

Jitendra Kumar Jaiswal

# Cloud Computing for Big Data Analytics Projects



Helsinki Metropolia University of Applied Sciences

Master of Engineering - Information Technology

Master's Thesis

15 Apr 2018

## PREFACE

I would like to express my sincere gratitude to **Ville Jääskeläinen**, Principle Lecturer at Metropolia for supervising this thesis, helpful criticism and advice. His incredible knowledge of variety of Information Technologies to face research problem was inspiring. I would like to thank **Jonita Martelius** for her help with the English language check for this thesis.

I would greatly appreciate my co-workers at my workplace who have always shared different perspective from their experience in the subject area.

Last but not the least my family's support has been unparalleled during my entire master's studies.

Espoo, 15 Apr 2018  
Jitendra Kumar Jaiswal

## ABSTRACT

Author(s) Title	Jitendra Kumar Jaiswal Cloud Computing for Big Data Analytics Projects
Number of Pages Date	41 pages 15 Apr 2018
Degree	Master of Engineering
Degree Programme	Information Technology
Specialisation option	Information Technology
Instructor(s)	<b>Ville Jääskeläinen, Principal Lecturer</b>
<p>Data alone has no significance unless relevant information is extracted from it to support decision-making. Data analysis is the process of extracting useful information from available data. This information in turn helps decision makers to take appropriate actions. Traditionally, the analysis on data was performed through a process known as Extract, Transform and Load (ETL) on relational database management systems (RDBMS), which were designed primarily for vertical growth ( i.e. adding more Central Processing Unit(CPU) and Random Access Memory(RAM) to systems). As the industry is already in “Exabyte &amp; Zetta-byte Age” of data, the traditional approaches like RDBMS have faced limitations to store and process this humongous data due to their architectural principles designed during 70’s. The large amounts of data, structured or unstructured, has been termed as “Big data”, having mainly five properties i.e. Volume, Velocity and Variety, Veracity and Value of which Value is the most important whose main purpose is to extract relevant information from the other four V’s. To solve this problem of humongous data and analysis, various technologies have emerged in past decades. “Cloud computing” is a platform where thousands of servers work together to meet different computing needs and billing is done as per ‘pay as you grow’ model.</p> <p>This thesis studies Cloud computing and Big data fundamentals and benefits of a Cloud computing platform for Big data analytics projects. Knowing that there are many Cloud Service Providers (CSP’s), this thesis explores Big data analytics solutions available from industry’s top three Cloud computing service providers. The study includes a demo of a Big data analysis project on a leading Cloud computing platform to validate the power of Cloud computing by utilizing publicly available data sets.</p>	
Keywords	ETL, RDBMS, CPU, RAM, Exabyte, Zettabyte, Big data

## Table of Contents

Preface

Abstract

List of Tables

List of Abbreviations

1	Introduction	1
2	Fundamentals of Cloud Computing and Big Data	2
2.1	Cloud Computing Architecture	4
2.2	Cloud Service Models	6
2.2.1	Infrastructure as a Service (IaaS)	6
2.2.2	Platform as a Service (PaaS)	7
2.2.3	Software as a Service (SaaS)	8
2.3	Cloud Deployment Models	9
2.3.1	Private Cloud	9
2.3.2	Public Cloud	10
2.3.3	Hybrid Cloud	11
2.3.4	Community Cloud	12
2.4	Big Data	13
2.4.1	Big Data Characteristics	13
3	Leading Cloud Providers	15
3.1	Top Three Providers	16
3.1.1	Amazon Web Services (AWS)	16
3.1.2	Microsoft Azure	17
3.1.3	Google Cloud Platform (GCP)	18
3.2	Big Data Technologies	19
3.3	Big Data Analytics Products from AWS, Azure and GCP	21
3.3.1	AWS Big Data and Analytics Products	21
3.3.2	Microsoft Azure's Big Data and Analytics Products	23
3.3.3	Google Cloud Platform's Big Data and Analytics Products	25
4	Benefits of Cloud Computing for Big Data Projects	27
4.1	Benefits of Cloud Computing	27
4.2	Criticism/Disadvantages of Cloud Computing	30

5	Project Demonstration	32
5.1	Setting up Google Cloud and Big Query Environment	33
5.2	Real Life Case Study	35
5.3	Results of Demonstration	39
6	Conclusion	40
	References	

## **List of Tables**

Table 1. AWS Big Data and Analytics Product List	22
Table 2. Microsoft Azure Big Data and Analytics Product List	23
Table 3. GCP Big Data and Analytics Product List	26

## 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	Microsoft
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

What if there is a need to sort 10 Gigabyte (GB) of data? Modern day computers have enough memory to hold this much amount of data and can easily process it through in-memory sorting algorithms such as quicksort. What if there is a need to sort 100 GB or One Terabyte (TB) of data? There are high-end configuration servers available to hold this much amount of data in memory but as those are quite expensive it is better to select disk-based systems, but in this case, algorithm such as mergesort could be used for sorting the data. However, what if there is a need to sort 50TB, 100 TB 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 concludes that same problem with a different size of data needs a different solution [1].

The amount of new data is growing exponentially in the world. This can be imagined with fact that the data generated between beginning of the time and the year 2000 is now generated every minute. With such humongous data comes the problem to handle and process the data. “Big data” is the term introduced for such larger scale data sets, be it structured (like RDBMS) or un-structured (like social media, organizations data etc.). Analysis process of such large-scale data or Big data is known as Big data analytics. Cloud computing is a delivery of computing services such as servers, storage, networks, databases, analytics, applications and much more through the medium of Internet.

This thesis explores the benefits of Cloud computing platform for Big data analytics projects. In this process, this thesis studies the fundamentals of Cloud computing and Big data. As there are multiple cloud providers in industry for Big data analytics solutions, this thesis studies solutions provided by top three cloud leaders i.e. Amazon Web Services, Microsoft Azure, Google Cloud Platform. A demo is performed to understand how easy it is to run a Big data analytics project in a Cloud computing environment.

This thesis is divided into six sections. The target and the scope of this thesis was first introduced. The second section explores fundamentals of Cloud computing and Big data. In the third section, it further explores leading Cloud computing platforms for Big data analytics projects and analytics solutions respectively. The fourth section of this thesis discusses some of the benefits of Cloud computing and Big data analytics projects in general. The fifth section of this thesis performs a demo project and in the sixth section, thesis conclusions are made.



## 2 Fundamentals of Cloud Computing and Big Data

This section explores the fundamental understanding of Cloud computing and Big data in general. Further it deep dives architectural understanding of Cloud computing, Cloud Service Models and Cloud Deployment Models. What is Big data and its characteristics are explored later in this section.

Cloud computing is one of fastest growing segment in Information Technology. During past two decades when virtualization and distributed computing has evolved along with superfast network connections, Cloud computing has achieved phenomenal growth and introduced countless possibilities.

“The worldwide Cloud computing market grew 28% to \$110B in revenues in 2015. Synergy Research Group found that public IaaS/PaaS services attained the highest growth rate of 51%, followed by private & hybrid cloud infrastructure services at 45%.” [2]

According to the official NIST definition,

"Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction." [3]

Above definition in its simplest form means one can avail computing resources over the Internet and pay only for real usage of the resources. Any individual having a smartphone with an Internet connection and some apps is already part of a Cloud computing platform; even they may not know how things are running in huge datacenters in a backend. The size of data centers are huge as some of them are of the size of many football stadiums, situated in different geographical regions of the world.

The best part of Cloud computing is that one can start almost instantly using services and unlike traditional computing where one has to go through Capex (capital expenditure) model. Capex model is one where needed computing resources are either procured or rented from own datacenter or available service providers through established or non-established (new) processes.

In Cloud computing, resources such as servers, storage, network, and many other business applications can easily be provisioned without any human interaction. The billing of these services is calculated either per minute or per hours depending upon different services offered by the different Cloud computing platforms. Companies offering these computing services are known as cloud service providers and typically, they charge for Cloud computing services based on usage, similar to electricity or water billing in houses.

This study revealed through some case studies that the cost benefit between traditional data-center computing and Cloud computing is one strong reason for its immense growth. One can immediately start a project with a small investment and scale up according to business growth. This reduces the costs associated with setting up or renting a data center and procuring the servers and other IT infrastructures, which takes weeks to months to begin. With Cloud computing, business can instantly spin up hundreds of servers in minutes.

## 2.1 Cloud Computing Architecture

This section explores the fundamental understanding of the architecture. The terms Cloud and Cloud computing are both used interchangeably in further sections to ease out the understanding.

The cloud architecture consists of five separate components which work together to provide on-demand services.

Figure 1 is taken from National Institute of Standards and Technology (NIST) Cloud computing reference architecture [4]. It shows cloud architecture and its five components i.e. cloud provider, cloud consumer, cloud carrier, cloud auditor and cloud broker. These are briefly described as further.

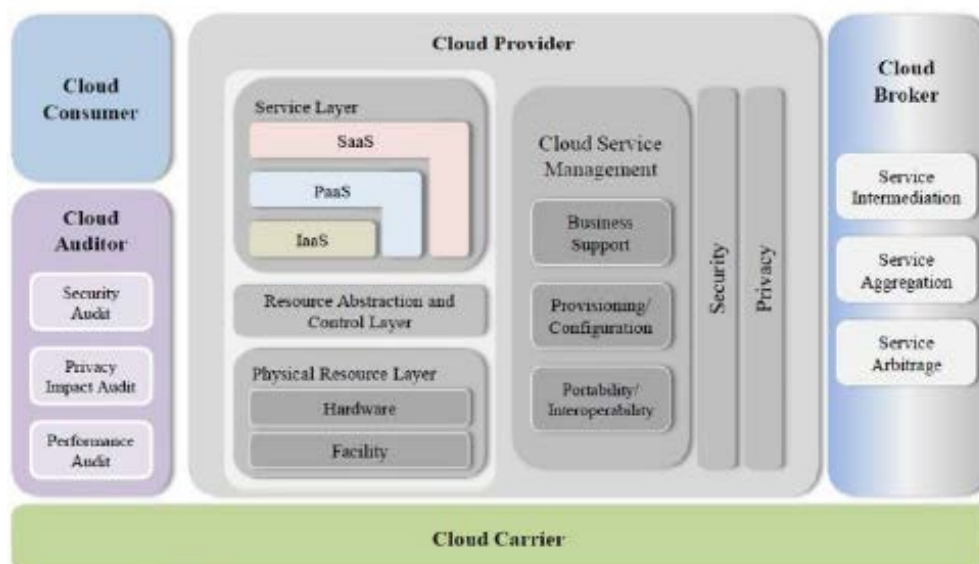


Figure 1: Cloud Architecture [4]

### Cloud Provider

The cloud provider are organizations that provides cloud services. The cloud provider has control over the IT infrastructure and manages the technical outages if planned or unplanned. Cloud provider also ensures that agreed Service Level Agreements (SLA's) are achieved. [5]

**Cloud Consumer**

A cloud consumer is a person(s) or an organisation using cloud service(s) and having an agreement with cloud provider or cloud broker. [5]

**Cloud Carrier**

The cloud carriers are networks and telecommunication companies, which ensure that services from cloud provider are available to cloud consumers. Cloud carrier work closely with cloud provider to meet agreed SLA's. [5]

**Cloud Broker**

The cloud brokers are third party companies, which work closely with both cloud providers and cloud consumers. Generally, these are consulting companies and so they can easily sell various cloud solutions to their existing customers as well as to new customers. [5]

**Cloud Auditor**

The cloud auditors are third parties who are specialized in independent assessment of cloud services offered by cloud providers. A cloud auditor can audit various areas such as security, privacy, performance, licensing, operations and other areas to highlight the gaps against various operations and data privacy standards. [5]

## 2.2 Cloud Service Models

There are multiple cloud services offered by various cloud providers. These are divided mainly into three categories of services i.e. Infrastructure as a service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS) also known as SPI (SaaS, PaaS, IaaS). [6]

Figure 2 represents three layers of Cloud service models. These three layers are further described as below.

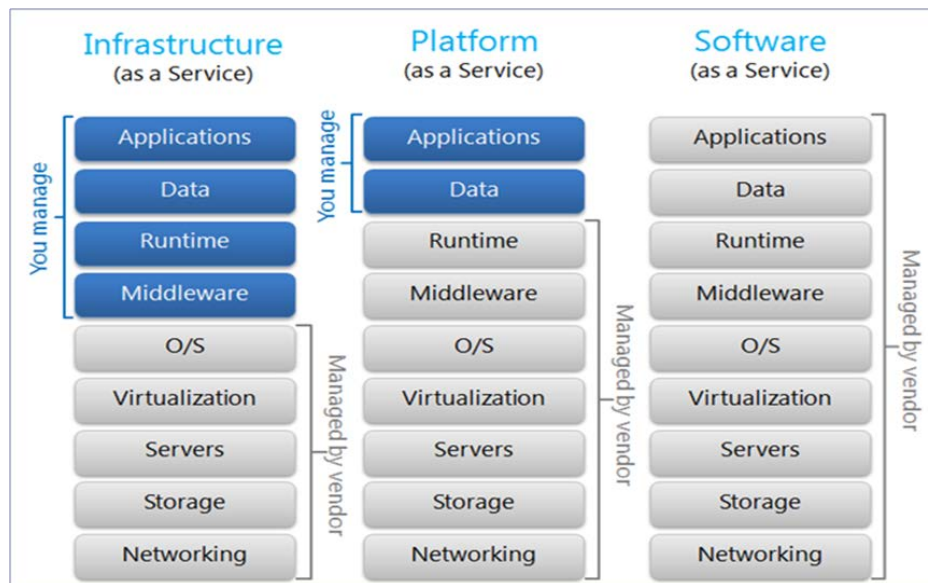


Figure 2: Three layers of Cloud service Models [39]

### 2.2.1 Infrastructure as a Service (IaaS)

The base layer of Cloud computing is known as Infrastructure as a Service (IaaS). This is most wide scale service used amongst the three service models. IaaS provides virtual infrastructure as well as raw hardware for creating, managing and removing storage, virtual machines, and virtual network over the Internet in cloud provider's infrastructure. Most of IaaS providers have capabilities to provide virtual servers with different configurations for example; one or more CPUs for computing with standard RAM as well as can add more RAM later as per demand. Like other service models infrastructure resources scale up or down and are billed as per real usage only. Leading cloud providers have capability to connect a virtual network in a cloud to company's network through VPN and it appears as a scalable IT infrastructure to existing IT infrastructure.

The popular advantage of an IaaS solution is rapid increase or decrease of an infrastructure on demand, lower risk in Return of Investment (ROI), reducing the human resource and hardware costs, automated scaling of computing power etc.

Example of these type of IaaS services are virtual machines, storage, servers, network, load balancers etc.

### 2.2.2 Platform as a Service (PaaS)

Platform as a Service is a layer on top of IaaS solution. The PaaS service provides a platform for creating new applications and software by developers or clients over the Internet. Customers can rent scalable virtualized servers, attached services, and can easily scale the demands as per requirement. Customers do not have control over network, storage, servers and OS but they can manage deployed applications and configurations. Costing model for PaaS could depend on multiple factors for example, number of I/O requests, storage usage in GB's, data transfer per GB etc.

There are advantages of PaaS model for example flexibility in load increase and decrease during a development process. In addition, in this case there is no need to manage infrastructure operations but to set proper rules for automatic load balance. Since PaaS has a design principle of one to many, an application can be configured by different customers from different locations. Security is a shared responsibility in PaaS between a cloud provider and a cloud consumer.

Example of these type of PaaS services are AWS elastic beanstalk, Google app engine, runtimes (like java runtimes), databases (like mySql, Oracle), web servers (tomcat) etc.

### 2.2.3 Software as a Service (SaaS)

This is a layer on top of PaaS solution. In SaaS, users or clients can use applications via thin/thick clients without installing those applications on local personal computers. The management of clients infrastructure and licensing of software applications are managed by SaaS provider. Some of the example of SaaS model are Microsoft OneDrive, Microsoft Office 365 etc. Network is a key component for better consumer experience and thus SaaS offering is good for lightweight applications compared to heavy weight applications such as 3D games. A cost model of SaaS based systems varies for applications as some are charged as per usage and some have a fix charge for a certain period of time.

An advantage of a SaaS model is a lower licensing cost since its design principle is one to many i.e. same application is used by many clients in parallel yet maintaining the isolation of each client. Other advantages are such as lower operations and maintenance cost which are also taken care by a SaaS provider as its infrastructure is controlled and managed by the SaaS provider.

Example of these type of SaaS services are Salesforce, Google Apps, Workday, Concur, Citrix GoToMeeting, Cisco WebEx, Microsoft Office 365 etc.

## 2.3 Cloud Deployment Models

A cloud deployment model represents a specific type of a cloud environment, primarily distinguished by ownership, size, and access control. Cloud deployment models are divided into Private Cloud, Public Cloud, Hybrid Cloud and Community Cloud.

### 2.3.1 Private Cloud

In Private Cloud (see Figure 3), the infrastructure is provisioned exclusively for a single organization having multiple clients for example business units, third parties, or vendors accessing the same resources.

Private clouds may be owned, operated and managed by single organization, a third party or a combination of them. It can be seen in Figure 3, that the infrastructure can be either on premise or off premise. In the last decade majority of large organizations, having their own data centres, have transformed into private cloud based solutions through virtualization technology.

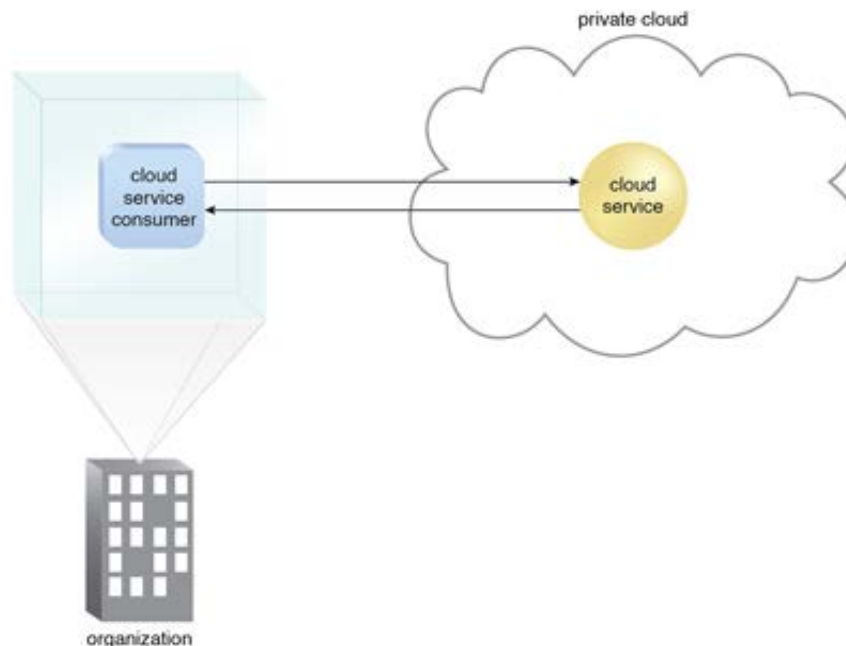


Figure 3 – Private Cloud [7]



### 2.3.2 Public Cloud

Public cloud (see Figure 4) is based on standard Cloud computing model where a provider assures availability of Virtual machines (VM's), storage or applications to public via Internet.

As per NIST, the definition for public cloud is described as:

“The cloud infrastructure is provisioned for open use by the general public. It may be owned, managed, and operated by a business, academic, or government organization, or some combination of them.” [8]

These cloud infrastructures are hosted on cloud provider's huge scale data centres spread across globe solving disaster recovery problems.

In public cloud, enterprises have no control over data and processing environments Also, because customers are connected to Public Clouds through Internet connections, Service Level Agreements (SLAs) may be difficult to meet if the Internet connections are not working properly. Some of the Public cloud examples are mentioned in figure 4.

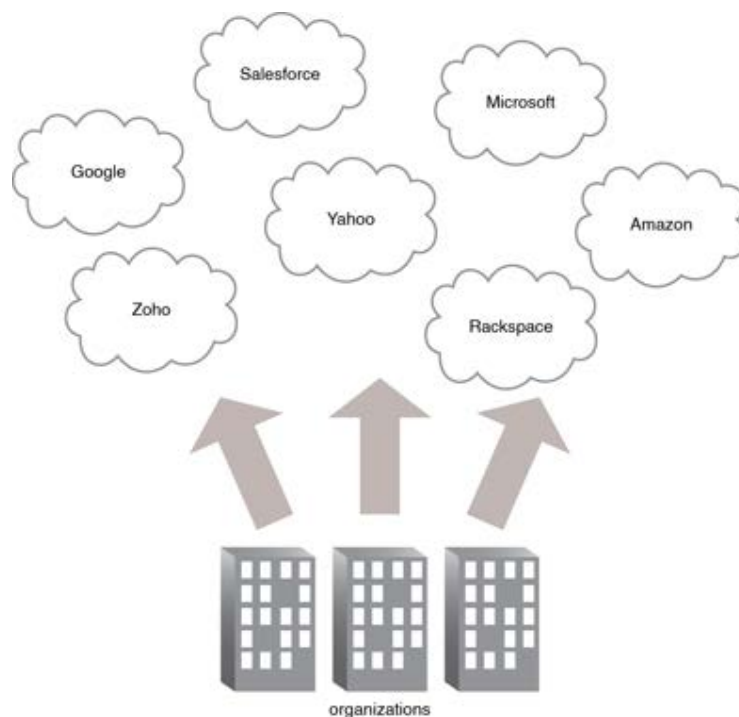


Figure 4 – Public Cloud [9]

### 2.3.3 Hybrid Cloud

A hybrid cloud (see Figure 5) consists of two or more different cloud providers for the same consumer. As an example, a banking customer may want to keep its business related financial data in a private cloud and it can operate marketing related tasks in a different cloud.

Hybrid cloud are quite popular among enterprise customers. This is due to their large existing infrastructure, which is not so straightforward to move onto the public cloud. So the enterprise either move a part of their existing resources into a cloud or start new projects directly in the cloud.

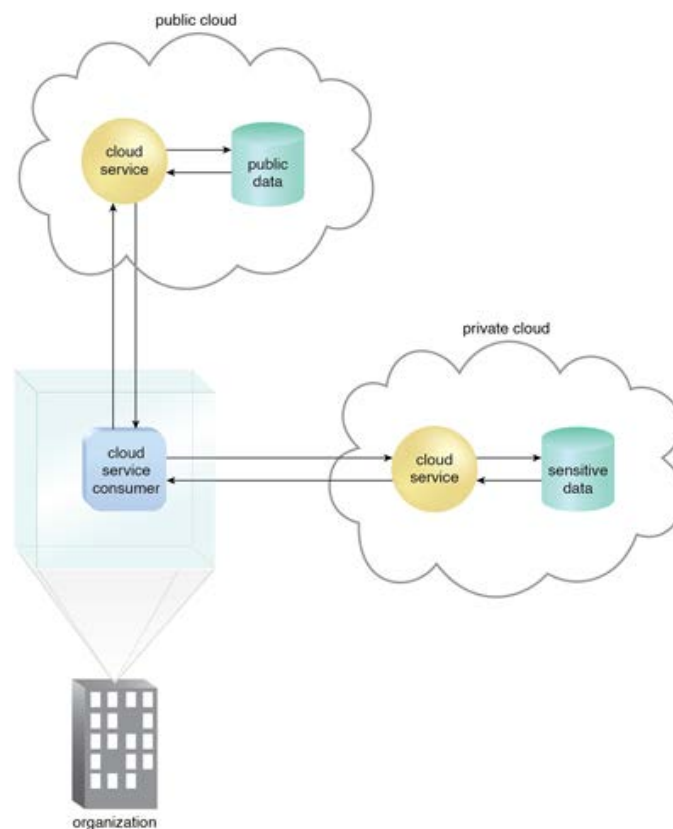


Figure 5 – Hybrid Cloud [10]

### 2.3.4 Community Cloud

Community Cloud (see Figure 6) is provisioned for a specific group of users or organisations.

The NIST definition for community cloud says:

“The cloud infrastructure is provisioned for exclusive use by a specific community of consumers from organizations that have shared concerns (e.g., mission, security requirements, policy, and compliance considerations). It may be owned, managed, and operated by one or more of the organizations in the community, a third party, or some combination of them, and it may exist on or off premises.”

[11]

A community cloud is almost like a public cloud but its usage is only for a specific community of cloud consumers. Community members not only share the responsibility of laying the foundation of the infrastructure and its evolution but also ensures that only allowed parties have access to it unless otherwise agreed.

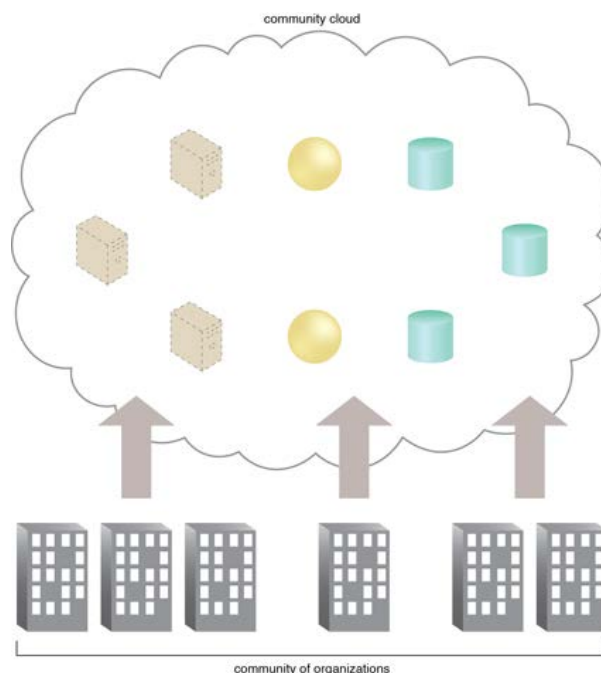


Figure 6 –Community Cloud [12]

Following section discusses Big data and its five key characteristics.

## 2.4 Big Data

In its simplest term, Big data is composed of data sets, which are huge in volumes that traditional data processing handling systems are incapable to process. The term Big data has been used since 1990's when John Mashey introduced the term.

The International Telecommunications Union (ITU) has defined Big data as follows:

“Big data is a paradigm for enabling the collection, storage, management, analysis and visualization, potentially under real-time constraints, of extensive datasets with heterogeneous characteristics.” [22]

Big data as a Service (BDaaS) is a cloud service category in which the capabilities provided to the cloud service customer are the ability to collect, store, analyse, visualize and manage data using Big data technologies. [22]

### 2.4.1 Big Data Characteristics

Three popular V's have described Big data i.e. Volume, Velocity, Variety. Additionally, two more V's were added later- Veracity and Value.

#### **Volume**

The amount of data generated and stored defines the Volume of data.

As an example, on Facebook alone, over 10 billion messages are sent every day, clicking “like” button over 4.5 billion times and over 350 million new pictures being added each day. Processing and storing such scale of humongous data is not easy with traditional relational database management systems. Big data is an answer to such example by distributing the data sets horizontally i.e. in multiple parallel networked computers and processing these data sets through latest algorithms.

### **Velocity**

Traditionally, corporations analyzed data using batch process, where a chunk of data is submitted to server for analysis and results were observed later. This approach is good when data input is slower than that batch processing job interval. To understand velocity, example from Facebook says a lot. Facebook has approximately 250 billion images as of March 2018. Facebook users uploads more than 900 million photos a day [40]. This data generation speed is represented as Velocity.

### **Variety**

Variety refers to different types of data. 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, with Big data technology, it is possible to load and process data types such as images, audio, video, JavaScript Object Notation (JSON), etc.

### **Veracity**

Veracity refers to data authenticity. With many different types of data, quality and reliability of data are less controllable. However, with Big data technology, it is possible to work with all kind of data.

### **Value**

Value is the most significant property in Big data. Big data is not useful unless one can create value out of it. So all other four V's i.e. Volume, Velocity, Variety and Veracity helps deriving Value or insight to support decision making. It is important to identify what Value a project is going to achieve with Big data analytics project.

### 3 Leading Cloud Providers

In year 2006, Amazon started AWS with virtual server Elastic Compute Cloud (EC2) instances and Simple storage service (S3) [43]. Their success led to a birth of many other cloud providers and Cloud computing has evolved into new business models. As per Gartner report June 2017, top three Cloud computing providers were Amazon Web Services (AWS), Microsoft (MS) Azure, and Google Cloud Platform. Besides these, many other cloud providers have entered into this business field. Alibaba, IBM, Oracle, Rackspace, Fujitsu are few to mention who have been evolving into Cloud computing business.

As can be seen in Figure 7, AWS and Microsoft are in the leader's quadrant and Google is the leader in the visionary's quadrant.



Figure 7: Gartner's Magic Quadrant for Cloud Infrastructure as a Service, Worldwide June 2017. [13]

## 3.1 Top Three Providers

In this section top three leading Cloud computing providers offering in general is discussed.

### 3.1.1 Amazon Web Services (AWS)

As on Dec 2017, AWS is the industry leader in a Cloud computing service providers list. Amazon's AWS has a wide range of products offering more than any other cloud provider. AWS also has more customers than any other.

As per AWS:

“Amazon Web Services provides a highly reliable, scalable, low-cost infrastructure platform in the cloud that powers hundreds of thousands of businesses in 190 countries around the world.” [14]

AWS claims to have their data centers spread across 18 regions in U.S., Europe, Japan, Singapore, Brasil, and Australia, ensuring customers from all industries to be benefitted from Cloud computing services. AWS claims to have advantages such as low cost, agility and instance elasticity, openness, flexibility, and at the same time capability to ensure security of the data with multiple security standards.

“The AWS Cloud operates 49 Availability Zones within 18 geographic Regions around the world, with announced plans for 12 more Availability Zones and four more Regions in Bahrain, Hong Kong SAR, Sweden, and a second AWS GovCloud Region in the US.” [15]

AWS has a large range of cloud services in many product categories:

- Compute,
- Storage,
- Database,
- Networking and content delivery,
- Developer tools,
- Management tools,
- Machine learning,
- Analytics,
- Security identity and compliance,

- AR & VR,
- Application integration,
- Customer engagement,
- Business productivity,
- Desktop and app streaming,
- Internet of things,
- Migration,
- Media services,
- Mobile services,
- Game development,
- Software and AWS cost management.

### 3.1.2 Microsoft Azure

Microsoft Azure is the second largest cloud service provider but it is gaining huge popularity amongst enterprise customers, as it has been originally an enterprise software company. Azure Cloud computing service was started in year 2010 and since then continuous innovation and growth has lead the service in leader's category in Gartner's magic quadrant June 2017. Microsoft Azure claims that 90 percent of fortune 500 customers trust on Microsoft cloud.

As per web definition of MS Azure:

“Azure is a comprehensive set of cloud services that developers and IT professionals use to build, deploy, and manage applications through our global network of data centers. Integrated tools, DevOps, and a marketplace support you in efficiently building anything from simple mobile apps to internet-scale solutions.”

[16]

Azure provides services in all three categories (SPI) i.e. SaaS, PaaS, IaaS at the same time it supports many different programming languages, tools and framework for Microsoft technologies as well as other third party software and systems. Azure claims to have data centers in 42 regions, more than any cloud provider does.

Azure has a huge list of products in following categories:



- Compute,
- Networking,
- Storage,
- Web and Mobile,
- Containers,
- Databases,
- Data and Analytics,
- AI and cognitive services,
- Internet of things,
- Enterprise Integration,
- Security + Identity,
- Developer's tools,
- Monitoring and management

### 3.1.3 Google Cloud Platform (GCP)

Google cloud platform is a Cloud computing offering from a search engine giant Google. It was started in year 2011, one year after Microsoft had launched its cloud services.

As per web definition of GCP:

“Google Cloud Platform is a suite of Cloud computing services that runs on the same infrastructure that Google uses internally for its end-user products, such as Google Search and YouTube.”[17]

GCP also includes several cloud-based services but not as many as AWS and Azure have at the time of writing this thesis:

- Compute,
- Storage and Databases,
- Networking,
- Big data,
- Data Transfer,
- Machine Learning,
- API platform and ecosystem,
- Identity and Security,
- Management tools,

- Internet of things,
- Developer tools etc.

### 3.2 Big Data Technologies

Cloud Providers have their own reference architecture for Big data analytics products, but in general two approaches are common to majority of analytics projects, Extract, Transform and Load (ETL) and Map Reduce.

Extract process in ETL is about reading data from a database. In this stage, data is collected from multiple different sources. Transform is a process where collected data in Extract phase is converted to needed form so it can be placed in another database. Finally, Load is process where converted data is written to target database.

Teradata corporation was first to analyse 1 TB of stored data on RDBMS in year 1992 when hard disk drives(HDD) were of 2.5 GB of size. As of 2017, Teradata claims to have installed an RDBMS of size over 50 Peta byte. Traditionally all the RDBMS has been practicing analytics process known as ETL (Extract, Transform and Load). With advent of Big data the requirement has evolved into Map Reduce architecture.

In 2004, Google published a paper on Map Reduce Architecture, which provides a parallel processing model to process huge amount of data. Hadoop is based on Map Reduce architecture. Map step is a process in which analysis queries are split and distributed across parallel nodes (having data) and processed simultaneously. Reduce step gathers and delivers the queried data. Map reduce technology is most popular in Big data analytics projects. [23]

#### **Big Data Analytics**

Big data analytics is a strategy for examining a large volume of data or Big data sets to decipher hidden patterns and useful information, which can help organizations to make more sensible decisions for business and societies.

These days data is gathered from many different sources such as social media, Internet of things (sensors), sales transactions, digital images, audios, videos, and more. Traditionally some of the advanced organizations have already been using analytics processes such as

ETL on RDBMS databases. As RDBMS have limitations to handle modern day data sources, a new term Big data analytics has evolved. The motivation behind this evolution has been to discover hidden patterns and correlations that might be otherwise invisible, and might provide valuable insights about the source of the data. Organizations may have edge due to this information which helps in making superior decisions. [24]

### 3.3 Big Data Analytics Products from AWS, Azure and GCP

The offering from top three cloud service providers on Big data analytics is a long list. Google has immense experience in search engine, where they already have been using analytics for their customers. AWS has some mature products such as Red shift, which is gaining popularity. Microsoft Azure has a long list of great products for Big data analytics solutions.

#### 3.3.1 AWS Big Data and Analytics Products

AWS has a broad spectrum of Big data services. Amazon Elastic Map Reduce (EMR) for example, runs Hadoop and Spark while Kinesis Firehose and Kinesis Streams provide a way to stream large data sets into AWS. A brief information about these product terms is as below:

- Hadoop or Apache Hadoop is an open source software framework for storing Big data and running applications on cluster of commodity hardware.
- Spark or Apache Spark is an open source cluster computing in memory computing environment used for running Analytics applications.
- Kinesis Firehose: Amazon Kinesis Data Firehose is the easiest way to load streaming data into data stores and analytics tools.
- Kinesis Streams: Amazon Kinesis is used to collect, process, and analyze real-time, streaming data to get timely insights and decide quickly to new information.
- Redshift: It is a petabyte-scale data warehouse, with data compression to help reduce costs.
- Amazon Elasticsearch is a service to deploy the open source Elasticsearch tool in AWS for analytics such as click-through and log monitoring. Kinesis Analytics complements this by analyzing data streams.

AWS has a larger set of data storage choices compared to Google. In addition to the massive AWS Simple Storage Service farm, it has DynamoDB, a low-latency NoSQL database; DynamoDB for Titan, which provides storage for the Titan graph database; Apache HBase, a petabyte-scale NoSQL database and relational databases. Graph database uses graph structure for semantic queries with nodes, edges and properties to represent and store data.

AWS also has a business intelligence (BI) service, QuickSight, which uses parallel, in-memory processing to achieve high speeds. QuickSight is further supported by Amazon Machine Learning and the AWS Internet of Things (IoT) platform, which connects devices to the cloud and can scale to billions of devices and trillions of messages, support it. Amazon Machine

Learning is a service on AWS, which provides tools, and services for developing Artificial intelligence applications. Internet of things (IoT) is a managed cloud platform that lets connected devices easily and securely interact with cloud applications and other devices. [18]

### **Analytics Services with AWS**

Table 1 has details about various services offered on AWS cloud platform for Big data analytics solutions. AWS has broad range of analytics services such as Data warehousing, Business Intelligence, Stream processing, Batch processing, Machine learning and Data orchestration. [19].

<b>Service</b>	<b>Product Type</b>	<b>Description</b>
Athena	Serverless Query service	Easily analyze data in Amazon S3, using standard SQL. Billing is based only for the queries one runs.
EMR	Hadoop	Provides a managed Hadoop framework to process vast amounts of data quickly and cost-effectively.
Elastic Search	Elastic Search	It makes it easy to deploy, secure, operate, and scale Elasticsearch for log analytics, full text search, application monitoring.
Kinesis	Streaming Data	Easiest way to work with streaming data on AWS.
Quick Sight	Business Analysis	Very fast, easy-to-use, cloud-powered business analytics with very low cost compared to traditional BI solutions.
Red Shift	Datawarehouse	Fast, fully managed, petabyte-scale data warehouse that makes it simple and cost-effective to analyze all data using existing business intelligence tools.
Glue	ETL	An ETL service that makes it easy for customers to prepare and load their data for analytics
Data pipeline	Dataf Workflow Orchestration	Helps to process and move data between different AWS compute and storage services, as well as on premise data sources, at specified intervals.

Table 1. AWS Big data and Analytics product list

### 3.3.2 Microsoft Azure's Big Data and Analytics Products

For analytics, Azure has Data Lake Analytics, which uses proprietary U-SQL language with SQL and C++, as well as HDInsight, a Hadoop-based service. There is also an Azure Stream Analytics service for analyzing streaming data, a Data Catalog service that identifies data assets using a global metadata system, and Data Factory service, which interlinks on-premises, cloud data sources, and manages data pipelines.

Azure's Big data storage service is Data Lake Store, a Hadoop file system. The cloud provider has a broad set of general-purpose storage offerings, including StorSimple, SQL and NoSQL databases and storage blobs. Blob storage is a feature from Microsoft azure which allows storing unstructured data in Azure cloud.

Azure also has Power BI and machine learning, lining up with AWS, and features an IoT Hub. Azure IoT hub is a fully managed service from Microsoft azure which allows reliable and secure bidirectional communication between millions of IoT devices and a solution backend. The cloud platform also includes a search engine. Microsoft's Cortana suite and Cognitive Services provide more advanced intelligence capabilities. [18]

Microsoft Azure Big Data and Analytics product list: Table 2 has popular analytics service solutions from Microsoft Azure platform [20].

<b>Service</b>	<b>Product Type</b>	<b>Description</b>
HD Insight	Hadoop	Azure HDInsight is a Hadoop-based service that brings an Apache Hadoop solution to the cloud. Gain the full value of Big data with a cloud-based data platform that manages data of any type and size.
Stream Analytics	Streaming Data	Azure Stream Analytics is an event-processing engine that helps you gain insights from devices, sensors, cloud infrastructure, and existing data properties in real-time.
Azure Bot Service	Serverless Bot Service	Azure Bot Service enables rapid intelligent bot development, bringing together the power of the Microsoft Bot Framework and Azure Functions.

Data Lake Analytics	Analytics	The Data Lake analytics service is a new distributed analytics service built on Apache YARN that dynamically scales so you can focus on your business goals, not on distributed infrastructure.
Data Lake Store	Repository	The Data Lake store provides a single repository where you can capture data of any size type and speed processing simply without forcing changes to your application as the data scales.
Data Factory	Data Transformation and Movement	Azure Data Factory is a managed service that lets a user to produce trusted information from raw data in cloud or on-premises sources.
Power BI Embedded	Data visualisation	Power BI Embedded provides stunning, fully interactive data visualizations in your customer-facing apps without the time and expense of having to build it from scratch.
Data catalogue	Enterprise Data Assets	Azure Data Catalog is a fully managed service that serves as a system of registration and system of discovery for enterprise data sources.
Log Analytics	Analytics	Azure Log Analytics lets a user to collect, correlate and visualize all machine data, such as event logs, network logs, performance data, and much more, from both on-premises and cloud assets.
Text Analytics API	Cognitive Service	Easily evaluates sentiments and topics to understand what users want.
Azure Analysis service	Analytics	Easily evaluates sentiments and topics to understand what users want.
Custom Speech service	Cognitive Service	Overcomes speech recognition barriers such as speaking style, background noise, and vocabulary.
Event hubs	Messaging	Azure Event Hubs enables elastic-scale telemetry and event ingestion with durable buffering

		and sub-second end-to-end latency for millions of devices and events.
SQL Data Warehouse	Datawarehouse	Azure SQL Data Warehouse is an elastic data warehouse as a service with enterprise-grade features based on the massively parallel SQL server processing architecture.

Table 2. Microsoft Azure Big Data and Analytics Product List

### 3.3.3 Google Cloud Platform's Big Data and Analytics Products

Google's BigQuery data service uses a SQL-like interface that is intuitive for most users (even nontechnical ones) to learn. It supports petabyte databases and can perform data streaming at 100,000 rows per second as an alternative to running data from cloud storage. BigQuery also supports geographic replication and users can select where they store their data.

BigQuery is a pay-as-you-go service without a dedicated infrastructure of instances, which allows Google to use a large number of processors to maintain fast query times. Integration with Spark, Hadoop, Pig and Hive is also supported. Pig is a high-level platform for creating programs that run on Hadoop. Hive is a data warehouse software which is built on top of Hadoop for reading, writing and managing large datasets stored in distributed datasets using SQL language. Organizations can also use Google Analytics and DoubleClick, a tool for the advertising industry that gathers statistics to feed BigQuery (as data sources). Google Cloud Dataflow allows users to sequence cloud data services.

Other Big data services offered by Google include Cloud Datastore, a NoSQL database for nonrelational data; Cloud BigTable, a massively scalable NoSQL database; Cloud Machine Learning, a managed platform for machine learning; and ancillary tools such as translators and speech converters.

One notable offering that Google is lacking for Big data is the Graphical Processing Unit (GPU) instance. Writing GPU code for data analytics is a high-value feature, given the incredible performance boosts that GPUs can offer. Google's lack of a GPU instance family is somewhat strange, especially with AWS having the feature since 2011 and Azure adding it in 2015 [18]

GCP Big data and Analytics product list: Google is third in Gartner's magic quadrant and it's analytics service solutions are captured in below table [21].



<b>Service</b>	<b>Product Type</b>	<b>Description</b>
BigQuery	Datawarehouse	Google's fully managed, low cost analytics data warehouse. BigQuery is serverless, there is no infrastructure to manage, no need to guess the needed capacity or overprovision, and no need for a database administrator.
Cloud Data Flow	Stream Analytics, ETL	A unified programming model and a managed service for executing a wide range of data processing patterns including streaming analytics, ETL, and batch computation
Cloud Dataproc	Hadoop and Spark	A managed Spark and Hadoop service, to easily process Big data-sets using the powerful and open tools in the Apache Big data ecosystem.
Cloud Datalab	Data Analysis	An interactive notebook (based on Jupyter) to explore, collaborate, analyze and visualize data.
Data Studio	Visualisation	Turns data into dashboards and reports that are easy to read, share, and customize.
Dataprep	Data Preparation Service	An intelligent data service for visually exploring, cleaning, and preparing structured and unstructured data for analysis.
Pub/Sub	Serverless Messaging service	Serverless, large scale, reliable, real-time messaging service that allows you to send and receive messages between independent applications.

Table 3. GCP Big Data and Analytics Product List

## 4 Benefits of Cloud Computing for Big Data Projects

This section introduces first the major benefits of Cloud computing, after that it also describes some criticism and disadvantages of Cloud computing.

### 4.1 Benefits of Cloud Computing

This section introduces the benefits of Cloud computing platform in general as well as the benefits it brings to Big data analytics projects.

#### **Flexibility**

Cloud computing is best optimised for fluctuating bandwidth demands. Unlike in traditional data centres where it takes weeks to scale up or scale down for any new computing requirement, this is much simpler in Cloud computing platform where this scaling up or down can be done almost in minutes and it is possible to automate the whole process. This flexibility is especially beneficial considering that cloud provider has enormous capacity hosted on giant data centers.

“Many organizations admit that the capability of promptly catering to business demands was one of the primary reasons why they shifted to Cloud computing.” [29]

#### **Cost Effective**

Cost is a major reason for acceptance of Cloud computing. The flexible cost models and operation expenditures makes cloud services very lucrative. Below are described some examples of Big data cases that have produced clear cost benefits.

Novartis:

“In 2013, Novartis ran a project that involved virtually screening 10 million compounds against a common cancer target in less than a week. They calculated that it would take 50,000 cores and close to a \$40 million investment if they wanted to run the experiment internally. Using Amazon Web Services (AWS) and the AWS Partner Network, Novartis built a platform leveraging Amazon Simple Storage Service (Amazon S3), Amazon Elastic Block Store (Amazon EBS), and four Availability Zones. The project ran across 10,600 Spot Instances (approximately 87,000 compute cores) and allowed Novartis to conduct 39 years of computational

chemistry in 9 hours for a cost of \$4,232. Out of the 10 million compounds screened, three were successfully identified.” [30]

Financial Times:

“By using Amazon Redshift, FT is supporting the same business functions with costs that are 80 percent lower than before. Headcount has not increased, and queries run much faster.” [31]

Dow Jones Case Study:

“The company has realized cost savings of 25 percent, more than \$40,000 per year, over the cost of leasing a data center—and the savings will continue each year that they use AWS. We will never have to refresh the hardware. That constitutes significant savings for Dow Jones.” [32]

Qatar Gas Transport:

“Qatari shipping and maritime company Nakilat has one of the world’s largest fleets of liquefied natural gas (LNG) carriers, transporting LNG from Qatar to global markets. To increase its competitive advantage, Nakilat wanted to improve employee productivity and mobility, without compromising on data security. It uses Microsoft 365 and Microsoft Cloud App Security to deliver highly secure cloud-first workplaces—shipboard and in the office. Nakilat also adopted the Microsoft Azure platform to optimize its operations and improve business continuity, reducing operating costs by 50 percent”. [33]

Carnegie Mellon University:

“The bank was very impressed by the energy savings it achieved using Carnegie Mellon’s dashboards and Power BI for Office 365,” says Lasternas. “It was able to reduce plug load energy consumption by 30 percent”. [34]

### **Disaster Recovery**

Unless there is a strong Disaster Recovery (DR) strategy in today’s competitive business environment, there is always a risk of impact to business in case the infrastructure fails.

According to Aberdeen Group,

“Small businesses are twice as likely as larger companies to have implemented cloud-based backup and recovery solutions that save time, avoid large up-front investment and roll up third-party expertise as part of the deal.” [35]

Cloud computing provides a solid Disaster Recovery infrastructure as majority of these solutions are based in different regions as we all as in the same region in different locations known as availability zones. Regions are independent geographic areas consisting zones. These zones are also called availability zones, which may contain one or more datacentres in a region.

### **Opex Based Instead of Capex Based**

Cloud services are based on pay as you go model, hence there is no requirement for big upfront investments. Now days there are providers such as Google and Amazon where they provide a couple of basic services for free for one year period. This helps start-ups and small companies to kick start at small scale and as the requirement grows they can anytime upgrade their services.

For enterprise services it can be an agreement to pay monthly for the usage or a yearly contract as per agreement with Cloud provider. This flexibility is another major reason for Cloud adoption from start-ups to enterprises.

### **Promoting a Greener Earth**

As per a study, in comparison to on-site server, cloud offers 30% lower energy consumption and subsequent carbon emission. Study also says that smaller organizations may even reduce 90% of energy usage as well as carbon emission. [29]

### **Rapid service Introduction**

Cloud services can be deployed rapidly in cloud environment and they are ready for use in a matter of minutes. It is easy to start using cloud services such as compute resources, storage capacity or application as a service etc. [36]

### **Improved Security**

Lost laptops are a severe business problem not because of the cost of piece of hardware but sensitive data inside it. Cloud computing gives a greater security when this happens. Because the data is stored in the cloud, one can access it no matter what happens to the machine. For an example, office365 is a group of subscriptions, which provides productivity software and services. Outlook, Microsoft Word, Microsoft Excel, Microsoft Power point etc. are few of its services where a user can keep the data on cloud so as to avoid any risk associated with laptop loss etc. [35]

## 4.2 Criticism/Disadvantages of Cloud Computing

Despite the benefits described in section 4.1, the Cloud Security Alliance has identified several barriers holding back cloud adoption. At 73% of companies, the security of data is the top concern holding back cloud projects. That has followed by concern about regulatory compliance (38%), loss of control over IT services (38%), and knowledge and experience of both IT and business managers (34%). As organizations address their security and compliance concerns by extending corporate policies to data in the cloud and invest in closing the cloud skills gap, they can fully take advantage of the benefits of cloud services. [38]

### **Network Connectivity**

There is always a dependency on Internet connectivity to access cloud services. Different services have different requirements for Internet connections with reference to Internet speed, network latency etc. [37]

### **Security Concerns**

One of the major issue while in the cloud is that of security issue. Before adopting this technology, it should be decided if the company willing to give sensitive information to a third-party cloud service provider. This could potentially put company to a great risk. Hence, one needs to choose the most reliable service provider, who will keep the information as secure as possible. [37]

### **Prone to Attack**

Storing information in the cloud could makes the company vulnerable to external hack attacks and threats. [37]

Cloud service providers are consistently targeted for attacks 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:

- Distributed denial of service attacks: Traditionally in DDoS, many systems at once overloads a target server, causing it to either be less effective or make its operations into cease. In 2016, Dyn attack demonstrated that large websites such as Amazon and Twitter's accesses were not available for customers. [41]

- Man in the cloud attack: This is a recently discovered method 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 in windows credential manager. The victim (user) is hit with malwares either via a website or email, which an attacker gains access to local files.

“By replacing the cloud synchronization token for one that points to the attacker's cloud account and placing the original token into the selection of files that will be synchronized, the victim is lead to unknowingly upload their original token to the attacker. That token can then be used by the attacker to gain access to the victim's actual cloud data”  
[42]

## 5 Project Demonstration

During thesis research it was found out that all major cloud providers have a 'free trial time period' to start and explore their cloud services with limited resources. For example, at present, Azure provides one month's time and \$200 as a free credit for using various services on their cloud platform for 30 days period. AWS provides 12 months of free trial with some limitations and some free limited products even after 12 months.

At the time of writing this thesis, the author used Googles Cloud Platform for exploring the functionality of Cloud computing for Big data sets. The Google Cloud platform provided \$300 as a free credit over 12 months for any GCP product.

As described earlier in Table 3, it was found out that Google has an Enterprise Datawarehouse service known as BigQuery Datawarehouse:

“BigQuery offers scalable, flexible pricing options to help fit your project and budget. BigQuery charges for data storage, streaming inserts, and for querying data, but loading and exporting data are free of charge.” [39]

The demo was based on analysis of data loaded in BigQuery Datawarehouse from Google cloud platform. Big Query is a petabyte scale, one of the fastest data warehouse solution for Big data analysis.

The main purpose of this demo is to demonstrate the quick access of Big data service in cloud. This demo was performed in below steps:

- Setting up Google Cloud and Big Query environment
  - Google Cloud Platform Account Creation
  - Login to Console
  - Login to Big Query
  - Browsing Publicly available sample tables
- Real life case study
  - Downloading Publicly available data set
  - Uploading on Google Big Query
  - Result set/ Query Execution
- Results on Demo

## 5.1 Setting up Google Cloud and Big Query Environment

This section introduces how to create an account in Google cloud, setup BigQuery environment and perform some queries on publicly available datasets in BigQuery.

1. Steps to create Google cloud account: Below steps are performed to create Google Cloud Account for free tier:

- a. Go to <https://cloud.google.com/>
- b. Click TRY IT FREE tab
- c. Sign up with Gmail (else it says can't find your google account)/password
- d. Enter the password of Gmail account
- e. Try cloud platform for free
  - i. Enter country if not selected by default
  - ii. Accept terms of services
- f. Customer info page appears
  - (i) Enter all details such as of Full name and address details
  - (ii) Enter payment method, preferably it accepts credit card
- g. Click Start my free trial

After this step, webpage of Google cloud platform's home console is displayed where the first step is to create a project.

2. Creating a project on Google Cloud Platform's one of the Analytics service named as BigQuery:

It is possible to access publicly available datasets and query it through structured query language (SQL) to see various outputs and also speed of data processing in BigQuery's datawarehouse.

3. Accessing publicly available sample datasets in BigQuery Datawarehouse:

- a. Click on product and services (top left)
- b. In Big data product category click on Big Query
- c. Click on bigquery-public-data-sets

It can be seen that there are many popular sources such as Wikipedia, Github etc. have datasets available in publicly available datasets category.

4. Browsing publicly available data-sets and running some queries with the query editor:



After clicking on any of the table, for example Wikipedia, one can see metadata about the table. Metadata represents information about data. In below Figure 8, column details can be seen about a Wikipedia table.

More sample tables can be seen on the left panel of the page. The tables can be queried by clicking 'Query Table' button on top right in web console.

The screenshot shows the Google BigQuery interface for the 'wikipedia' table. On the left, a sidebar lists various sample datasets, with 'wikipedia' highlighted. The main content area is titled 'Table Details: wikipedia' and contains a table with the following columns and descriptions:

Column Name	Data Type	Nullability	Description
title	STRING	REQUIRED	The title of the page, as displayed on the page (not in the URL). Always starts with a capital letter and may begin with a namespace (e.g. "Talk:", "User:", "User Talk:", ...)
id	INTEGER	NULLABLE	A unique ID for the article that was revised. These correspond to the order in which articles were created, except for the first several thousand IDs, which are issued in alphabetical order.
language	STRING	REQUIRED	Empty in the current dataset.
namespace	INTEGER	REQUIRED	Wikipedia segments its pages into namespaces (e.g. "Talk:", "User:", etc.). MEDIA = 202, 0 = 2 in IFP XML, but these values must be >0 SPECIAL = 202, 0 = 1 in non XML, but these values must be >0 MAIN = 0 TALK = 1 USER = 2 USER_TALK = 3
is_redirect	BOOLEAN	NULLABLE	Versions later than ca. 230938 may have a redirect marker in the XML.
revision_id	INTEGER	NULLABLE	These are unique across all revisions to all pages in a particular language and increase with time. Sorting the revisions to a page by revision_id will yield them in chronological order.
contributor_id	STRING	NULLABLE	Typically, either_ip or (_id and _username) will be set. IP information is unavailable for edits from registered accounts. A (very) small fraction of edits have neither_ip or (_id and _username). They show up on Wikipedia as "(Username or IP removed)".
contributor_ip	INTEGER	NULLABLE	Typically, either (_id and _username) or_ip will be set. A (very) small fraction of edits have neither_ip or (_id and _username). They show up on Wikipedia as "(Username or IP removed)".
contributor_username	STRING	NULLABLE	Typically, either (_id and _username) or_ip will be set. A (very) small fraction of edits have neither_ip or (_id and _username). They show up on Wikipedia as "(Username or IP removed)".
timestamp	INTEGER	REQUIRED	In Unix time, seconds since epoch.
is_minor	BOOLEAN	NULLABLE	Corresponds to the "Minor Edit" checkbox on Wikipedia's edit page.
is_bot	BOOLEAN	NULLABLE	A special flag that some of Wikipedia's more active bots voluntarily set.  If this edit is a revision to a previous edit, this flag records the revision_id that was reverted. If the same article had no previous revisions, then

Figure 8. Sample dataset of Wikipedia on BigQuery

In the following section, a real data set is taken from a publicly available dataset. It is then uploaded to BigQuery Datawarehouse and then queries are executed for desired results.

## 5.2 Real Life Case Study

The objective of this section is to find a publicly available dataset, upload it into BigQuery Datawarehouse and then run query to find result.

For this purpose, sample data source of TED talks was selected from [www.kaggle.com](http://www.kaggle.com) in CSV format. These datasets contain information about all audio-video recordings of TED Talks uploaded to the official TED.com website until September 21st, 2017. [40] This dataset downloaded has information about all the recordings which were uploaded on Youtube on various dates. But what TED represents here?

“TED (Technology, Entertainment, and Design) is a media organization which posts talks online for free distribution, under the slogan "ideas worth spreading"” [44]

**Problem Statement:** The main objective is to find top 10 topics from Ted Talks at YouTube having maximum views of all time from dataset downloaded

The following steps were performed for achieving desired result:

### 1) Finding the Datasets

After some research on google, a website named as [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 below link, a CSV file having all records for Ted Main Dataset was downloaded on local computer:  
<https://www.kaggle.com/rounakbanik/ted-talks>

### 2) Uploading the datasets to BigQuery Datawarehouse

This involves below steps in sequence:

- a) Logging into BigQuery on below URL:  
<https://bigquery.cloud.google.com/welcome/mimetic-core-181107>

### b) Creating new datasets in BigQuery

After logging into BigQuery, clicked on my first project (Figure 9 below).

In this figure, the default page of BigQuery is highlighted. It can be seen that drop down menu from 'My First Project', highlights few options and first option is to create a new dataset. Creation of dataset is a process to upload data on BigQuery Datawarehouse.

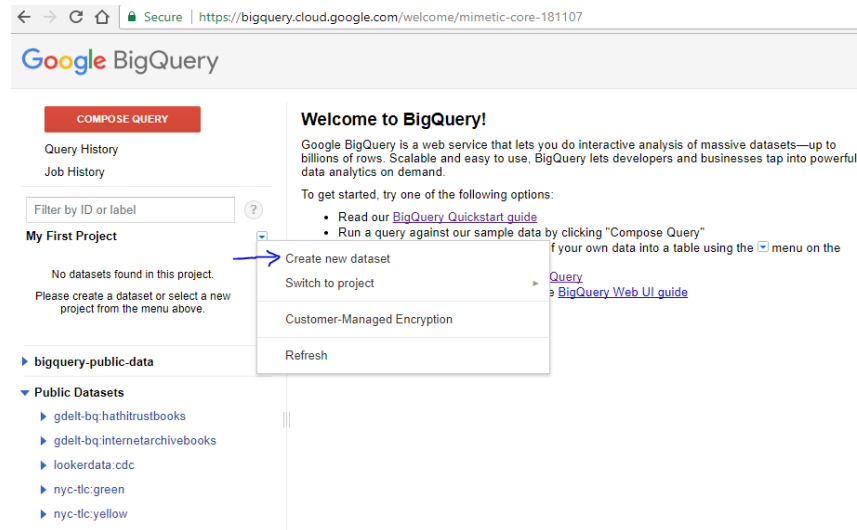


Figure 9. Process to create a new datasets.

After clicking create dataset option below window in figure 10 appears on the screen:

In this Figure 10, the key details such as Dataset ID, Data location and Data expiration details are entered to create a dataset in BigQuery.

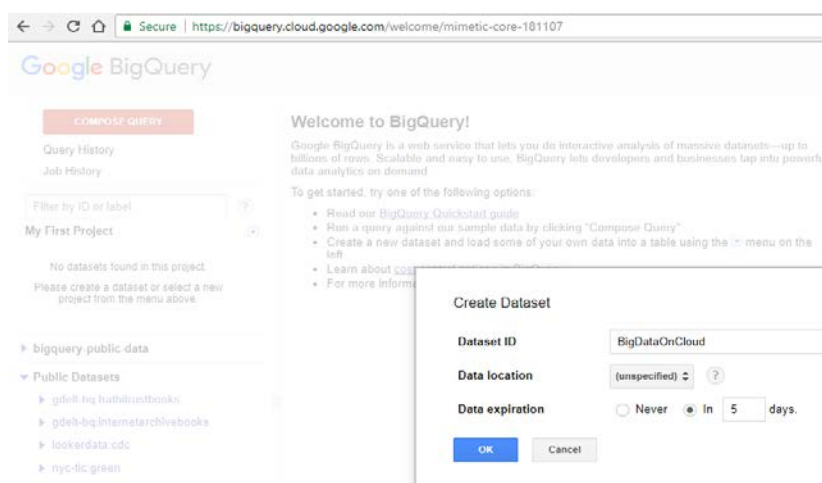


Figure 10. Creating a data set in BigQuery

In Figure 11, more details are added for table creation based on available source data, i.e. CSV file and uploading it from local computer. In next row table name is entered and create table button on bottom of page is clicked to create table in BigQuery Datawarehouse. This step completes Dataset creation process on BigQuery.

The next step is to upload of data source on BigQuery Datawarehouse. In Figure 11, file path is given, which was downloaded from [www.kaggle.com](http://www.kaggle.com) in earlier step in this section.

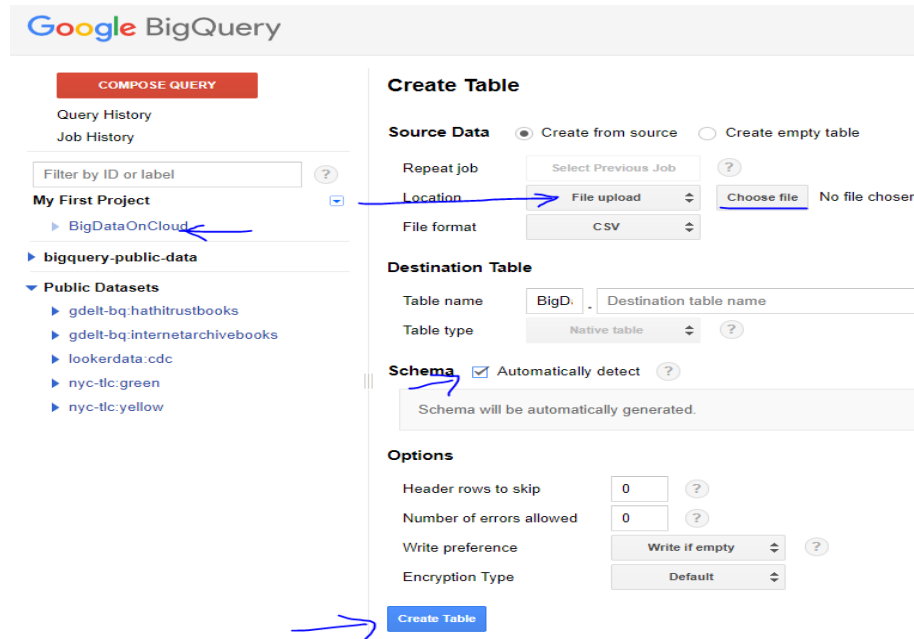


Figure 11. Uploading file to BigQuery Datawarehouse.

In this Figure 12, table name is added which will be used for querying the data.

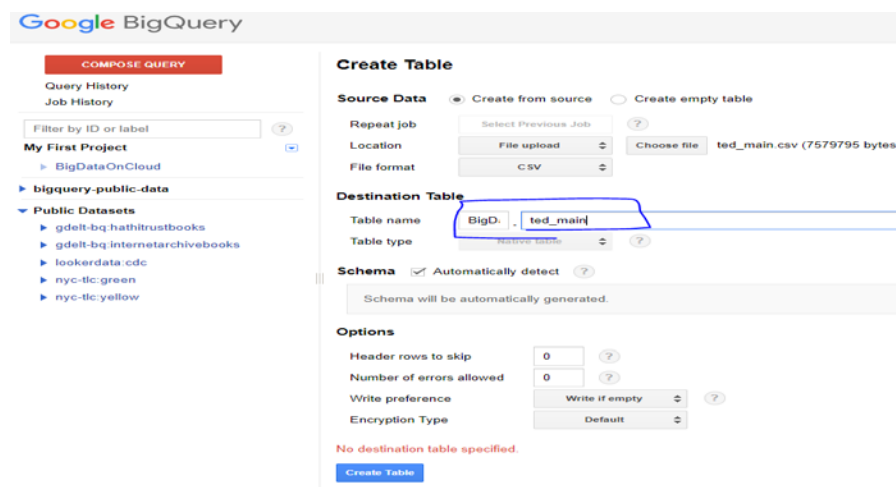
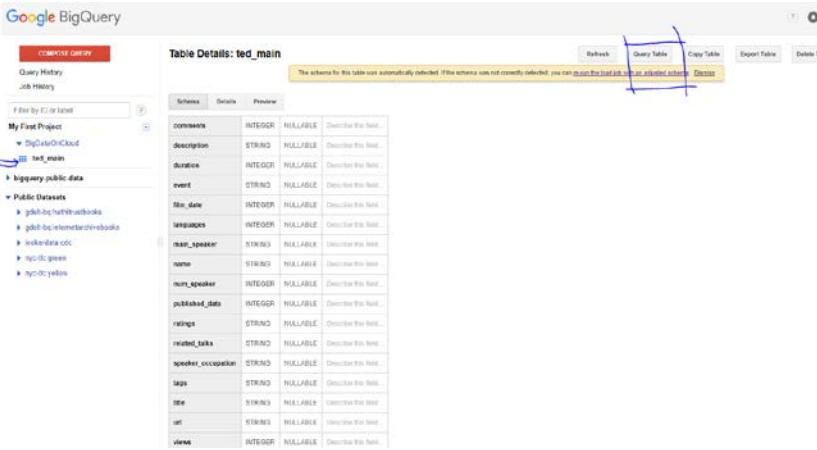


Figure 12. Adding table name

### 3) Querying table in editor

At this stage table is ready for query and finding top 10 topics viewed by maximum count. This is achieved as per below query in  figure 13

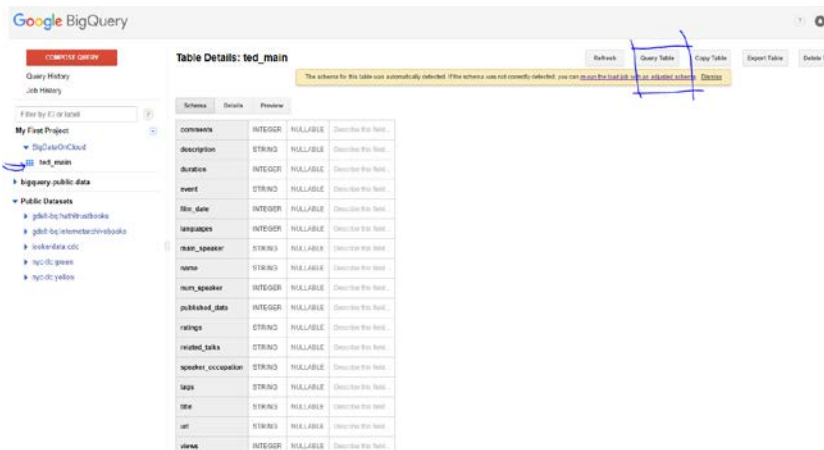


Figure 13. Querying table on BigQuery Datawarehouse on Created Datasets

**Final result:** Click query table as per Figure 13, and writing below SQL query resulted into needed output for finding out top 10 topics in TED event having maximum views:

Select name, views, title, languages from [mimetic-core-181107:BigDataOn-Cloud.ted\_main] order by views desc limit 10;

Below SQL query format also worked for finding out top 10 topics in TED event having maximum views:

SELECT name, views, title, languages FROM BigDataOnCloud.ted\_main order by views desc limit 10.

Figure 14 displays the results after writing SQL query in Query editor:

The screenshot shows the Google Cloud BigQuery Query Editor interface. At the top, there is a 'New Query' header with a help icon. Below it, the SQL query is displayed in a text area: `SELECT name, views, title, languages FROM [aiemtic-core-181107:bigdataoncloud.ted_main] order by views desc limit 10`. A status bar indicates the query is valid and will process 264 KB. Below the query, there are buttons for 'RUN QUERY', 'Save Query', 'Save View', 'Format Query', and 'Show Options'. A message states 'Query complete (1.2s elapsed, 264 KB processed)'. The results are displayed in a table with columns: Row, name, views, title, and languages. The 'views' column is highlighted with a blue box. Below the table, there are options to 'Download as CSV', 'Download as JSON', 'Save as Table', and 'Save to Google Sheets'. The table footer shows 'Table JSON' and 'First < Prev Rows 1 - 9 of 10 Next > Last'.

Row	name	views	title	languages
1	Ken Robinson: Do schools kill creativity?	47227110	Do schools kill creativity?	60
2	Amy Cuddy: Your body language may shape who you are	43155405	Your body language may shape who you are	51
3	Simon Sinek: How great leaders inspire action	34309432	How great leaders inspire action	45
4	Brené Brown: The power of vulnerability	31168150	The power of vulnerability	52
5	Mary Roach: 10 things you didn't know about orgasm	22270883	10 things you didn't know about orgasm	37
6	Julian Treasure: How to speak so that people want to listen	21594632	How to speak so that people want to listen	45
7	Jill Bolte Taylor: My stroke of insight	21190883	My stroke of insight	49
8	Tony Robbins: Why we do what we do	20685401	Why we do what we do	36
9	James Voitch: This is what happens when you reply to spam email	20475972	This is what happens when you reply to spam email	43

Figure 14. Query output

### 5.3 Results of Demonstration

This demonstration highlights how easy is to start an analytics project on cloud platform. This demo has taken sample data available from one of the publicly available data sets on website [www.kaggle.com](http://www.kaggle.com). The size of CSV file was small of size of 12 megabytes. Creating a cloud account was much easier on Google cloud platform and process to upload CSV file was quite straightforward. During the upload process of CSV file, a table is created in BigQuery' s Datawarehouse. Finally, a SQL query displayed the results of top 10 topics with maximum view count. It was observed from this demo that getting on boarded on cloud is fairly quick and easy. GoogleQuery is Datawarehousing solution from Google cloud platform.

## 6 Conclusion

Data is a new capital for enterprises and organisations. The analysis process ensures that hidden facts are highlighted for better decision making. Storing and processing RDBMS data was good enough solution in the previous decade. With the evolution of data types and its volume and velocity created a new phenomena named as Big data. Traditional servers and databases have limitations to process these Big data sets and thus evolution to Cloud computing begun.

Cloud computing is a platform, which is flexible, efficient and scalable and adds strategic value to organisations of all scale. Executing Big data analytics projects require scalable computing as well as storage to process a huge amount of data. Cloud computing with examples such as Novartis, it has been 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 as without it the credibility can be on stake should there be a downtime due to production failures. Cloud computing offers disaster recovery capability to handle major accidents and recovering faster as stated in agreed SLA's. One can start small and grow big through Cloud computing platform, as one strong feature of cloud services is 'pay as you grow'. This is even better for start-ups 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 these needs.

Amazon has been leading the industry in Cloud computing services but other players such as Azure, Google, Alibaba, IBM, Fujitsu, Oracle etc. are catching up fast to provide huge opportunity for end customers. From long list of cloud providers this study covered Big data analytics products from top three cloud service providers i.e. Amazon Web Services, Microsoft Azure and Google Cloud Platform.

Big data analytics projects are implemented in all business areas but an important question is to find out the best solution that meets the requirements of the company and the project. The Value is the most important character in all Big data projects since all other characters, Volume, Velocity, Variety and Veracity helps deriving the hidden information. The demo highlighted one example where a publicly available datasets was analysed on Google's BigQuery Datawarehouse service. This demo has demonstrated the comfort of starting an analytics project on Cloud computing platforms almost instantly.

For all the analytics project the most important question lies in value which needs to be created from available datasets.

|



## References

1. Cloud Computing for large scale data analysis  
<https://melmeric.files.wordpress.com/2010/02/thesis-proposal.pdf>
2. The-benefits-of-cloud-computing-for-the-enterprise  
<https://cloudacademy.com/blog/the-benefits-of-cloud-computing-for-the-enterprise/>
3. Final-version-nist-cloud-computing-definition-published  
<https://www.nist.gov/news-events/news/2011/10/final-version-nist-cloud-computing-definition-published>
4. NIST Cloud Architecture  
[https://www.researchgate.net/figure/273945605\\_fig1\\_Figure-1-NIST-Cloud-Computing-Reference-Architecture-CCRA-2](https://www.researchgate.net/figure/273945605_fig1_Figure-1-NIST-Cloud-Computing-Reference-Architecture-CCRA-2)
5. Cloud-carrier  
<https://www.trustingthecloud.eu/joomla/index.php/cloud-carrier>
6. Cloud Computing service models: A comparative study  
<http://ieeexplore.ieee.org/document/7724392/>
7. Private Clouds  
[http://whatiscloud.com/cloud\\_deployment\\_models/private\\_clouds](http://whatiscloud.com/cloud_deployment_models/private_clouds)
8. Cloud Service and Deployment Models  
[https://cloudcomputing.ieee.org/images/files/education/studygroup/Cloud\\_Service\\_and\\_Deployment\\_Models.pdf](https://cloudcomputing.ieee.org/images/files/education/studygroup/Cloud_Service_and_Deployment_Models.pdf)
9. Public Clouds  
[http://whatiscloud.com/cloud\\_deployment\\_models/public\\_clouds](http://whatiscloud.com/cloud_deployment_models/public_clouds)
10. Hybrid Cloud  
[http://whatiscloud.com/cloud\\_deployment\\_models/hybrid\\_clouds](http://whatiscloud.com/cloud_deployment_models/hybrid_clouds)

11. NIST definition Community Cloud  
[https://cloudcomputing.ieee.org/images/files/education/studygroup/Cloud\\_Service\\_and\\_Deployment\\_Models.pdf](https://cloudcomputing.ieee.org/images/files/education/studygroup/Cloud_Service_and_Deployment_Models.pdf)
12. Community Cloud  
[http://whatiscloud.com/cloud\\_deployment\\_models/community\\_clouds](http://whatiscloud.com/cloud_deployment_models/community_clouds)
13. Gartner\_confirms\_what\_we\_all\_know  
[https://www.theregister.co.uk/2017/06/19/gartner\\_confirms\\_what\\_we\\_all\\_know\\_aws\\_and\\_microsoft\\_are\\_the\\_cloud\\_leaders\\_by\\_a\\_fair\\_way/](https://www.theregister.co.uk/2017/06/19/gartner_confirms_what_we_all_know_aws_and_microsoft_are_the_cloud_leaders_by_a_fair_way/)
14. About AWS  
<https://aws.amazon.com/about-aws/>
15. AWS Global Infrastructure  
<https://aws.amazon.com/about-aws/global-infrastructure/>
16. What is Azure?  
<https://azure.microsoft.com/en-us/overview/what-is-azure/>
17. Google Cloud Platform  
[https://en.wikipedia.org/wiki/Google\\_Cloud\\_Platform](https://en.wikipedia.org/wiki/Google_Cloud_Platform)
18. Compare-AWS-vs-Azure-vs-Google-big-data-services  
<http://searchcloudcomputing.techtarget.com/tip/Compare-AWS-vs-Azure-vs-Google-big-data-services>
19. Analytics Services with AWS  
[https://aws.amazon.com/products/analytics/?nc2=h\\_I3\\_db](https://aws.amazon.com/products/analytics/?nc2=h_I3_db)
20. Azure products Intelligence-Analytics  
<https://azure.microsoft.com/en-us/services/?filter=intelligence-analytics>

21. Big Data solutions  
<https://cloud.google.com/products/big-data/>
22. Functional architecture for Cloud Computing and Big Data  
<https://www.itu.int/en/ITU-T/studygroups/2017-2020/13/Pages/q18.aspx>
23. Big Data Architecture  
[https://en.wikipedia.org/wiki/Big\\_data#Architecture](https://en.wikipedia.org/wiki/Big_data#Architecture)
24. Big-data-analytics  
<https://www.techopedia.com/definition/28659/big-data-analytics>
25. Big Data Case Studies  
[https://en.wikipedia.org/wiki/Big\\_data#Case\\_studies](https://en.wikipedia.org/wiki/Big_data#Case_studies)
26. How-our-likes-helped-trump-win  
[https://motherboard.vice.com/en\\_us/article/mg9vvn/how-our-likes-helped-trump-win](https://motherboard.vice.com/en_us/article/mg9vvn/how-our-likes-helped-trump-win)
27. IoT Mid-Year Update From IDC And Other Research Firms  
<https://www.forbes.com/sites/gilpress/2016/08/05/iot-mid-year-update-from-idc-and-other-research-firms/#33425d9755c5>
28. 6 Predictions For The \$203 Billion Big Data Analytics Market  
<https://www.forbes.com/sites/gilpress/2017/01/20/6-predictions-for-the-203-billion-big-data-analytics-market/#65172c612083>
29. Benefits-of-cloud-computing  
<http://dataconomy.com/2014/11/7-benefits-of-cloud-computing/>
30. Novartis Case Study  
<https://aws.amazon.com/solutions/case-studies/novartis/?hp=tile>
31. AWS Case Study: Financial Times  
<https://aws.amazon.com/solutions/case-studies/financial-times/>
32. Dow Jones Case Study  
<https://aws.amazon.com/solutions/case-studies/dow-jones/>

33. Shipping company navigates to the cloud, boosts security, cuts operating costs by 50 percent  
<https://customers.microsoft.com/en-us/story/nakilat-process-mfg-resources-microsoft365>
34. University Improves Operational Efficiency, Cuts Energy Consumption by 30 Percent with BI Solution  
<https://customers.microsoft.com/en-us/story/university-improves-operational-efficiency-cuts-energy>
35. Why-move-to-the-cloud-10-benefits-of-cloud-computing.html  
<https://www.salesforce.com/uk/blog/2015/11/why-move-to-the-cloud-10-benefits-of-cloud-computing.html>
36. Cloud Business Benefits  
<https://www.wired.com/insights/2012/10/5-cloud-business-benefits/>
37. Cloud Computing and Is it Really All That Beneficial?  
<https://www.lifewire.com/cloud-computing-explained-2373125>
38. 11 Advantages of Cloud Computing and How Your Business Can Benefit From Them  
<https://www.skyhighnetworks.com/cloud-security-blog/11-advantages-of-cloud-computing-and-how-your-business-can-benefit-from-them/>
39. Cloud Service Models  
<https://rajivramachandran.wordpress.com/2012/06/19/cloud-service-models-iaas-vs-paas-vs-saas/>
40. Inside Facebook's Blu-Ray Cold Storage Data Center  
<https://datacenterfrontier.com/inside-facebooks-blu-ray-cold-storage-data-center/>
41. Four-common-cloud-attacks-and-how-to-prepare-for-them  
<https://searchcloudsecurity.techtarget.com/tip/Four-common-cloud-attacks-and-how-to-prepare-for-them>

42. Four-common-cloud-attacks-and-how-to-prepare-for-them

<https://searchcloudsecurity.techtarget.com/tip/Four-common-cloud-attacks-and-how-to-prepare-for-them>

43. Amazon\_Elastic\_Compute\_Cloud

[https://en.wikipedia.org/wiki/Amazon\\_Elastic\\_Compute\\_Cloud](https://en.wikipedia.org/wiki/Amazon_Elastic_Compute_Cloud)

44. TED\_(conference)

[https://en.wikipedia.org/wiki/TED\\_\(conference\)](https://en.wikipedia.org/wiki/TED_(conference))