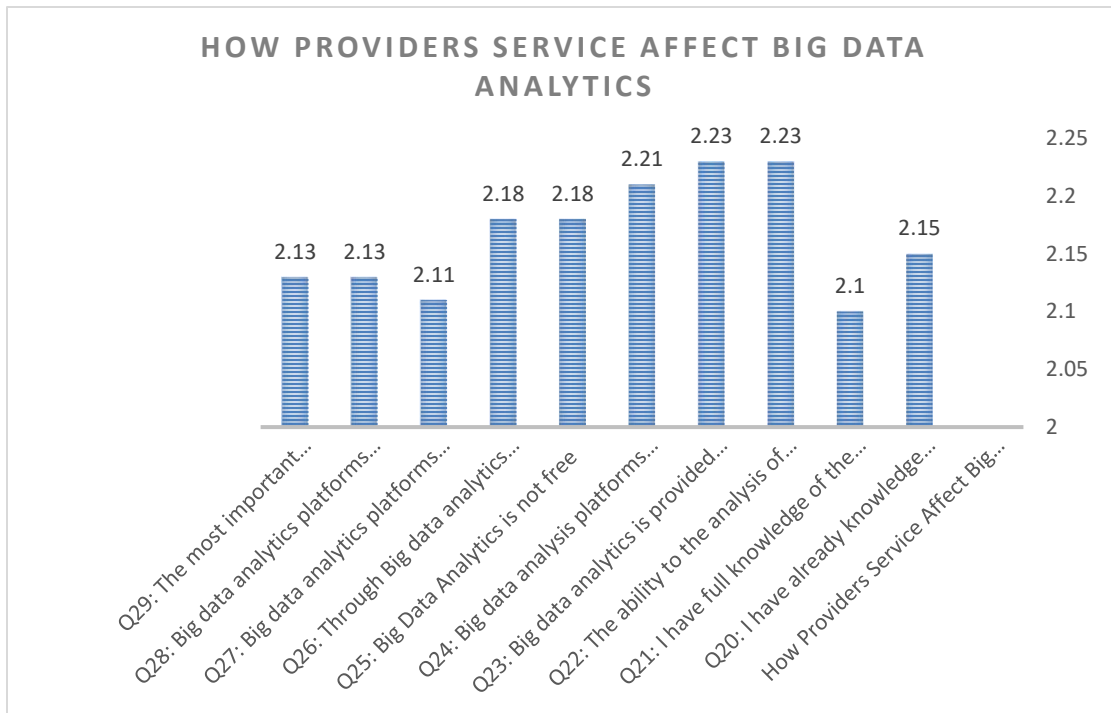
So the average of the How Providers Service effect Big Data Analytics? is 2.1589 which consider a low level, since the intervals of level as follow:

Low Level [ 1.181 - 2.60].

Moderate Level [ 2.60 - 3.39].

High Level [ 3.40 - 5].

This is results shown clearly in the below figure ( 5.13 ).



**Figure 5.13:** Showing The Results of Section III

By looking at the results of the questionnaire using the SPSS program, it was found that the general trend of the analysis is that most of the answers are in the low level which represents the ( Disagree ) of the questions. This gives a general indication that many users do not know about modern techniques, especially in the field of big data analytics. According to the results, this research will give a lot of motivation for the use of modern technologies to capture big data, most importantly cloud platforms.

## 5.8.2. Querying data

Two data types, which are handled through the BigQuery service are: The first one, which is included in the BigQuery platform page through the GCP, doesn't need to create and configure. The second type should be manually uploaded, for example, CSV file. This type of file should be manually uploaded on the BigQuery platform page [32].

### 5.8.2.1. BigQuery's public datasets

As the name suggests, the main purpose of BigQuery is to create queries about the data, so we

can try some queries. It is possible through Cloud Console and from the left-side navigation menu, BigQuery should be chosen. Unlike the APIs, a new page (or tab) is exclusively focused on BigQuery. Click Public Datasets, which show some of the datasets (Figure 5.14) [32].

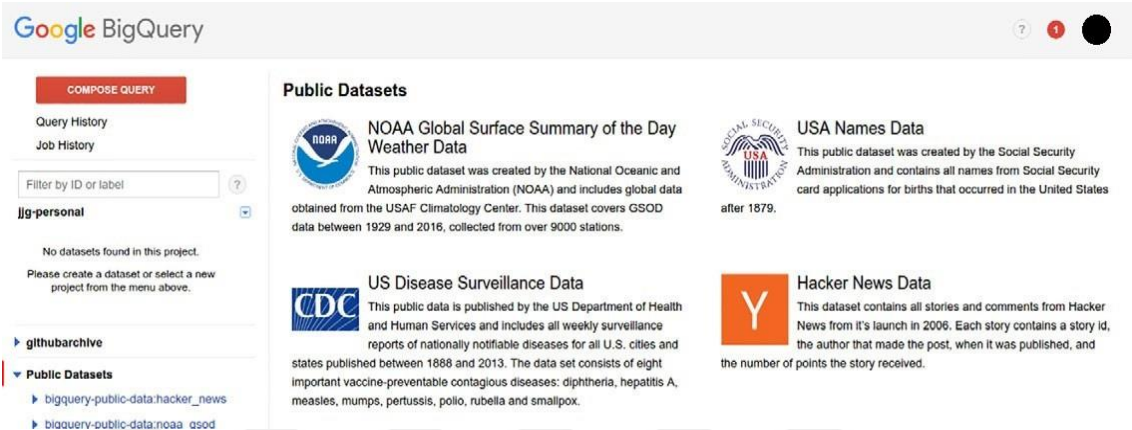


Figure 5.14: BigQuery’s public datasets [32]

5.8.2.2. BigQuery’s uploading dataset

BigQuery jobs support multiple types of operations; one of them is for loading new data. But there are multiple ways of shifting data from a source into a BigQuery table. Additionally, BigQuery tables themselves may be used in other data sources, such as Bigtable, Cloud Datastore, or Cloud Storage. Now, we’ll see how a chunk of data can be selected (such as a big CSV file) and loaded into BigQuery as a table [31]. In Figure 5.15, the sample of CSV dataset file has been given:

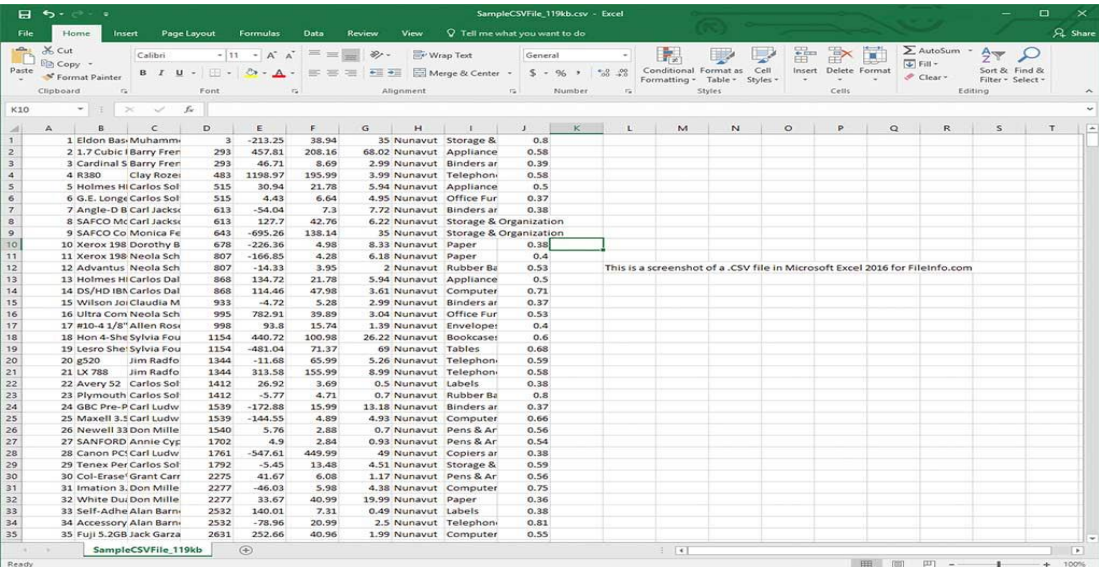


Figure 5.15: Sample CSV file

### 5.8.3. Real-Life Case Study

In this section, a public dataset from Google Cloud Services will be used in addition to uploading a dataset from a website that provides a wide range of the publicly available datasets.

#### 5.8.3.1. Real-life case study of a public dataset

In this section, the general data file is showing the details of a yellow taxi, as shown in Figure 5.16. If one of the choices is clicked (in this case, try out the yellow taxi dataset), BigQuery will open a data summary, which includes some details of the dataset itself and a list of the tables. If you click on the tables, BigQuery will bring you to a page that shows the most important piece: The schema (Figure 5.17.) shows the list of available fields, their data types, and a short description of the data that exists in each field. Notice that all of these fields are NULLABLE, so there's no guarantee that a value will be in there [32]. If a user clicks the Details tab at the top, it will provide an overview of the table, which, in this case, shows that it contains about 130GB in total, which is spread across over a billion rows. In Figure 5.12, complete Table ID is given, which is a combination of the project (in this case, NYC- TLC), the dataset (yellow), and the table (trips).

#### Table Info

<b>Table ID</b>	nyc-tlc:yellow.trips
<b>Table Size</b>	130 GB
<b>Long Term Storage Size</b>	130 GB
<b>Number of Rows</b>	1,108,779,463
<b>Creation Time</b>	Sep 25, 2015, 2:29:01 PM
<b>Last Modified</b>	Dec 24, 2015, 10:34:53 AM
<b>Data Location</b>	US
<b>Labels</b>	None <input type="button" value="Edit"/>

**Figure 5.16:** The yellow taxi trips schema [32]



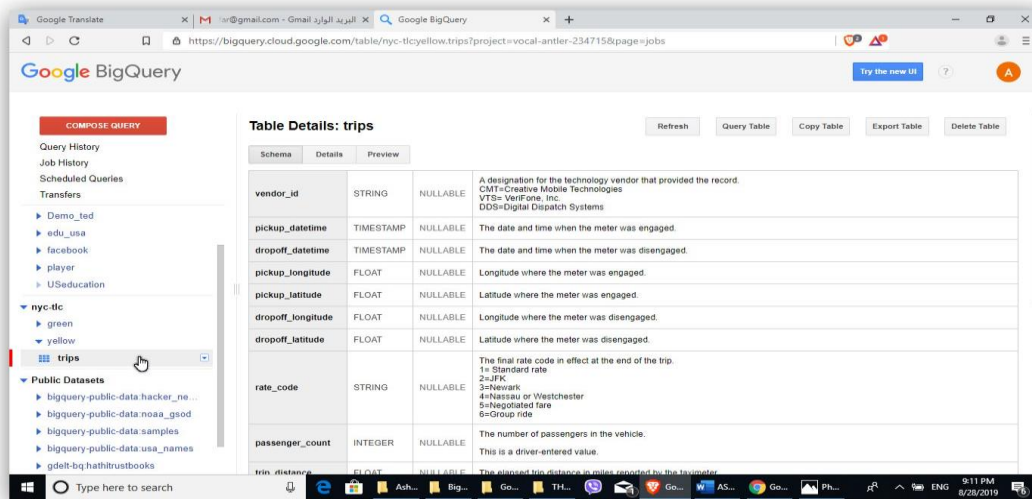


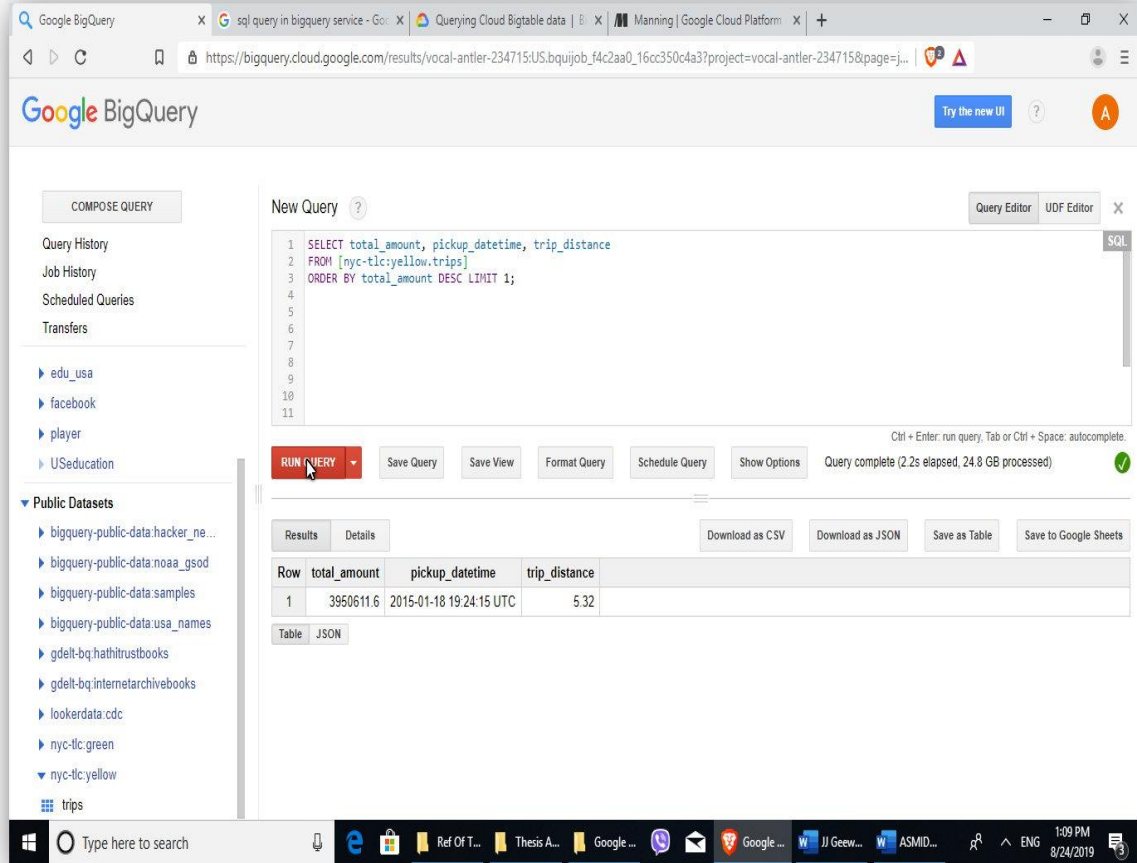
Figure 5.17: The yellow taxi trips table [32]

A few queries can be run that would be an interesting thing to do, but often the initial worry is “Won’t this take a few minutes?” That’s a reasonable first thought—after all, querying 1.1 billion records in PostgreSQL would probably take a while—so a query should be launched on any other database that can easily handle it as long as there was an index: the most expensive ride. To run this query over the table, Query Table button should be clicked at the top right and the following information can be added:

```
SELECT total_amount, pickup_datetime, trip_distance
FROM `nyc-tlc.yellow.trips`
ORDER BY total_amount
DESC LIMIT 1;
```

In case you’re unfamiliar with SQL, this query asks the table for some details sorted by the total trip cost but only gives you the first (most expensive) trip. Before you run this query exactly as it’s formatted, you’ll need to tell BigQuery not to use the legacy (old) SQL-style syntax. The newer syntax uses backticks for escaping table names rather than the square brackets since when BigQuery was first launched. This setting is accomplished by clicking Show Options and then unchecking the Use Legacy SQL box. When “Run Query” is clicked, BigQuery will get to work and should return a result within around two seconds. (In my case, it was 1.7 seconds.) It’ll also show how much data it queried within that time, which, in my case, was about 25 GB, and it

means that BigQuery sorted about 15GB per second to give back this result (Figure 5.18).

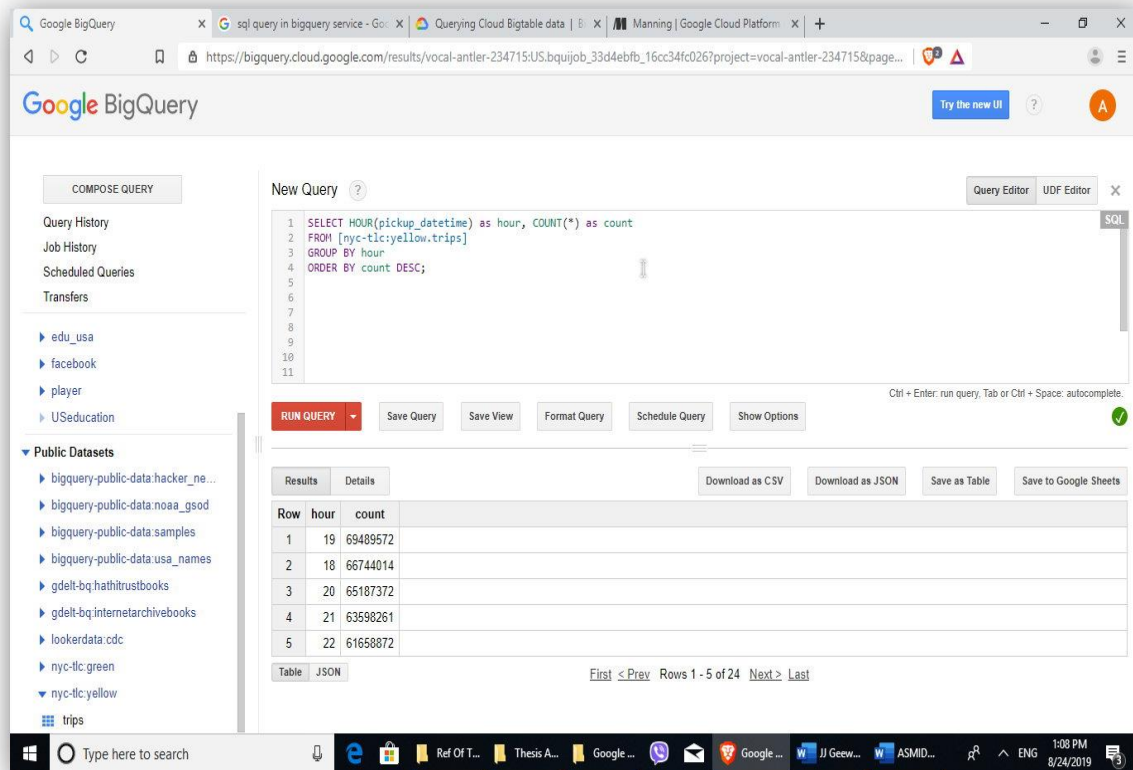


**Figure 5.18:** BigQuery results of the most expensive trip [32]

What if someone is trying to figure out what was the most common hour of the day that people were picked up? You'd have to take the pickup time and group by the hour part, then sort by the number of trips falling in each hour. In SQL, this isn't that complicated:

```
SELECT HOUR(pickup_datetime) as hour, COUNT(*) as count
FROM `nyc-tlc.yellow.trips`
GROUP BY hour
ORDER BY count DESC;
```

Running this query shows that the evening pickups are the most common (6–10 p.m.) and the early morning pickups are least common (3, 4, and 5 a.m.). See Figure 5.19.



**Figure 5.19:** Results of querying with grouping by pickup time [32]

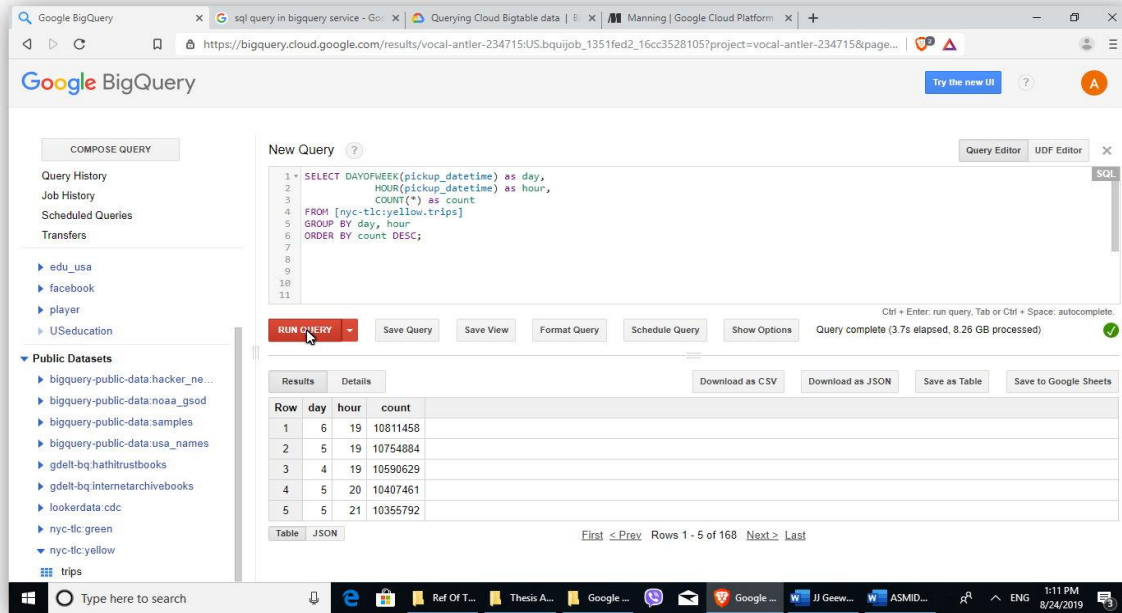
Perhaps more information is available here if there is a break down by the day of the week. Try adding that into the mix:

```

SELECT DAY OF WEEK (pickup_datetime) as day, HOUR (pickup_datetime) as hour,
COUNT(*) as count
FROM `nyc-tlc.yellow.trips`
GROUP BY day, hour
ORDER BY count DESC;

```

Running this query shows that the evening hours are most popular toward the end of the week (Thursday and Friday at 7 p.m. on the charts). See Figure 5.20.



**Figure 5.20:** Results showing the day and hour with most pickups [32]

Right now, a reader might be thinking “So what? MySQL can do all of this.” If so, BigQuery has done its job. The whole purpose of BigQuery is to create the need to run an analytical query with any other SQL database, but way faster. It is capable of scanning over a billion records stored in BigQuery. To make things even cooler, if a user needs to enhance his data size with respect to magnitude (10x (what it is today) to 10 billion rows), these queries would take about the same amount of time as they do now.

Running queries in the UI is fine, but what if a user wants to build something that displayed the data taken from BigQuery? This is where the client library (@google-cloud/bigquery) comes in. To see how it works with BigQuery, some code should be written, for example, that finds the most expensive ride. In order to use BigQuery, a user should log on to npm install @google-cloud/bigquery@1.0.0.

### 5.8.3.2. Real-life case study by uploading the dataset

After being able to register the service of Google Cloud Provider as well as after learning about the features of this service and identifying the type of datasets used in this service, it is possible to offer this project. The objective of this section is searching for a public dataset and uploading it onto BigQuery Datawarehouse. After this, a query will be run for finding the result in GCP. For this purpose, sample data source of TED talks was selected from [www.kaggle.com](http://www.kaggle.com) in CSV format. This kind of datasets possesses audio-video recordings and the information of TED Talks. These videos were uploaded at TED.com until 21 September, 2017. This downloaded dataset contains about certain recordings, which were uploaded on YouTube on various dates. But what TED represents here? The TED promotes design, entertainment, and technology.

The main objective is to find the top 10 topics from Ted Talks at YouTube having maximum views of all time from the downloaded dataset. To obtain the desired result, the following steps were performed:

#### 1 - Finding the Datasets

After some research on Google, a website [www.kaggle.com](http://www.kaggle.com) was found having multiple publicly available datasets. There are two steps needed for dataset download:

- A. A login account was created with an email ID and password on [www.kaggle.com](http://www.kaggle.com)
- B. With a link given below, a CSV file having all records for Ted Main Dataset was downloaded on a local computer: <https://www.kaggle.com/rounakbanik/ted-talks>. Figure 5.21 shows us how to find the TED dataset on Kaggle website.



Figure 5.21: Finding the TED dataset on Kaggle website

2 - Uploading the datasets to BigQuery Datawarehouse. This involves following steps in sequence:

A. Logging onto BigQuery by URL: <https://bigquery.cloud.google.com/welcome/mimetic-core-181107>

B. Creating new datasets in BigQuery after logging onto BigQuery, and clicking on my first project. Figure 5.22 explains that the default page of BigQuery has been highlighted. It can be seen that the drop down menu from “My First Project” highlights few options and the first option is to create a new dataset. Creation of dataset is a process to upload data on BigQuery Datawarehouse.

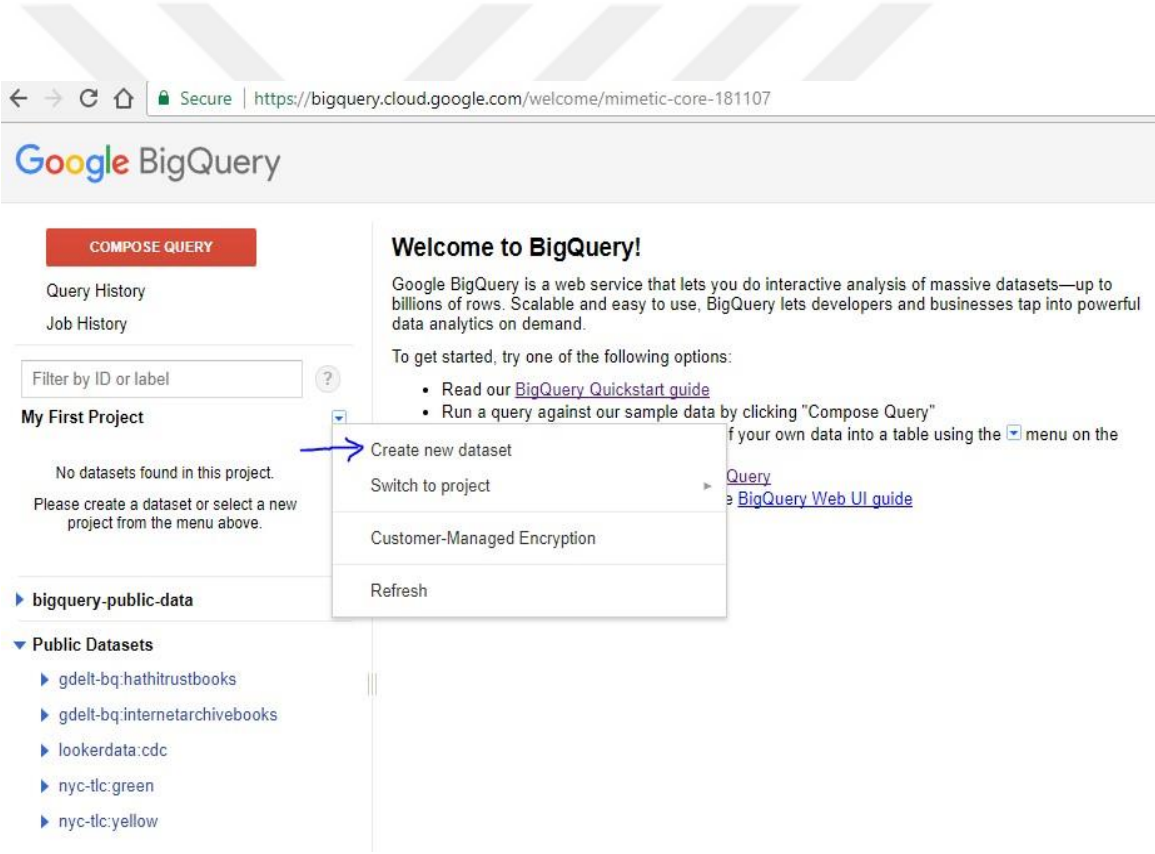
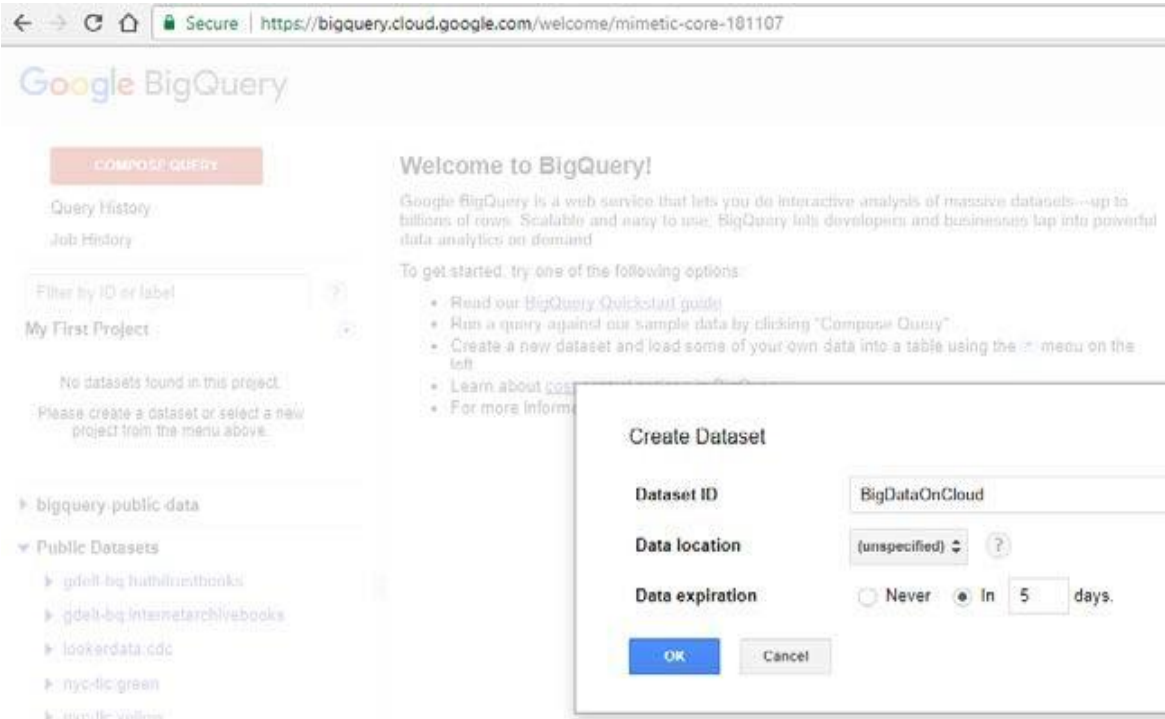


Figure 5.22: Process to create new datasets [32]

After clicking the “create dataset” option, the following window (Figure 5.23) appears on the screen: In this Figure, the key details such as Dataset ID, Data location and Data expiration details are entered to create a dataset in BigQuery.



**Figure 5.23:** Creating a data set in BigQuery [32]

In Figure 5.24, more details are added for table creation based on available source data, i.e. CSV file and it is uploaded from a local computer. In the next row, table name is entered and table button is created on bottom of the page, which is clicked to create table in BigQuery Datawarehouse. This step completes Dataset creation process on BigQuery.

The next step is to upload the data source on BigQuery Datawarehouse. In Figure 24, a file path is given, which was downloaded from [www.kaggle.com](http://www.kaggle.com) in an earlier step of this section.

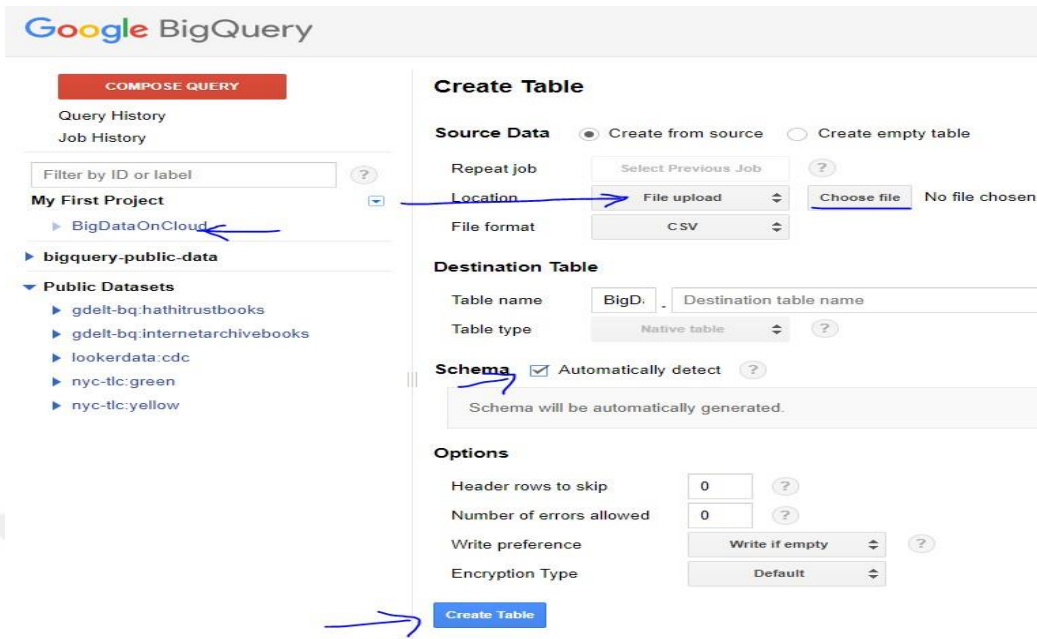


Figure 5.24: Uploading file to BigQuery Datawarehouse. [32]

In this Figure 5.25, table name is added, which will be used for querying the data.

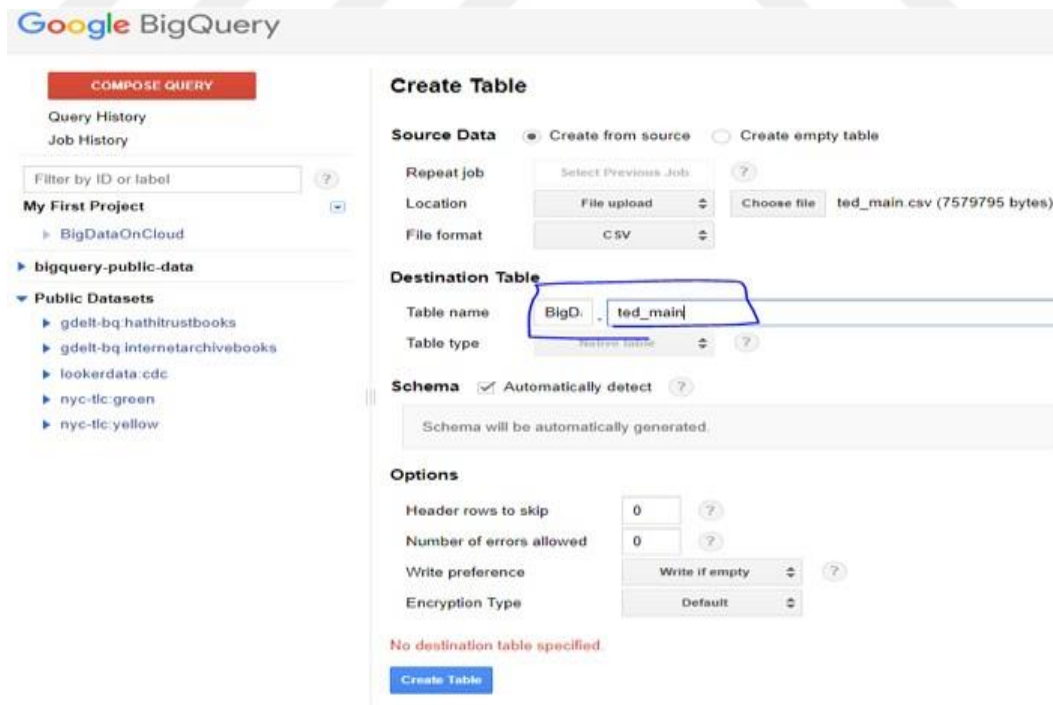
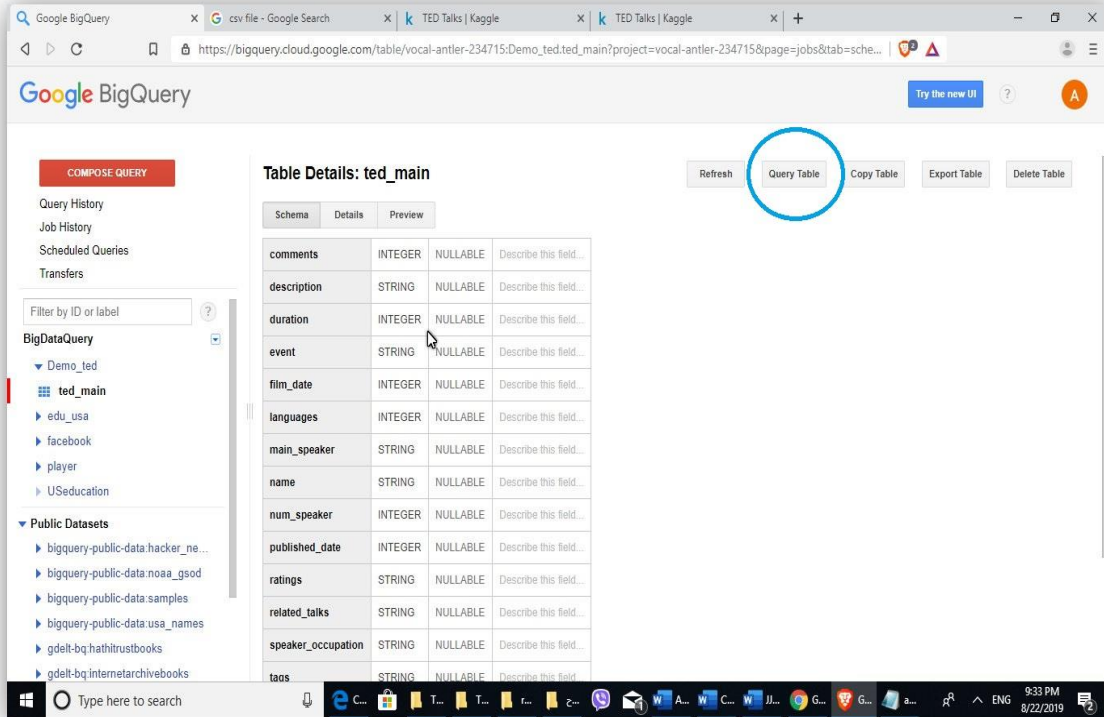


Figure 5.25: Adding table name [32]



### 3 - Querying table in editor

In Figure 5.26, the table is ready for query and finding top 10 topics are viewed by maximum count. This is accomplished according to the query given below:



**Figure 5.26:** Querying table on BigQuery Datawarehouse on Created Datasets [32]

**Final result:** Click query table, which is given in Figure 5.27, and writing below SQL query resulted in the needed output for finding out top 10 topics in TED event that has maximum views:

```
SELECT name, views, title, languages
FROM [Demo_ted.ted_main]
ORDER BY views
DEDC limit 10;
```

Figure 5.27 displays the results after writing SQL query in the Query editor:

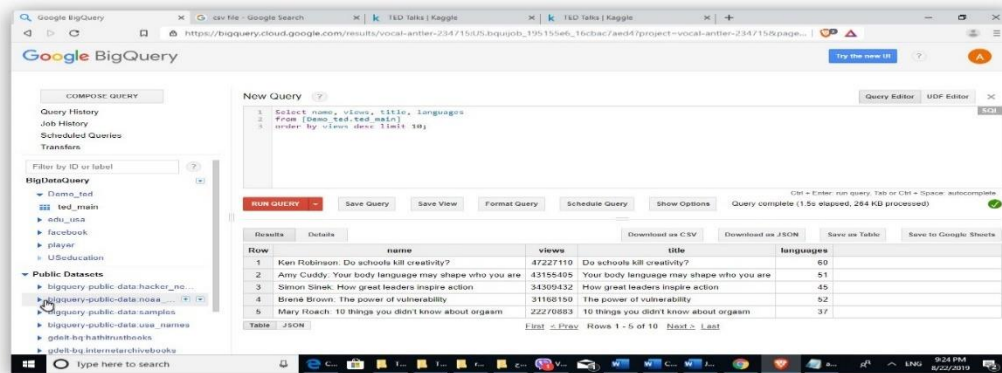


Figure 5.27: Query output [32]

## 5.9. Results of the Case Study

The survey shows that a large percentage of the users, who were targeted by the questions of a survey, did not know anything regarding cloud computing services and the big data analysis service. The respondents had significant proportion of male and female respondents and most of the respondents were students and engineers. The result of the survey was encouraging to proceed with and complete this research. We conducted a case study that highlights how easy it is to start an analytics project on cloud platform. The first of the demos have taken sample data available from the general data available on the Google Cloud Services platform, which is the yellow taxi dataset. Three queries were run on this data set:

- Query on the most expensive trip: This query was executed and the result was obtained within 2.2 seconds despite 24GB data size.
- The second query was about the possibility of knowing the most hours of service in a day. The result was obtained within seconds.
- The third query, in addition to the second query, was: When people were picked up during the week. All results were obtained in very little time.

These demos have taken the data from one of the publicly available data sets on website [www.kaggle.com](http://www.kaggle.com). The size of CSV file was small, only 12 megabytes. It was easy to initiate a cloud account on Google cloud and the process to upload CSV file was quite straightforward. During the upload process of CSV file, a table was created in BigQuery's datawarehouse. Finally, the SQL query displayed the results of top 10 topics with maximum views. It was observed from this demo that getting boarded on cloud is fairly quick and easy. GoogleQuery is a datawarehousing solution offered by the Google cloud platform.

## 6. CONCLUSION

Data is a new capital for enterprises and organizations. The analysis ensures that hidden facts are highlighted for better decision-making. Storing and processing RDBMS data was a sufficient solution in the previous decade. The evolution of data types, gigantic volumes, and velocity created a new phenomenon named as big data. Traditional servers and databases have limitations to process these big data sets; therefore, evolution of cloud computing took place.

Cloud computing is a platform, which is flexible, efficient and scalable, and it adds strategic value to organizations of all scales and types. Executing Big data analytics projects require scalable computing as well as storage to process a huge amount of data. Cloud computing has been massively used by organizations like Novartis, which has clearly shown that traditional million dollar projects can be completed with thousands of dollars, and that is a huge plus for businesses (though similar returns cannot be expected for every project). As data is becoming the key business driver, it's important to have high availability of data because without it, the credibility can be on stake during downtimes or production failures.

Cloud computing facilitates users by offering fast disaster recovery, which is helpful to deal with accidents. A user can start small and grow big through cloud computing platform, as one strong feature of cloud services is 'pay as you grow.' This is encouraging for start-ups and companies looking for more powerful computing resources. Security is a key area to meet data compliance and local regulatory requirements but at least big cloud providers are capable to comply with them.

Amazon has been leading the industry in cloud computing services but other players such as Azure, Google, Alibaba, IBM, Fujitsu, and Oracle etc. are catching up fast to provide huge opportunities for customers. From long list of cloud providers, this study has covered most reliable and cutting-edge big data analytics products, which are offered by top three cloud service providers i.e. Amazon Web Services, Microsoft Azure and Google Cloud Platform. Big data analytics projects are implemented in all the business areas but an important question is to find out the best solution that meets the requirements of every company and every project. Value is the most important characteristic in all the big data projects despite the fact that all the other characteristics such as volume, velocity, variety, and veracity help deriving the hidden information. We have

practically demonstrated processing publicly available datasets, which was analyzed on Google's BigQuery Datawarehouse service. It has demonstrated the comfort of instantly starting an analytics project on cloud computing platforms, and getting quick responses to complex queries. For all the analytics projects, the most important question lies in value, which needs to be created from the available datasets.



## REFERENCES

- [1] H. Liu and D. Orban, "Gridbatch: Cloud computing for large-scale data-intensive batch applications," in *2008 Eighth IEEE International Symposium on Cluster Computing and the Grid (CCGRID)*, 2008, pp. 295–305.
- [2] F. M. Groom and S. S. Jones, *Enterprise Cloud Computing for Non-Engineers*. CRC Press, 2018.
- [3] S. Sakr, A. Liu, D. M. Batista, and M. Alomari, "A survey of large scale data management approaches in cloud environments," *IEEE Commun. Surv. Tutorials*, vol. 13, no. 3, pp. 311–336, 2011.
- [4] N. Sawant and H. Shah, "Big Data Application Architecture," in *Big data Application Architecture Q & A*, Springer, 2013, pp. 9–28.
- [5] B. S. P. Mishra, S. Dehuri, and E. Kim, *Techniques and environments for big data analysis: parallel, cloud, and grid computing*, vol. 17. Springer, 2016.
- [6] J. Fan, F. Han, and H. Liu, "Challenges of big data analysis," *Natl. Sci. Rev.*, vol. 1, no. 2, pp. 293–314, 2014.
- [7] R. Gupta, H. Gupta, and M. Mohania, "Cloud computing and big data analytics: what is new from databases perspective?," in *International Conference on Big Data Analytics*, 2012, pp. 42–61.
- [8] T. Yang and Y. Zhao, "Application of cloud computing in biomedicine big data analysis cloud computing in big data," in *2017 International Conference on Algorithms, Methodology, Models and Applications in Emerging Technologies (ICAMMAET)*, 2017, pp. 1–3.
- [9] V. N. Inukollu, S. Arsi, and S. R. Ravuri, "Security issues associated with big data in cloud computing," *Int. J. Netw. Secur. Its Appl.*, vol. 6, no. 3, p. 45, 2014.
- [10] N. R. Kishor, "International Journal of Advance Research in Computer Science and Management Studies," *Int. J.*, vol. 2, no. 3, 2014.
- [11] R. B. Bohn, J. Messina, F. Liu, J. Tong, and J. Mao, "NIST cloud computing reference architecture," in *2011 IEEE World Congress on Services*, 2011, pp. 594–596.
- [12] P. Mell and T. Grance, "The NIST definition of cloud computing (draft)," *NIST Spec. Publ.*, vol. 800, no. 145, p. 7, 2011.
- [13] M. J. Kavis, *Architecting the cloud: design decisions for cloud computing service models (SaaS, PaaS, and IaaS)*. John Wiley & Sons, 2014.
- [14] R. S. Kenett, "Reliability and Availability Of Cloud Computing," *Qual. Prog.*, vol. 46, no. 10, p. 54, 2013.

- [15] T. Erl, R. Puttini, and Z. Mahmood, *Cloud computing: concepts, technology & architecture*. Pearson Education, 2013.
- [16] N. Japkowicz and J. Stefanowski, *Big Data Analysis: New Algorithms for a New Society*. Springer, 2016.
- [17] “Best 10+ Cloud Service Providers - 2017 Reviews | Clutch.co.” [Online]. Available: <https://clutch.co/cloud#survey>. [Accessed: 16-Aug-2019].
- [18] “About AWS,” Amazon Web Services, Inc. [Online]. Available: <https://aws.amazon.com/about-aws/>. [Accessed: 16-Aug-2019].
- [19] “Global Infrastructure.” [Online]. Available: [https://aws.amazon.com/about-aws/global-infrastructure/?nc1=h\\_ls](https://aws.amazon.com/about-aws/global-infrastructure/?nc1=h_ls). [Accessed: 16-Aug-2019].
- [20] “What is Azure—Microsoft Cloud Services | Microsoft Azure.” [Online]. Available: <https://azure.microsoft.com/en-us/overview/what-is-azure/>. [Accessed: 16-Aug-2019].
- [21] S. P. T. Krishnan and J. L. U. Gonzalez, *Building Your Next Big Thing with Google Cloud Platform: A Guide for Developers and Enterprise Architects*. Springer, 2015.
- [22] A. Sathi, “Big Data Analytics: Disruptive Technologies for Changing the Game, 2012.” S, 2012.
- [23] S. Srinivasa and V. Bhatnagar, *Big Data Analytics: First International Conference, BDA 2012, New Delhi, India, December 24-26, 2012, Proceedings*, vol. 7678. Springer Science & Business Media, 2012.
- [24] “What is Big Data Analytics? - Definition from Techopedia,” Techopedia.com. [Online]. Available: <https://www.techopedia.com/definition/28659/big-data-analytics>. [Accessed: 17-Aug-2019].
- [25] K. Krishnan, “Introducing Big Data Technologies Data Warehousing in the Age of Big Data Chapter 4,” *Sci. com*, 2016.
- [26] A. Anthony, *Mastering AWS Security: Create and maintain a secure cloud ecosystem*. Packt Publishing Ltd, 2017.
- [27] B. Beach, S. Armentrout, R. Bozo, and E. Tsouris, *Pro PowerShell for Amazon Web Services*. Springer, 2014.
- [28] T. Hunter and S. Porter, *Google Cloud Platform for Developers: Build highly scalable cloud solutions with the power of Google Cloud Platform*. Packt Publishing Ltd, 2018.
- [29] J. Xu, E. Huang, C.-H. Chen, and L. H. Lee, “Simulation optimization: A review and exploration in the new era of cloud computing and big data,” *Asia-Pacific J. Oper. Res.*, vol. 32, no. 03, p. 1550019, 2015.

- [30] D. Cukier, “DevOps patterns to scale web applications using cloud services,” in *Proceedings of the 2013 companion publication for conference on Systems, programming, & applications: software for humanity*, 2013, pp. 143–152.
- [31] H. E. Chen, “The Implementation Of Network Teaching Resources Sharing Center Based On Google Cloud,” in *Advanced Materials Research*, 2014, vol. 989, pp. 5580–5583.
- [32] J. Tigani and S. Naidu, *Google BigQuery Analytics*. John Wiley & Sons, 2014.
- [33] “Data Center Innovation,” Google Cloud. [Online]. Available: <https://cloud.google.com/about/data-centers/>. [Accessed: 28-Aug-2019].
- [34] “Global Locations - Regions & Zones,” Google Cloud. [Online]. Available: <https://cloud.google.com/about/locations/>. [Accessed: 28-Aug-2019].
- [35] A. Londhe and P. P. Rao, “Platforms for big data analytics: Trend towards hybrid era,” in *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*, 2017, pp. 3235–3238.
- [36] N. Mouthaan, “Effects of big data analytics on organizations’ value creation,” *Univ. Amsterdam*, 2012.
- [37] S. B. and T. Employment, “Benefits of cloud computing,” 15-Jun-2011. [Online]. Available: <https://www.business.qld.gov.au/running-business/it/cloud-computing/benefits>. [Accessed: 28-Aug-2019].
- [38] A. Alzahrani, N. Alalwan, and M. Sarrab, “Mobile cloud computing: advantage, disadvantage and open challenge,” in *Proceedings of the 7th Euro American Conference on Telematics and Information Systems*, 2014, p. 21.
- [39] “What is SPSS (Statistical Package for the Social Sciences) ? - Definition from WhatIs.com.” [Online]. Available: <https://whatis.techtarget.com/definition/SPSS-Statistical-Package-for-the-Social-Sciences>. [Accessed: 03-Nov-2019].
- [40] “What are the advantages of SPSS? - Quora.” [Online]. Available: <https://www.quora.com/What-are-the-advantages-of-SPSS>. [Accessed: 03-Nov-2019].
- [41] I. D. Constantiou and J. Kallinikos, “New games, new rules: big data and the changing context of strategy,” *J. Inf. Technol.*, vol. 30, no. 1, pp. 44–57, 2015.
- [42] A. McAfee, E. Brynjolfsson, T. H. Davenport, D. J. Patil, and D. Barton, “Big data: the management revolution,” *Harv. Bus. Rev.*, vol. 90, no. 10, pp. 60–68, 2012.
- [43] M. Gupta and J. F. George, “Toward the development of a big data analytics capability,” *Inf. Manag.*, vol. 53, no. 8, pp. 1049–1064, 2016.
- [44] H. Yang and M. Tate, “Where are we at with cloud computing?: a descriptive literature review,” in *20th Australasian conference on information systems*, 2009, pp. 2–4.

- [45] R. L. Villars, C. W. Olofson, and M. Eastwood, “Big data: What it is and why you should care,” *White Pap. IDC*, vol. 14, pp. 1–14, 2011.
- [46] C. Weinhardt, A. Anandasivam, B. Blau, and J. Stößer, “Business models in the service world,” *IT Prof.*, no. 2, pp. 28–33, 2009.
- [47] L. M. Vaquero, L. Rodero-Merino, J. Caceres, and M. Lindner, “A break in the clouds: towards a cloud definition,” *ACM SIGCOMM Comput. Commun. Rev.*, vol. 39, no. 1, pp. 50–55, 2008.
- [48] S. F. Wamba, S. Akter, A. Edwards, G. Chopin, and D. Gnanzou, “How ‘big data’ can make big impact: Findings from a systematic review and a longitudinal case study,” *Int. J. Prod. Econ.*, vol. 165, pp. 234–246, 2015.
- [49] P. P. Maglio and C.-H. Lim, “Innovation and big data in smart service systems,” *J. Innov. Manag.*, vol. 4, no. 1, pp. 11–21, 2016.
- [50] G. C. Fox, A. Ho, E. Chan, and W. Wang, “Measured characteristics of distributed cloud computing infrastructure for message-based collaboration applications,” in *2009 International Symposium on Collaborative Technologies and Systems*, 2009, pp. 465–467.