A Review Study of Big Data and Cloud Computing

Hina Zafar
Raja Balwant Singh Engineering Technical Campus, Bichpuri Agra

Brajesh Kumar Singh
Raja Balwant Singh Engineering Technical Campus, Bichpuri Agra

ABSTRACT
Big data and Cloud Computing are the latest technologies that are high in demand. Big data refers to the massive amount of data generated by various internet sources such as retail applications, banking services, social media, stock exchanges, airlines, etc. A huge amount of data is produced which cannot be handled by traditional database software like RDBMS. Cloud computing is a game changer in the way we store, process, analyze data and access it. It provisions infrastructure, hardware, and software on the internet as per user demand. This paper has reviewed various research papers. What is big data? Its meaning with respect to different authors, characteristics, deployment on cloud, its various models, scope, and challenges are discussed.

Keywords
Big data, Cloud Computing, Data Analytics

1. INTRODUCTION
Big data is defined as a massive amount of data generated from a vast variety of internet sources. It is characterized by 3V's – volume (size of data), velocity (data production and processing speed), and variety (types of data). In the present scenario, the internet is accessible to almost every human hence data generated per day is from terabyte to zettabyte. In research, it is found that Facebook generates 500+ Tb of data every day. Another research says New York Stock Exchange produces about one TB of new trade data per day. This size of data is large and in various formats such as images, video, text, gifs, etc. It is categorized as structured, unstructured, and semi-structured data. Traditional database software works only on structured data; hence the need arises to develop new software that handles the other two forms of data. Cloud computing is the best technology to address this issue. It offers a secure, reliable, and scalable service to store and process big data. Computing services such as infrastructure setup (IaaS), hardware installation (PaaS), and tool deployment (SaaS) are done on remote machines which are accessed through the internet. The setup is done with a low upfront cost. Cloud Computing offers a solution for small to large-scale businesses. There are three basic cloud models Private, Public and Hybrid Cloud. Each organization can choose from these according to their requirement. However, it's seen that most organizations prefer Hybrid Cloud as it's a combination of both. In this paper, we will the review work of various authors, and analyze and compare them. We will discuss tools used to store and process big data, advantages and challenges faced. Figure 1 shows the processing of big data through various software using the cloud.

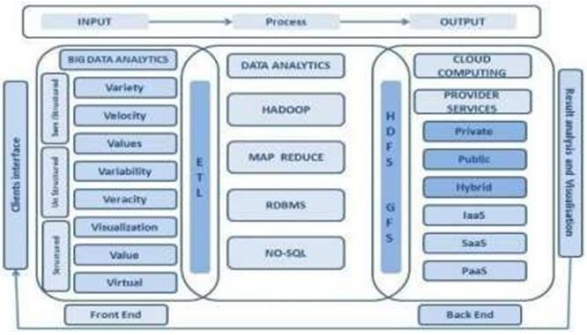

Fig 1: Relationship Model- Big data and Cloud Computing [1]

2. RELATED WORK

2.1 Mohaiminul Islam, Shamim Reza[3] – The Rise of Big data and Cloud Computing

- The above paper "The Rise of Big Data and Cloud Computing" has explained in detail the definition, characteristics, and types of Big Data and Cloud Computing and the relation between them. As said big data is a huge volume of data that cannot be processed by traditional software. Cloud computing offers the capability to store and process data efficiently. Big data was first coined in 2010.[4][5]
- It refers to a large volume of data generated from the Internet. It can be of different types Structured, Unstructured and Semi-Structured. Structured data is one of the types of big data that can be stored, processed, and fetched in a fixed format. The Unstructured data doesn't have any format like email. Semi-structured data is a combination of structured and unstructured data e.g., a Google Search can result in text, video, gif, etc., hence it's difficult to store, process, and analyze using a traditional database system like RDBMS.
- Cloud computing is known as the "Democratization of Information System"[6]. It provides infrastructure, software, and platform as a service. One can build up complete infrastructure on a remote machine with no upfront cost, a pay-as-you-go model, and high availability, scalability, security, and cost-efficient system. For example, Amazon's "Elastic Map Reduce" exhibits the use of EC2 in Big Data processing. Hadoop is an open-source framework used to effectively store and process large datasets ranging in size from gigabytes to petabytes of data. Instead of using one large computer to store and process the data, Hadoop allows clusters of multiple computers. IBM Interconnect 2013 discussed how a hybrid cloud can be used in the analytics of unstructured data. Service providers like AWS, Google, and Microsoft are offering their own big data software in a cost-efficient manner that can be used for businesses of all sizes.

- Cloud offers three types of architecture Public, Private and hybrid to cater needs of all types of businesses.
- A new model Analytics as a Service (AaaS) has been developed to serve the need for big data. This model will work on all types of data and process them faster than previous models.
- The Relationship between Big Data and Cloud Computing: -
    - Cloud Computing provides an infrastructure to save processes and store big data.
    - Since the cloud environment is scalable, it can provide an efficient data management solution if volumeincreases rapidly.
    - Cloud computing offers security features to protect the company's confidential data.
    - Cloud vendors are responsible for the global management of servers hence the overall cost is less compared to local servers. This helps to process data from different locations.
    - Cloud computing has high-level network servers that process data faster.
- In addition, this cloud provides a low-cost simple virtualized environment with security and privacy.
- Cloud can fulfill the sudden demand to handle the exponential spike in data, but provisioning resources for a short period of time can increase the overall project cost.

2.2 Charlotte Castelino [7], Dhaval Gandhi [7], Harish G. Narula [8], Nirav H. C - Integration of big data and Cloud Computing

- In this article big data is defined as a new phenomenon that was introduced because of large data generated from various sources. It discusses in brief about 3 Vs of big data i.e., Volume, Variety, and Velocity. Cloud computing, its services, and architecture are discussed in short. The main emphasis of this paper is on the Integration of Big data and the Cloud Environment and its potential benefits.
- The author suggests the use a of hybrid cloud for big data storage and processing. ConPaaS is a cloud environment for big data processing, which was developed by Vrije Universiteit Amsterdam, the University of Rennes 1, Zuse-Institut Berlin, and XLAB. It facilitates cloud deployment efficiently and supports web hosting, SQL and NoSQL databases, and data storage. MapReduce and TaskFarming are two services specifically dedicated to Big Data. MapReduce provides parallel programming facilities while TaskFarming provides automatic execution of independent units.
- Cloud Computing offers potential benefits such as a parallel scalable secure computing solution. However, it possesses major challenges such as: -
    - Transfer of data from on-premises to the cloud introduces latency.
    - Data retention is governed by multiple laws, which is a major challenge
    - Cloud doesn't offer 100% isolation management guarantee
    - Data recovery is difficult at times of disaster

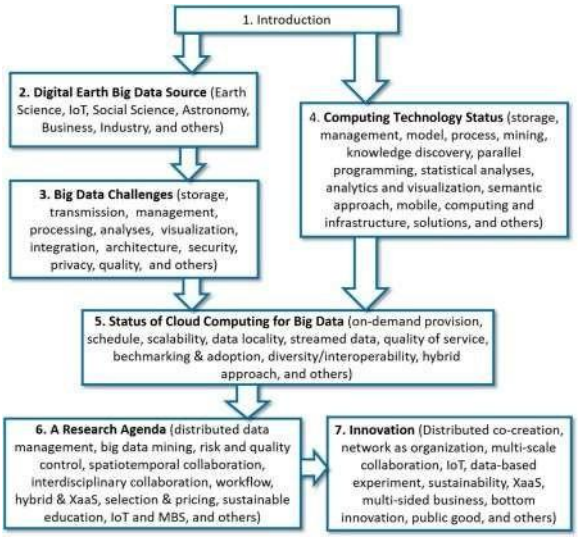

Fig 2: Tackling Big Data challenges with cloud computing for innovation [7]

2.3 Divyakant Agrawal [9] Sudipto Das [9] Amr El Abbadi [9] - Big Data and Cloud Computing: New Wine or just New Bottles?

- This paper summarizes the statements of main points with no elaboration or explanation. It starts with the cloud and its architecture and introduces the concept of DaaS (Database as a service).
- However major emphasis is laid on the design of databases for single and large multitenant databases, small and heavy workloads. Since RDBMS cannot fulfill the need for fast-growing data, new technologies have been used such as Key-Value Stores in Bigtable, PNUTS, and Dynamo.
- The major challenge found is designing a perfect secure cloud infrastructure for a multitenant database system. No such system has been designed yet.

2.4 Bernice M. Purcell Holy Family University - Big Data using Cloud Computing.

- In this paper big data is defined as data analysis methodology. Major emphasis is laid on big data infrastructure based on a cluster known as NAS (Network Attached Storage) and analysis is done through MapReduce.
- Cloud offers inexpensive scalable secure services for all kinds of business. Cloud offers various platforms such as Infrastructure as a Service (IaaS), Software as a Service (SaaS), Platform as a Service (PaaS), and Hardware as a Service (HaaS). Based on business requirement Private, Public and Hybrid Cloud can be implemented. One of the challenges for public cloud computing are security and loss of control.

2.5 Samir A. El-Seoud [10] Hosam F. El-Sofany [10] Reham Mohamed British University [10] - Big Data and Cloud Computing: Trends and Challenges.

- This paper elaborates on big data, its characteristics 5V's (volume, velocity, variety, value, and veracity); Cloud computing concepts, models, and architecture. It

also defines the 6 stages of how big data analytics is deployed on the Cloud.

- The paper suggests the storage options for big data can be Google File System, Amazon (S3), Nirvanix, OpenStack and Windows Azure (Blob) storage. Hadoop, Hadoop Distributed File System (HDFS), MapReduce, NAS, and Taskfarming are preferred for processing. There are a few cloud environments for big data analytics like Canpaas that which developed by Vrije Universiteit Amsterdam, the University of Rennes 1, Zuse-Institut Berlin and XLAB. The environment hosts web applications written in PHP or java, as well as manages different dbms including SQL or NoSQL.
- Despite multiple benefits big data on cloud computing offers, there is certain risk and challenges. One of them is security while other is data and its availability. Big data may be available in any part of the world but there is no surety that cloud server and its application support will be present too.

2.6 Logica Bănică [11], Viorel Păun[11], Cristian Ștefan[11] - Big Data leverages Cloud Computing Opportunities.

- This paper has well explained the concepts of Big Data and Cloud computing with references, facts and examples.
- Cloud computing offers various models, other than that mentioned before like Business Desktop as a Service (DaaS), Storage as a Service (StaaS), Security as a Service (SECaaS), Process as a Service (BPaaS), and architecture such as Community Cloud.
- The term Big Data was used in 1997, by two NASA researchers, Michael Cox and David Ellsworth [12], to describe massive amounts of information that cannot be processed and visualized. But this is not true today. The author favors Oracle 4V's Volume, Velocity, Variety and Value and a C for Complexity.
- NoSQL is used for storing big data and Hadoop is used for processing it. The workflow process has been explained in detail with diagram.
- Besides all the advantages, big data and Cloud Computation integration is very expensive.

2.7 Md. Golam Morshed [13] Ling Yuan [13] – Big Data in Cloud Computing: An Analysis of Issues and Challenges.

- This paper defines big data, cloud, their characteristics, storage, and processing software.
- However, challenges are explained in detail since data is coming from a variety of sources its heterogeneous in nature.
- While the other challenges are Availability, Scalability, Privacy, Integrity, Transformation, and Administration of data.
- Still there is a large scope for research in this area such as Higher-Level Computing systems, Distributed Database systems, Data Analytics, Quality, and Security.

2.8 Chamandeep Kaur[14], Fatima Farhan[14] , Ali Almaliki[14] , Wejdan Mohommed Therwi[14], Alaa Mohommed Alhassan Tomihe[14], Samar Mansour Hassen[14] - The Study of Resource Management in Big Data Using Cloud Computing.

- This paper presents a resource management study with respect to big data and cloud computing.
- Big data is a huge amount of data and cloud computing provides a platform to store and process it. Cloud computing provides an internet-based platform where servers and applications are provisioned on remote machines provided at minimum or no upfront cost.
- Since data grows exponentially multiple servers are used.
- Hadoop is preferred for processing big data.
- Cloud computing offers a more scalable, flexible, cost-effective model than an on-premises deployment.

2.9 P.M. Kokila[15], P. Saravanan[15], Dr. B. Jagadeesan[15], Sharmila[15] – Big Data and Cloud Computing Service Models and NoSQL Deployment.

- National Institute of Standards and Technology (NIST), defines cloud computing as "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" [16].
- Cloud Computing characteristics are defined as On-Demand Self Service, Broad Network Access, Resource Pooling, Rapid Elasticity, Measured Service.
- There are 4 deployment models are: Public Cloud, Private Cloud, Community Cloud, and Hybrid Cloud.
- NoSQL database is used for big data storage.
- Challenges are interoperability, transferring data to the cloud, and portability.

2.10 Mohammad Tabrez Quasim [17] , Prashant Johri[17] , Mohammad Meraj[17] , SK.Wasim Haider[17] - 5 V'SOF BIG DATA VIA CLOUD COMPUTING: USES AND IMPORTANCE

- This paper has defined big data and explained its 5V's (Volume, Velocity, Variety, Value, Veracity).
- There are various big data applications such as Healthcare, Education and Government, Banking and Securities, Communications, Media, and Entertainment.
- Cloud vendors offer a different solution for big data storage and processing as shown below:

Table 1: Comparison of various big data providers [17]

	Google	Microsoft	Amazon	Cloudera
Big Data Storage	Google Cloud Services	Azure	S3	
Map Reduce	AppEngine	Hadoop on Azure	Elastic Map Reduce (Hadoop)	Map Reduce YARN
Big Data Analytics	Big Query	Hadoop on Azure	Elastic Map Reduce (Hadoop)	Elastic Map Reduce (Hadoop)
Relational Data Base	Cloud SQL	SQL Azure	My SQL or Oracle	My SQL,Oracle, PostgreSQL
NoSQL Database	AppengineDataStore	TableStorage	DynamoDB	Apache Accumulo
Streaming Processing	Search API	Streaminsight	Nothing Prepackaged	Apache Spark
Machine Learning	Prediction API	Hadoop+ Mahout	Hadoop+ Mahout	Hadoop+ Oryx
Data import	Network	Network	Network	Network

COMPARATIVE ANALYSIS OF PREVIOUS RESEARCH PAPER

After reviewing previous research papers, I have concluded that most scholars use IAAS model for Big Data followed by PAAS & SAAS and only few recommend other types of cloud models.

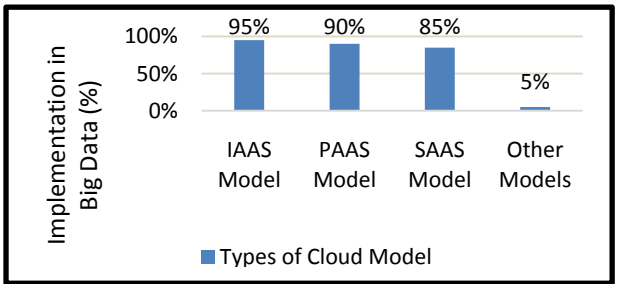

Fig 3: Comparative analysis of previous research paper.

3. CONCLUSION

In this paper we have reviewed the work of different authors, understanding their perspectives on big data and cloud computing. Big data and cloud computing have various definitions according to the use and understanding of authors. As we have seen Cloud offers various models, but the selection is based on requirements. Big data and Cloud Integration is a must for today. However, there are multiple ways to accomplish it. One can use Amazon S3, GFS, Nirvanix Cloud Storage for storage while NoSQL, Hadoop, TaskFarming for processing and analysis. Public Cloud, Private Cloud, or Hybrid Cloud can be set up depending upon the privacy of data and type of access given to users. Cloud offers multiple benefits – Allows pay as you use, On-Demand service, Multi-tenancy, Virtualization, Location, Device Independence, Security, Scales automatically to adjust to the increase in demand, etc. Nevertheless, challenges never end– the same internet connectivity cannot be achieved in all parts of the world, cloud vendor's downtime should be considered while project implementation and execution, before adopting cloud technology, you should be aware of the fact that you are sharing company's private information to a third-party cloud computing service provider. Hackers may misuse this information.

4. REFERENCES

[1] Mohaiminul Islam, Shamim Reza.,2019, The Rise of Big Data and Cloud Computing.

[2] C. L. Philip, Q. Chen and C. Y. Zhang, 2014, A survey on big data, Information Sciences.

[3] K. Kambatla, G. Kollias, V. Kumar and A. Gram, 2018, Trends in big data analytics.

[4] Charlotte Castelino, Dhaval Gandhi, Nirav H. Chokshi,2014, Integration of Big Data and Cloud Computing.

[5] Harish G. Narula.2017, Integration of Big Data and Cloud.

[6] Divyakant Agrawal, Sudipto Das and Amr El Abbadi,2010, Big Data and Cloud Computing: New Wine or just New Bottles?

[7] Samir A. El-Seoud, Hosam F. El-Sofany and Reham,2017, Big Data and Cloud Computing: Trends and Challenges.

[8] Logica Bănică, Viorel Păun, Cristian Ștefan - Big Data leverages,2014, Cloud Computing opportunities.

[9] Cox, M., Ellsworth, D., 1997. Application-Controlled Demand Paging for Out-of-Core Visualization.

[10] Md. Golam Morshed and Ling Yuan, 2017,- Big Data in Cloud Computing: An Analysis of Issues and Challenges.

[11] Chamandeep Kaur, Fatima Farhan, Ali Almaliki , Wejdan Mohommed Therwi, Alaa Mohommed Alhassan Tomihe, Samar Mansour Hassen,2019,The Study of Resource Management in Big Data Using Cloud Computing.

[12] P.M. Kokila, P. Saravanan, Dr.B.Jagadeesan, Sharmila,2016, - Big Data and Cloud Computing Service Models and Nosql Deployment.

[13] NIST the NIST Definition of Cloud Computing,2011,http://csrc.nist.gov/publications/nistpubs/8 .

[14] Mohammad tabrez quasim , prashant johri , mohammad meraj , sk.wasim haider, 2011,5 v's of Big Data via Cloud Computing: Uses and Importance.