IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-7, July 2018 E-ISSN: 2347-2693

# Big Data and Hadoop Computational Challenges- A Comparison Study

**Mamoon Rashid[1*], Mir Aamir[2], Balwant Ram[3]**

[1,3]Assistant Professor, School of Computer Science & Engineering, Lovely Professional University, Punjab, India.
[2]Research Scholar, School of Computer Science & Engineering, Lovely Professional University, Punjab, India.

[*]*Corresponding Author: mamoon873@gmail.com*

**Available online at: www.ijcseonline.org**

Accepted: 06/Jul/2018, Published: 31/Jul/2018

***Abstract*:** Big Data refers that data which cannot be stored and processed efficiently by traditional techniques. Storage and processing of big data is a real challenge. To solve the big data problem, Hadoop provide an optimum framework. Hadoop framework is based on java and it can process data in large quantities. It is able to process unstructured data as well and that too in a distributed manner. In this paper, the authors have studied different dimensions of Big Data and tried to provide the comparison study existing between the resource findings related to it. The authors have further highlighted the various challenges where researchers need to put focus to enhance functionality and power of Big Data Computation.

***Keywords:*** Big Data, Hadoop, HDFS, Map Reduce, Security

## I. Introduction

Big Data is a concept and a concept can be interpreted in many ways. We can say that Big Data means that data which cannot be handled by traditional database processing techniques or that is difficult to process using traditional database and software techniques. Any data which cannot be processed is big data in that scenario. Usually the term big data is associated with data in terabytes or more.

In today's era, huge volumes of data are generated every day. Data is generated from multiple sources that include social networks, physical sensors, aircraft data, stock exchange data, data from experiments etc. However if we are unable to process this data, it is of no use. This data should be properly and efficiently stored and then processed so that we can analyse the underlying patterns. These patterns are of great importance in business decisions, elections, marketing strategy etc.

Characteristics of Big Data
Big Data is characterised by some parameters. Mainly they are:
a) Volume
Volume means the quantity of data generated. The parameter of size is very important. Usually this parameter determines what is the value and potential of the data. Various researches show that data volumes are increasing year-by-year. This data contains hidden information, so organizations now store a significant amount of historic data.

**b) Variety**
Another important aspect of Big Data is the variety. Nowadays data is generated in heterogeneous types i.e. both structured and unstructured data is generated. Researches have also indicated that larger portion of this data is unstructured. Managing this data is tedious task for organizations. Hence the category to which data belongs to is a very important factor that data analysts should know as it helps in effectively using the data.

c) Velocity
Velocity means the speed at which data is generated and can also refer to speed at which it needs to be processed. Nowadays data is being generated at humongous speeds. Hence, the speed at which data is generated and needs to be processed is increasing.

There are other parameters too that characterise big data. Some of them are:
a) Variability
The amount of data generated nowadays is very much variable. Variability refers to consistency in data. Data can show inconsistency sometimes which hinders the process of efficient management and utilization of data.

b) Value
Value refers to the quality of data i.e. how important the data is. Big data has enormous value. Value is an extremely important parameter for big data. Value can refer to the important and valuable information that is extracted. Value is most important factor for Big Data Analytics.
**c) Veracity**
Veracity refers to the accuracy of data. Every data has its inherent abnormalities. When the data volume is large, these abnormalities tend to grow. Good veracity will result in good data analysis. Veracity in data analysis is the main challenge relative to volume and velocity. In order to analyse the data, we need to clean the data to get meaningful and useful results.

d) Volatility
Volatility refers to the timeframe in which data is valid and how long should it be stored. Data is generated in huge volumes daily and we need to determine when a particular dataset is not important.

Challenges in Big Data
Big data is a real challenge in today's technological oriented world. It needs to be analysed before its value expires. The main challenges are:
a) How to store the large volume of data.
b) How to determine which data is important and needs analysis.
c) How to process this data then.
d) How to manage cost involved in Big Data storage and processing.
e) How to determine for how long particular data important.

Where Big Data finds its usage
Big Data finds diverse usage in all fields. Some areas where we can find useful applications related to big data are Social Media, Healthcare Systems, Defence Systems, Banking Sector, Universities etc.

## II. HADOOP

Hadoop is a framework based on java which has the capability to process large amounts of data. It is a part of the apache project of the Apache Software Foundation. Hadoop can be referred to as an engine for processing extremely high volumes of data in any structure. One of the major benefits of Hadoop is its ability to process complex and unstructured data. Hadoop works in a distributed computing environment.

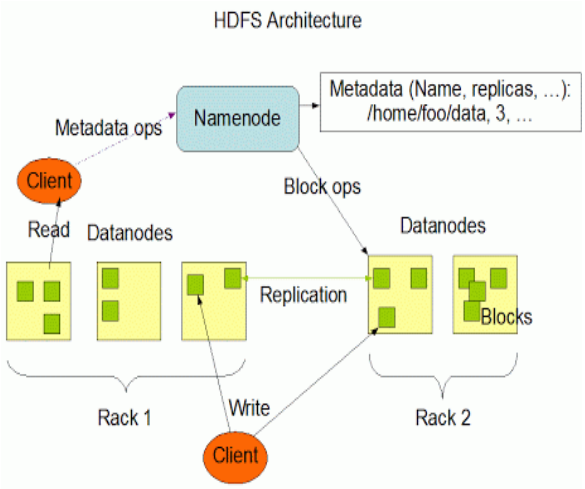
Components of Hadoop
Two major components of Hadoop are Map Reduce and the Hadoop Distributed File System (HDFS). HDFS provides the storage portion of the Hadoop framework, while MapReduce is used for analysis. The Hadoop framework makes use of these two components to distribute data across multiple servers and process them parallely.

**Hadoop Distributed File System (HDFS)**
HDFS is a filesystem for the Hadoop framework. It is distributed, scalable and portable. It is a fault-tolerant storage system. HDFS has scale-out capabilities because of two main components: Name Nodes, which is responsible for management of file system and DataNodes, which are responsible for storing and retrieving the data. Depending on the design of HDFS, the files are stored on many nodes redundantly to ensure high availability.

These nodes operate in a master slave relationship. The NameNode reflects the master and keeps track of all the DataNodes where the blocks for a given file are located. Data nodes also update the NameNode with lists of the actual blocks that they are responsible for.

HDFS has to great qualities, scalability and fault-tolerance. It is due to its ability of scaling out with the adding of commodity servers. These servers hold the data being collected as well as multiple copies to ensure reliability.

Figure 1: HDFS Architecture




HDFS Cluster
Hadoop has many nodes grouped in clusters and work is distributed manner between them. If one fails, Hadoop continues to operate the cluster without losing data or interrupting work, by shifting work to the remaining machines in the cluster. HDFS manages storage on the cluster by breaking incoming files into pieces, called blocks and storing each of the blocks redundantly. In the common case, HDFS stores three complete copies of each file. Each node in a Hadoop instance has a single datanode; a cluster of datanodes form the HDFS cluster.

Map Reduce
The MapReduce is a software framework which is associated with processing of data through a cluster of nodes. The framework divides the problem and data, and runs it in parallel. In Hadoop, these operations are written as MapReduce jobs. The outputs of these jobs is written back to the HDFS.
There are two functions in MapReduce: Map function and Reduce function. Every map and reduce functions are independent of each other. The processing occurs on those nodes that hold the data for that particular function. The Map divides a large job across the worker nodes of a cluster. Multiple map functions can be executed at once. The reduce function then takes the output of the map functions, and does some processing on them to generate the desired result.
When a MapReduce program is run by Hadoop, it sends the job to a master node, the JobTracker, which has multiple slave nodes or TaskTrackers. The JobTracker divides the map tasks amongst the TaskTrackers, so that they all work in parallel. Also, the JobTracker keeps track of which TaskTrackers fail, so their tasks are redistributed to other task trackers.
The Map function is divided into two steps. First a key is assigned (K1) using which data is. The output from this creates

a list based on another key (K2), which is a group key for all the pairs of the same key.
Reduce then aggregates the list based on K2. It combines outputs of every map. It processes the inputs from the map functions and creates a new list to form the final output.

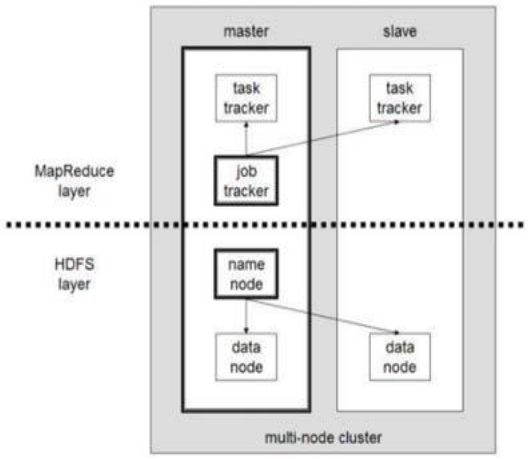



Figure 2: Nodes in a Hadoop Cluster

### III. RELATED WORK

Seref Sagiroglu et.al. [1] did a review on big data. According to them, big data processing is the main thing in businesses. Big data is generated from multiple sources like online transactions, emails, videos, audios, images, click streams, search queries, health records, social networking interactions, sensors and mobile phones. So databases grow massively and it becomes difficult to store, manage, share, and analyze data using traditional database software tools.
Various components of big data are variety, volume and velocity. Big data comes from a great variety of sources and generally is of three types: structured, semi
structured and unstructured. Volume or the size of data now is larger than terabytes and petabytes. This calls for a new approach as traditional methods will not work with this large data. Velocity is required not only for big data, but also all processes. It means the rate at which data is generated and also the rate at which data is processed.
Big data is applicable in every field like genomics, biological science and research, life sciences, medical records, scientific research, government, military surveillance, social networks, text, document, photography, audio, video, click streams, call detail records, mobile phones and sensor networks. But five main areas are healthcare, public sector, retail, personal location data and manufacturing.
They also discussed about map reduce. Map Reduce can be divided into two stages:
Map Step: The master node data is sliced into many smaller sub-problems. A small piece of sub problems is given to worker node for processing and it is controlled by Job Tracker.
Reduce Step: This step analyses and merges input data from the map steps. Multiple reduce tasks run parallely on the worker nodes to speed up this merging.
They discussed what tools Hadoop offers. It includes HDFS, Map Reduce, Flume, Zookeeper, HBase, Pig, Hive, Sqoop, Oozie and Flume.

According to Dinesh D. Jagtap et.al.[2] Big data is data whose volume is too large to be processed using traditional methods. He also discussed how the data is getting big as time goes by. According to him, managing and replication across many database servers is too slow. MapReduce is a divide and conquer technique which can be used. MapReduce takes a job to be performed, divides it, and runs it parallely.
According to a cloudera white paper [3] there are various challenges of big data. They include scalability, locality and flexibility. Hadoop has a flexible storage, scalability, good performance, cost efficiency and security. HDFS has two main internal components. HDFS consists of Name Node and Data Nodes. Name Node responds to the request from the client by giving locations of Data Nodes. Data Nodes handle all storage and retrieval of data. HDFS automatically replicates data for fault tolerance.
According to Sreedhar C, Dr. D. Kavitha et.al. [4] Hadoop is a free framework which is able to process large data sets in a distributed computing environment. Two key components of Hadoop include Map Reduce and the Hadoop Distributed File System (HDFS). The Hadoop framework makes use of these two components to distribute data among many nodes and process them parallely. It processes huge amounts of data and also processes heterogeneous types of data. The main benefit of Hadoop is its ability to process unstructured data.
HDFS cluster is formed by a number of data nodes. HDFS allows scale-out which provides fault tolerance and scalability. HDFS provides a mechanism of splitting the data into many pieces and provides methods to access this
data so that the processing logic is distributed as well; each and every node contains a piece of the file and logic. HDFS also provides cluster rebalancing which helps in fully utilizing the data nodes.
HDFS focuses on fault tolerance. It provides replication of data. Also each data node sends a signal to name node. Absence of this signal prompts the name node to deem the data node as failed.
In HDFS each cluster contains of one name node which controls many data nodes; like a master-slave architecture. The file system is maintained by the name node. When a file is created, it is split into blocks of data, each going into the same or separate data nodes. Name node maps and store the mapping of each block of data. Currently, HDFS only supports a single writer and multiple reader architecture which helps to prevent data corruption.
Map Reduce is the software framework that is associated with processing of data through a cluster of nodes. The Map Reduce system is divided into two parts: Map and then Reduce. Map divides job and distributes it to different worker nodes. The reduce function combines all results and in a sense reduces them into one complete whole. This final result is the answer to

the original job submitted by client. The master node sends a signal to the worker node for a response. Failure to respond means that worker node is unavailable. The master node will then redistribute this worker node's assigned work to another worker node.

Mapping takes the unstructured data. The results of the mapping is used as input for the reduce function. MAP function is actually divided into two steps. It first assigns a key to sort all the data and then a processor to process the data sets using those keys. The output from this creates a list based on another key. Reduce then takes this list based on second key and aggregates them. It processes the inputs from the map functions and creates a new list to form the final output.

Components of Map Reduce include the input reader, partition function and a comparison function. The input reader takes all the data from the source and generates key/value pairs for the input of the map function. It splits the data into for each map function as well. The partition function distributes the output of the map function to the reducers. In the meantime, the comparison function sorts the input of the map function to make it ready for the reducers.

Hadoop has many benefits over traditional databases. They include scalability, fault tolerance and ability to store unstructured data. However it is inefficient for small amounts of data and when there are less nodes in a cluster.

Yashika Verma et.al [5] did a review on big data. According to them big data has many obstacles to address. Data nowadays is mostly in unstructured form and is growing at a faster rate. The approach of using a powerful but single computer will not work. So we need to divide data into many parts and process them parallely. Also storing this huge amount of data will require lot of resources and money. Also big data is not interoperable and its ETL process is time consuming and expensive.

Hadoop provides a solution for storage and processing. The storage portion of the Hadoop framework is provided by HDFS, while the processing functionality is provided by Map Reduce.

HDFS is a distributed file system and consists of name node and data node. Name node is master and maintains the file system. It also keeps track of all the data nodes. The Data Nodes store and retrieve data when requested by the clients or the Name Node.

Map Reduce is a framework in which large volume of unstructured data is processed across various clusters. The Map Reduce framework works in two main phases i.e. Map phase and the Reduce phase.

In Hadoop clusters are used to speed up data analysis, throughput and provide fault tolerance.

According to **Swarupkumar Shivaji Suradkar [6]** in a cluster, there is one master node and many slave nodes. Master node runs the masters; name node in HDFS and Job tracker in Mapreduce. Slave node runs Data Node for the HDFS and Task Tracker for Mapreduce.

According to Anusha B. Dhakite et.al.[7] big data consists of both structured and unstructured data which is difficult to process by traditional techniques. Big data has characteristics like volume i.e. quantity of data, velocity i.e. rate of data growth, variety i.e. types of data, veracity i.e. accuracy of data, variability i.e. inconsistency in data and complexity.

Prabhjot Singh et. al [8] proposed a new approach to control and monitor the home appliance by integrating two technologies internet of things and the third party cloud server. Internet of things makes ordinary device as smart device and cloud computing provide the easy access to monitor and control smart devices on remote basis.

According to Ranjana Bahri [9] value is also a characteristic feature of big data. Value means the usefulness of data. Big data has challenges to address. It is heterogeneous but machine learning algorithms process only homogenous data. Also data is growing manifold and storage becomes a challenge. Due to large size, processing and analysis of data becomes a difficult task. Also, there is a threat to personal privacy as personal data is being collected and transmitted through networks.

Hadoop is used by most of the organisations to manage large quantities of unstructured data. It has many components but the main two components are HDFS and Map Reduce. Map Reduce is a programming model for processing data.

According to **Prof. R.A.Fadnavis et.al.** [10] Map Reduce & Hadoop are the most widely used models used today for Big Data processing. Two main challenges regarding big data are storage and processing. A map reduce model can be used. The map function takes a key, value pair as input
and a list of intermediate values is obtained. The map function is written in such a way that multiple map functions can be executed at once. The reduce function then takes the output of the map functions and combines the values to generate output. When a Map Reduce program is run by Hadoop, the job is sent to the job tracker, which has multiple task trackers. Job tracker divides the map tasks amongst the task trackers, so that they all work in parallel. Also, the job tracker keeps track of which task trackers fail, so their tasks are redistributed to other task trackers. If some tasks are still running and there are no new tasks left to be assigned to machines that have finished their tasks already, tasks can be moved to them.

According to Simmi Bagga et.al [11] Big Data is data which has volume, diversity, and complexity and it require new techniques and algorithms to manage it and extract knowledge from it. Big data is a voluminous amount of structured, semi-structured and unstructured data that can be mined for information. Main characteristics of big data are volume, variety and velocity. But there can be other characteristics as well. They include veracity i.e. abnormality in data, validity i.e. accuracy of data and volatility i.e. how long data is valid.

Hadoop is a processing engine that can handle high volumes of data in any structure. Hadoop provides distributed storage and distributed processing for very large data sets. Hadoop has two main components:

The Hadoop distributed file system (HDFS), which supports data in any form. HDFS is a distributed file system that provides high throughput access to data.

The Map Reduce programing is a framework for performing distributed data processing and is based on divide and aggregate technique.

HDFS, or the Hadoop Distributed File System, gives the programmer unlimited storage. The advantages of HDFS are that it can be scaled out horizontally and it uses commodity hardware which is cheap.
Hadoop also deal with hardware failures. Since there are large number of nodes, node failure may be common. HDFS replicates the data on multiple data nodes, so data can be restored from a replica if one of the nodes fail.
Map Reduce deals with distributed computing. It divides the processing between different nodes. It is also fault-tolerant. If some of nodes fail, Hadoop by default assigns the incomplete work to another node. It also is able to combine the results of the computation in one place.
A set of nodes running HDFS and Map Reduce is known as a Hadoop Cluster. More nodes in the cluster mean better is the performance. These components of Hadoop provide a solution to handle the problem of big data.
**Deepika P et.al. [12]** studied the Hadoop related tools and techniques. Hadoop has many tools which help it in dealing with data. Ambari maintains the Hadoop
clusters and system admin can use it to manage and monitor a Hadoop cluster. HBase is a distributed database. It stores large quantities of data and provides fault tolerance. Tajo is a data warehousing functionality associated with big data. Map Reduce is a programming framework for processing the data in clusters. HDFS is a distributed file system which stores data in splits redundantly. Hive is also the data warehouse. Its programming is same as sql programming.

**Abhinandan P Shirahatti et.al.[13]** have proposed a system of using twitter data for sentiment analysis using Hadoop. Flume is used to get data from twitter after creating a twitter application. Then it is queried using Hive.

**Sukhpal Kaur et. al [14]** have worked on Big Data News Clusters and managed Web News Big Data and categorizing it on the basis of text and content to provide the accurate news with less running time.

The various kinds of research works related to Big Data and Hadoop are tabularized in Table 1.

## IV. RESEARCH OUTCOMES

Based on the study in literature, the authors summed up the following challenges in Big Data:
1. Dealing with the computation of Increased Data Growth- Most of data is in unstructured form and thus cannot be stored in structured databases. Data is in the form of audio, video, documents and images which can be difficult to analyse and search. Moreover the data stored in unstructured data is huge and a result very difficult to compute. So in real sense, storage of data was never a problem but its computation of stored data.
2. Faster rate of Insights in Big Data- No organization wants to store huge amounts of data but the major crux of storing huge amounts of data is to gain insights and to achieve organisational interests. Data sources are finance, banking, healthcare and insurance. In these areas, faster processing is important to enhance business.
3. Providing security for Big Data- Security of Big Data is one of prime concern for organizations dealing with Big Data. It has been observed that big data sources often turn targets for various treats and hackers.
4. Solution to Data Governance Issue- Organizations dealing with Big Data often get similar information of data from different sources and the agreement of data among such sources often turn to be problem. Hence there is need of strategies and policies which will fix and ensure the data accuracy of these Big Data Sources.

## V. CONLUSION

Data is generated in large quantities and in future, data generation will increase manifold. This leads to problems of data storage and data processing within expected timelines. This Big Data is characterised by many features like Volume, Variety and Velocity. Hadoop offers a way of storing and processing this data. Hadoop has HDFS which stores data in distributed manner; and MapReduce which processes data in parallel over many clusters. Also, it provides fault tolerance, has a distributed architecture, processes data in parallel and can be scaled out. In this paper, the authors have compared various dimensions of Big Data and structured it to get some insights for specific purposes in the related field. The authors have further highlighted the various challenges where researchers need to put focus to enhance functionality and power of Big Data Computation.

Table 1. Summarization of Hadoop and Big Data Research

| S.No | Title | Author(s) | Key Points |
|---|---|---|---|
| 1 | Big Data: A review | Seref Sagiroglu and Duygu Sinanc | Data can be structured, semi structured an unstructured, big data is used in many fields hadoop offers many tools to deal with big data. |
| 2 | Big Data using Hadoop | Dinesh D. Jagtap and Prof. B.K. Patil | Data is growing and will grow rapidly. Ma reduce processes data parallely. |
| 3 | Hadoop and HDFS | Cloudera white paper | Big data has challenges like scalability, localit and flexibility. HDFS has name nodes and dat nodes. |
| 4 | Big Data and Hadoop | Sreedhar C, Dr. D. Kavitha and K. Asha Rani | HDFS allows scale-out which provides fau tolerance and scalability. |
| 5 | A Review Paper on Big Data and Hadoop | Yashika Verma and Sumit Hooda | HDFS deals with storage while MapReduce deal with processing. Map Reduce has Map an Reduce phases. |
| 6 | Analysis of Big Data with Hadoop Framework | Swarupkumar Shivaji Suradkar | A cluster has one master node and many slav nodes. |
| 7 | Big Data: An Overview | Anusha B. Dhakite and Prof Nitin V. Wankhade | Big data characteristics include volume, velocity variety, veracity, variability, complexity. |
| 8 | Smart Home Automation Deployment on Third Party Cloud Using Internet of Things | Prabhjot Singh, Mamoon Rashid | Control and monitor the home appliance by integrating two technologies internet of things and the third party cloud server. Internet of thing makes ordinary device as smart device and cloud computing provide the easy access to monitor an control smart devices on remote basis. |
| 9 | Big Data: Concept, Challenges and Management Tools | Ranjana Bahri | Big Data is heterogeneous but machine learnin algorithms process only homogenous data. Dat is growing, storage becomes a challenge Processing and analysis of data becomes difficult |
| 10 | Big Data Processing using Hadoop | Prof. R.A. Fadnavis and Samrudhi Tabhane | The map function takes a key, value pair as inpu and a list of intermediate values is obtained. Th reduce function then takes the output of the ma functions and combines the values. Jobtracke divides the map tasks amongst the tasktrackers. |
| 11 | Big Data and Hadoop | Simmi Bagga and Satinder Kaur | Validity and volatility are features of big data Hadoop provides storage and processing in distributed manner. HDFS uses commodity hardware which is cheap. |
| 12 | A Study of Hadoop-Related Tools and Techniques | Deepika Pand Anantha Raman G R | Hadoop has many tools which help it in dealin with data. They include HBase, HDFS, Hive an MapReduce. |
| 13 | Sentiment Analysis on Twitter Data using Hadoop | Abhinandan P Shirahatti, Neha Patil, Durgappa Kubasad and Arif Mujawar | Big data can be used in sentiment analysis as contains various patterns and is generated fron social networks also. |
| 14 | Web News Mining using Back Propagation Neural Network and Clustering using K-Means Algorithm in Big Data. | Sukhpal Kaur, Mamoon Rashid | Managing Web News Big Data and categorizing it on the basis of text and content to provide the accurate news with less running time. |

**REFERENCES**

[1] Seref Sagiroglu and Duygu Sinanc, "Big Data: A review" , IEEE, 2013, pp. 42–47.
[2] Dinesh D. Jagtap and Prof. B.K. Patil, "Big Data using Hadoop", International Journal of Engineering Research and General Science, Oct-Nov 2014, Volume 2, Issue 6, pp. 467 472.
[3] Cloudera white paper, "Hadoop and HDFS", version 106.
[4] Sreedhar C, Dr. D. Kavitha and K. Asha Rani, "Big Data and Hadoop", International Journal of Advanced Research in Computer Engineering and Technology, May 2014, Volume 3, Issue 5, pp. 1571–1575.
[5] Yashika Verma and Sumit Hooda, "A Review Paper on Big Data and Hadoop", International Journal of Scientific Research and Development, 2015, Volume 3, Issue 2, pp. 682–684.
[6] Swarupkumar Shivaji Suradkar, "Analysis of Big Data with Hadoop Framework", International Journal of Advanced Research in Computer Science and Software Engineering, Nov 2016, Volume 6, Issue 11, pp. 167–169.
[7]Anusha B. Dhakite and Prof. Nitin V. Wankhade, "Big Data: An Overview", International Journal of Advanced Research in Computer Science and Software Engineering, June 2016, Volume 6, Issue 6, pp. 841–845.
[8] Prabhjot singh, Mamoon Rashid, "Smart Home Automation Deployment on Third Party Cloud Using Internet of Things", published in Journal of Bioinformatics and Intelligent Control, Vol. 4, 1–4, 2015.
[9] Ranjana Bahri, "Big Data: Concept, Challenges and Management Tools", International Journal of Advanced Research in Computer Science and Software Engineering, Feb 2015, Volume 5, Issue 2, pp. 986–990.
[10] Prof. R.A. Fadnavis and Samrudhi Tabhane, "Big Data Processing using Hadoop", International Journal of Computer Science and Information Technologies, 2015, Volume 6(1), pp. 443–445.
[11] Simmi Bagga and Satinder Kaur, "Big Data and Hadoop", International Journal of Advanced Engineering Research and Applications, Oct 2015, Volume 1, Issue 6, pp. 233–238.
[12] Deepika Pand Anantha Raman G R, "A Study of Hadoop-Related Tools and Techniques", International Journal of Advanced Research in Computer Science and Software Engineering, September 2015, Volume 5, Issue 9, pp. 160–164.
[13] Abhinandan P Shirahatti, Neha Patil, Durgappa Kubasad and Arif Mujawar, "Sentiment Analysis on Twitter Data using Hadoop", International Journal of Emerging Technology in Computer Science and Electronics, April 2015, Volume 14, Issue 2, pp. 831–837.
[14] Sukhpal Kaur and Mamoon Rashid, "Web News Mining using Back Propagation Neural Network and Clustering using K-Means Algorithm in Big Data, published in Indian Journal of Science and Technology, Vol 9(41), DOI: 10.17485/ijst/2016/v9i41/95598, November 2016.


## Authors Profile


*Mr. Mamoon Rashid* is currently working as an Assistant Professor in the School of Computer Science & Engineering, Lovely Professional University, Jalandhar, India. The main area of interest is Cloud Computing, Big Data, Data Sciences, Machine Learning and Information Security. The author has published several research papers related to Cloud Computing and Big Data in Scopus, Web of Science and Thomson Reuters based Journals.

*Mr Aamir Hamid is* Research Scholar in the School of Computer Science & Engineering, Lovely Professional University, Jalandhar, India. The author has main research interests in Big Data and Machine Learning Technologies.

*Mr Balwant Ram* is Currently working as an Assistant Professor in the School of Computer Science & Engineering, Lovely Professional University, Jalandhar, India. The main area of interest is Pattern Recognition, Machine Learning and Web Services. The author has published several research papers on Machine Learning and Web Services in Scopus and Thomson Reuters based Journals.