A Study of Big Data in Cloud Environment with their Related Challenges
Anshu Dhabhai1, Yogesh Kumar Gupta2
Department of Computer Science
Banasthali Vidyapith, Jaipur, Rajasthan, India

Abstract: Big Data, it is a word which is used for the huge amount of data sets which have the big, extra various and difficult structure facing some difficulties in storing the data then analyzing and visualizing the data for the final outcome [1]. Big data analytics is the procedure of probing huge amounts of data. Here we use cloud computing technology so firstly acquaint the general environment of big data and some associated technology like Hadoop. This powerful technology is used to execute the huge scale and complicated data computing. It eradicates the requirements of hardware computing which is costly dedicated space and software’s [3]. It is a resolution for analyzing Big Data by deploying Big Data softwares like Hadoop on cloud background which gives users an innovative pattern of service called as Analytics as a Service [7]. Hadoop uses the map-reduce paradigm. We also talk about various advantages, challenges, and issues associated with Big Data.

Keywords: Big Data, cloud computing, Data Mining, challenges, analytics tools.

I Introduction

Big Data be a new concept present in IT companies, industries, academics and medical science and continuously increasing in the volume and the detail of some data are structured and mainly non-structured crossed the order of petabytes, Exabyte, and zetabyte and further rising year by year are generally referred as Big Data. These data obtained by health data records, science data records, online transactions, videos, email, images, audios, posts, logs, searching queries, some social networking sites, sensors, mobiles with their apps and must update time to time quickly. Big data are characterized with the help of (1) various data (2) it cannot classify in relational databases (3) data are originated, catches and carry out quickly for a final result [3]. They are saved in databases which are expanding profoundly and turn into very complicated to an acquisition, storage, formation, managing, analyzing, contributing and visualizing through a relational database and some software tools [1]. Handling, Managing, analyzing the huge data to get insight knowledge is a large challenging task. Nowadays almost every individual is carrying the mobile phone with some data plan and these mobile phones are one of the biggest sources to generating the real-time data through mobile sensors. Not only mobile but sensors are embedded in all type of things like vehicles, humans, and devices like home TV, hospitals which generates networks connected with the internet infrastructure and generating huge data [7]. Cloud computing has a great structural design to execute big-scale and complicated data. The cloud computing advantages contain virtual resources, security, data service combination and parallel computing with scalable storage space [3]. Cloud computing offers compact infrastructure maintenance cost, user access, and proficient management along with less charges and automatic control by individual or enterprises. Virtualization is the single base technology to implement cloud computing. Virtualization is a procedure of sharing the resources and separates the hidden hardware for increasing the utilization of computer resources, scalability, and efficiency [3]. Cloud computing provides an attractive model for analytics, where consumers pay the only total amount of services they used.

Big Data Characteristics

It is a word refers to inflation of the contents that are not easy to save, process, examine with typical database technologies. Big data needs novel techniques and technologies to form new patterns of combinations to find out the largely hidden values from the huge collections of data sets that are massive, dissimilar and complex.

Figure 1 – Four V’s of Big Data [3]

(1) Volume- It is used to describe the size of data or the total quantity of different kind of data produced from various sources so continuously extends. The Advantage of accumulation of the huge data quantity is that it contains the formation of secret facts, patterns from an examination of data [3].

(2) Variety- variety makes big data actually big. It mentions the various kinds of data that are composed through social communications, sensors, mobile phones which contains video, audio, text, image, and data logs that are structured, unstructured and semi-structured format [3].

(3) Velocity- It speaks about the data transmission rate, the variations in the data contents continuously growing by reason of the complexity in data accumulation, presentation
of formerly collected data and stream data incoming from different sources [3].

(4) Value - It is very valuable characteristic of large data. It mentions the procedure of uncovering the large unseen information from huge datasets by different kind of data such as statistical, events and hypothetical values [3].

**Big data Classification**

Big data categorized into classes’ greater understanding of their components. The classification based upon five classes: (1) sources of data, (2) formats of contents, (3) storage of data, (4) enhancement of data, and (5) processing of data [3].

**Figure 2 - Big Data classifications [3]**

**II Big Data in Cloud Environment**

“Cloud computing is a model for allowing ubiquitous, convenient, and on-demand network access to a number of configured computing resources (e.g., networks, server, storage, application, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction”. It provides the access to applications as utilities above the internet. Big data requires the bunch of servers to support the tools and technologies that process on huge volume, the variety of formats and high velocity. Clouds already have the pool of servers, networking resources, and storage that can scale up and down when desired. It offers the cost-effective way to hold big data technologies.

In cloud computing, we have cloud server. With the cloud server, many of the Cloud service providers are connected and they all are works parallel to each other and manage/balance the load.

**Figure 3 - Load balancing on Cloud Server**

In Big data, job tracker allots the jobs to the Task tracker and HDFS (Hadoop distributed file system) works along the Job tracker. So with the job tracker many of the task trackers are connected and work parallel and the load is maintained at the each node after that we use the MapReduce function.

**Figure 4 - Load balancing on Job Tracker**

So in both framework, the work are done in parallel and load balance also maintained, that’s why we can easily run the big data on cloud framework.

**Figure 5 - Distributed Environment of HDFS**

**Cloud Service Model**

Cloud computing provides some cloud service models: (1) software as a service (Saas), (2) platform as a service (Paas), and infrastructure as a service (Iaas) to the users (can be developer or researchers) to apply some data conclusive solution in the cloud.

**Data analytics Saas** - here the applications are delivered software as a service over the internet. It provides quick knowledge discovery technique and clear data mining algorithm to the client as an Internet service. In this level deploying the software without concern about the installing and maintaining and a user can simply access via the internet [2], [7]. Example- Gmail.

**Data analytics Paas** - it supports a platform to the builders for building their own data analytics applications or we can say that it provides libraries and tools to developers to build, deploy, test and run applications on the cloud environment [2],[7]. For example- Google app Engine.

**Data analytics Iaas** - It supports the set of virtual resources as computing infrastructure to the developers to execute their applications and run data analytics [2], [7]. Example- Amazon EC2.
III Big Data Analytics Tools

Hadoop Framework

Hadoop is generally an open-source Programming framework which provides support for the huge data sets processing in a distributed environment. It created by Google’s MapReduce. MapReduce is also a framework which uses divide and conquer method to break down an application in different fragments. Hadoop has two prime components that are Map Reduce programming framework and HDFS [9],[4].

Map Reduce Programming Model

Map Reduce it is also a software programming structure used for scattered computing environment. It was developed on the basis of GFS (Google File System) and adopted by an open-source Hadoop implementation. It uses the method of divide and conquers to divide the difficult big problems into their small sub problems and execute each of them in parallel. It works in two phases –

- **Map phase**: here large data contains the master node is splits into several smaller problems. Then smaller problems subsets are processed by worker node and it is controlled by job tracker node and for saving the results it used local file system so reducer can easily access it [9],[4].

- **Reduce phase**: it accesses the results comes from the map steps. those results are analyzes and merges. Then the process of multiple reduces tasks aggregations are performed on worker nodes, and all tasks are executed to the Job Tracker control [9],[4].

![Figure3-Processing of Map Reduce programming Model](image)

Hadoop Distributed File System

HDFS stands for Hadoop distributed file system, it planned to execute on the local file system so the extremely large files and cluster nodes can easily access the streaming data. It has an extreme power to tolerate the fault and can easily up the scale from one server node to thousands of server nodes when needed; every node provides own storage and estimation. It has the two kinds of nodes - a nave node say as “master node” and some data nodes say as “slave node”. It also contains a secondary nave node. The nave node completely manages the whole hierarchy of director namespace and file systems [9].

IV Benefits of Big Data

There are some benefits when the organization applied the big data analytics -

- Client-based analysis
- Identification of market and sales chances
- Computerized decisions for real-time processes
- Risk Quantification
- Better knowledge of business changes
- Better scheduling and forecasting
- Fraud detection
- Detection of consumer behavior from click streams etc [10].

V Big Data challenges

Scalability

It is the capability to handle the increasing volume of data storage in a proper manner. The critical portion of cloud infrastructure is scalable distributed data storage system so to handle the storage and increasing amount of data we have a database that is NoSql database. This NoSql database contains schema-free, simple API, easy replication support, flexible and reliable nodes. There is different kind of NoSql databases like the key-value database, document-oriented database, and column-oriented database to support big data [3].

Availability

It speaks about the availability of the resources ready for use to make a request by the client or some authorized person. CSP concern about the availability of data storage, which is the principle issue in the cloud environment. The cloud users are increasing and they create a request for data so the CSP needs to deal with this issue of providing the high-quality data services to the clients.

Data integrity

Big data security has a feature of integrity. Integrity explains incorruptibility, it means data cannot modify by everyone and only modified by the authorized person. The cloud applications give the chance to users to manage as well as store the data in data centers of cloud and have a guarantee of data integrity. The correctness of the user data is a major challenge in the cloud [3].

Data quality

There are different sources of originating data and all sources are not verifiable or well-known. Poor quality of data has made a sober difficulty for several CSP in data collecting from various sources. It is a challenge to generating high data quality from huge sets of data origin.

Heterogeneity

Variety is the big data characteristics have major aspects of the advancement in the data from various sources. This expansion leads the diverse nature of data. Users stored their structured data, semi-structured data, or unstructured data format in the cloud environment. Today’s the formats of structured data are suitable in a database, but the formats of
semi-structured data are suitable for few level and format of unstructured data are not suitable because of the complicated format [3].

Privacy
In cloud computing the privacy, as well as security are major challenges. It worries about the deficiency of faith and CSP, deficient information regarding the SLAs and some probability of various attacks on different locations that are few challenges, which cannot support cloud infrastructure. Dealing with fault tolerance is a key challenge that related to the use of cloud computing infrastructure along the capability to discover and rediscover not including the little casualty in the final outcome.

VI Big Data Issues
The amount of data continuously increasing day by day but there is not any progress in the processing mechanisms. In an environment in the cloud only some tools exist to handle the processing issue of big data.

Data staging
Staging means the enhancement of data after gathering the data from different sources have different nature; all data are not in the same structural format. For example, mobile applications, social networking, and blogs are incorrectly structured like pieces of text messages, images, and videos. For analysis, transposing and purification of such kind of data is very challenging tasks before storing them in a warehouse.

Analysis of Big Data
The selection of a suitable model for analysis the huge data is dangerous. To process such data we required some effective tools for analysis of data. The rate of data stream coming from various sources of data needs to be processed and distinguish with the previously stored data in a definite interval. Data analysis is a complex task which integrates the data from various sources has dissimilar formats.

Security of Big Data
Many unresolved security threats exist and they are exaggerate by the variety of big data, the velocity of data and increasing volume of data. When we are using cloud computing with big data, many of threats are present in the data confidentiality, availability, integrity and data privacy. In cloud environment, we need powerful cryptography, to sum up the unstable data. So we need to develop a new algorithm to manage the keys and securely exchange the keys between two parties [8].

VII Conclusion
In present time the size of data is continuously increasing having variety, velocity, volume and value. We present a review of the big data that are the rise in a cloud environment. Here we design the big data classification, and big data theoretical vision, and service models of a cloud. We also discuss the Hadoop background technology and its main components – HDFS and MapReduce. We also review of some big data processing challenges which wrap volume, availability, data protection, scalability, an integrity of data, quality of data, heterogeneity in data, and privacy of data. And some related issues are described.

VIII References