iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** IV **Month of publication:** April 2023

DOI: https://doi.org/10.22214/ijraset.2023.49619

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue IV Apr 2023- Available at www.ijraset.com

Big Data in Cloud Computing: An Overview

Shiva Sharma[1], Simranjeet Singh[2], Shyam Sharma[3], Neerja Negi[4]
Department of Computer Applications, Manav Rachna International Institute of Research & Studies, Haryana, India

Abstract: Cloud computing is a potent tool for sophisticated and massive-scale computation. It removes the need for expensive hardware, specialized space, and software maintenance. It has been noticed that cloud computing has resulted in a massive increase in the volume of data, or big data. Managing massive amounts of data is a complex and time-consuming operation that requires an extensive computer infrastructure for effective data processing and analysis. Many industries, including minor and major organizations, healthcare, education, and many more, are attempting to harness the potential of big data. In healthcare, for example, big data is used to reduce treatment costs, predict pandemic outbreaks, and prevent infections, among other things. This article discusses comprehensive data processing strategies from system and application perspectives to offer an orderly picture of the issues that application developers and database management system (DBMS) designers face while designing and deploying internet-scale applications. While big data has various uses in various industries, it has challenges.
Keywords: Big Data, Cloud Computing, DBMS, Data Processing.

I. INTRODUCTION

Cloud computing has shown to be a practical paradigm for SOAP. This development has ushered in changes in the abstraction and use of computer infrastructure. The flexibility, pay-as-you-go pricing model, cheap initial investment, and risk transferability of cloud computing make it the go-to platform for establishing cost-effective business infrastructure. For several decades, distributed databases have been the holy grail of scientific inquiry. However, as data patterns and applications evolve, a new form known as key-value storage has emerged and is now extensively employed by many businesses. Hadoop, an open-source version of MapReduce, is widely utilized in business and academia [1]. In terms of usability and efficiency, Hadoop is a game-changer. HDFS has become a beneficial technology for managing and archiving large, complicated datasets. It is becoming easier for computers to access and make sense of big data. Today is a data-driven world. They are everywhere these days due to the fantastic technological advances of recent years [2]. The pace of digitization has accelerated, and the term "digital information societies" has entered common parlance. Whereas just 1% of information created 20 or 30 years ago was digital, now more than 94% of information arrives in digital form from a wide variety of digital sources. Large data sets that exceed the capacity of existing technologies are a hallmark of the "big data" phenomenon, which represents the evolution of human cognition [3]. Fast, heterogeneous data calls for novel processing forms to facilitate decision-making, insight discovery, and process optimization. We must be able to safely store, handle, and share complex data on the cloud so that we can analyse the data and identify trends. Given the cloud's inherent complexity, we believe that focusing on incremental improvements to cloud security is preferable to presenting comprehensive approaches.

II. BIG DATA

Big data refers to the enormous, intricate, and varied databases that are challenging to handle and process using conventional data processing techniques. Volume, Velocity, and Variety are the three Vs that define it. The enormous quantity of data produced by numerous sources, including social media, sensors, and other digital devices, is referred to as volume. Velocity is the measure of how quickly data must be handled in order to be used in real-time. Data that is diverse includes all kinds and forms, including organized, semi-structured, and unstructured data. The difficulties of handling and studying these sizable files have given rise to big data technologies like Hadoop, Spark, and NoSQL databases. [4]. These tools enable businesses to gather insightful data and make data-driven choices in a variety of industries, including marketing, finance, and healthcare.

A. Big Data and its features

Volume, value, variety, velocity, and veracity often define big data as a compilation of several sources.

1) *Volume:* The size and scope of a company's big data operations.
2) *Value:* From a commercial perspective, the most important "V" is value, and the value of big data is created when new insights and patterns are uncovered, which in turn lead to increased productivity, stronger customer relationships, and other tangible benefits [5].

3) *Variety:* Raw data, semi-structured data, and unprocessed data all contribute to the vastness and variety of available information.
4) *Velocity:* Speed with which information is gathered, stored, and processed by an organisation; for example, the number of social media postings or search queries received daily, hourly, or in any other time period.
5) *Veracity:* Executive confidence is typically influenced by the "truth" or accuracy of data and information assets.

Processing in Velocity may be done in two main ways: in a batch, or in a continuous stream. It is common practice to process data in batches that have been saved for later use. Data handled in batches tend to be quite useful. As a result, their processing time will increase. For large amounts of data, Hadoop MapReduce is the best framework available. This technique works well when processing large volumes of data is more important than obtaining real-time analytics.

However, stream processing is fundamental for real-time data processing and analysis. With the use of stream processing, new information may be examined as it comes. Rapid ingestion of this data into analytics tools enables rapid output of findings. The ability to spot anomalies that point to fraud in real time makes this approach promising in a number of contexts. Furthermore, online firms would profit from real-time processing since it would enable them to keep detailed records of consumer transactions and provide real-time product recommendations [6].

III. CLOUD MANAGEMENT FOR MASSIVE DATA SETS

The Cloud Computing ecosystem is built on the use and provision of services. There are several groups into which service-oriented systems might be grouped. The abstraction level supplied to the system's user is one of the most common criteria for categorising these systems. Typically, three distinct tiers are separated in this manner: Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS) (SaaS). Cloud Computing provides scalability regarding resource utilisation, cheap administration effort, price model flexibility, and software user mobility. Under these conditions, it is clear that the Cloud Computing paradigm is advantageous for big projects, such as those involving Big Data and BI [7].

Considering the nature of the data management industry, the optimal management organisation design may be built on a four-layer architecture and include the following elements:

A file system for storing Big Data, i.e., many big-sized archives. This layer is implemented at the IaaS level since it specifies the fundamental architecture structure for the subsequent layers [8].

A DBMS for efficiently arranging and gaining access to data. It is situated between IaaS and PaaS since it has properties with both systems. Developers utilise it to access the data, although its implementation is hardware-based. A PaaS serves as an interface, offering its capabilities on the top side and the implementation for a specific IaaS on the lower side. This functionality enables the deployment of apps on several IaaS without rewriting them.

A tool for distributing the computing workload among the cloud's processors. Clearly connected to PaaS, this layer functions as a "software API" for encoding Big Data and BI applications [9].

Users need a query mechanism for knowledge and information extraction between the PaaS and SaaS levels.

Computing services like as hosts, memory, databases, infrastructure, applications, analytics, and many more are distributed across the Internet to provide scalability, rapid innovation, and cost savings. Cloud computing has transformed the abstraction and use of computer infrastructure. The scope of cloud concepts has been expanded to include anything that may be deemed a service. The many advantages of cloud computing, including flexibility, pay-as-you-go or pay-per-use models, cheap initial investment, and many more, have made it a feasible and desired option for storing, administering, and analytics of large amounts of data [10]. Amazon, Google, and Microsoft provide their own cost-effective big data platforms since big data is increasingly crucial for many enterprises and disciplines. These technologies are scalable for organisations of all sizes. That has led to the popularity of Analytics as a Service (AaaS) as a quicker and more effective method to connect, manipulate, and display various kinds of data. Data Analytics [11].

IV. BIG DATA ANALYTICS CYCLE

According to experts, processing massive data for analytics varies from regular transactional data. In conventional setups, data is analyzed before creating a model design and database structure. As can be seen, it begins by collecting information from several sources, including different files, systems, sensors, and the Internet. This data is stored on a medium capable of processing the volume, diversity, and velocity of data, known as the "landing zone." Typically, this is a distributed file system. After data is saved, it undergoes many modifications to retain its efficiency and scalability. Then they are incorporated into specific analytic activities, operational reporting, databases, or raw data extraction [12].

A. Advantages of Big Data Analytics

For companies seeking to harness the power of data to drive business outcomes, big data analytics has become a crucial instrument. The following are some benefits of big data analytics:

Decision-making is improved thanks to big data analytics, which give businesses insights into consumer behavior, market patterns, and other important data elements. Organizations can find patterns and trends that would be difficult to find through manual analysis by studying big datasets.

1) *Saving Money:* Through data analysis, businesses can pinpoint areas where they can reduce expenses, reorganize their processes, and make better use of their resources. Big data analytics, for instance, can assist businesses in cutting waste, enhancing supply chain effectiveness, and lowering delay [13].
2) *Enhanced Effectiveness:* Big data analytics can assist companies in automating routine processes and improving the effectiveness of their operations. Machine learning algorithms, for instance, can be used to automate monotonous chores and increase output. Organizations can open up resources to concentrate on more important projects by automating procedures [14].
3) *Improved Customer Experience:* By analyzing customer data, businesses can better comprehend customers' requirements and preferences and adjust the content and delivery of their goods and services. Big data analytics, for instance, can be used to customize marketing campaigns, enhance client support, and find new product possibilities.
4) *Better Risk Management:* Big data analytics can support businesses in identifying possible hazards and mitigating them before they develop into significant problems. Predictive analytics, for instance, can be used to spot theft or cybersecurity risks before they cause serious harm [15].
5) *Competitive Advantage:* By using big data analytics, businesses can make quicker, data-driven choices that give them a competitive edge. Big data analytics, for instance, can be used to spot market patterns and openings, giving businesses an advantage over rivals.

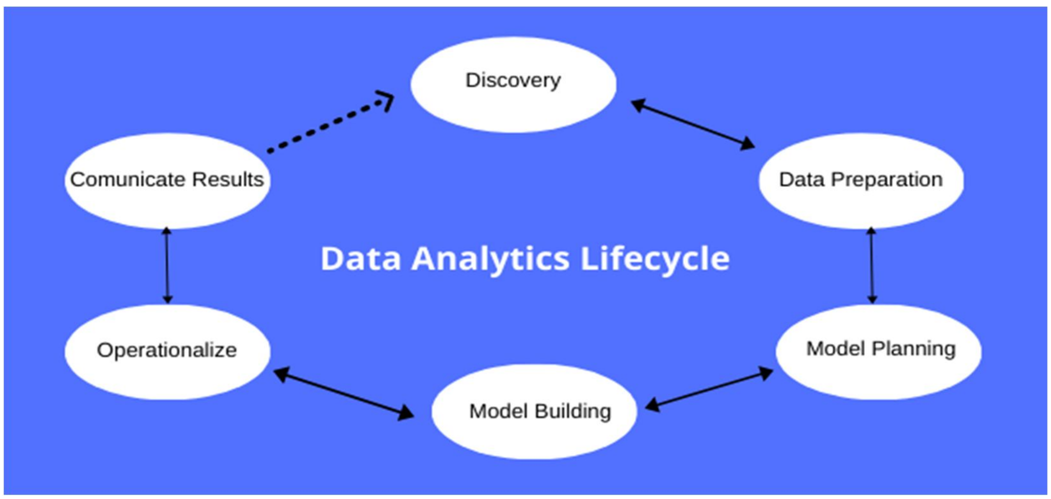


Figure 1. Big Data Analytics Lifecycle.

V. BIG DATAT MANAGEMENT

The demands of big data cannot be met by present technology, and the rate of storage capacity expansion is substantially slower than the data growth rate. Consequently, a revolutionary redesign of the information framework is essential. For this, we must develop a hierarchical storage architecture. Existing efficient algorithms do not effectively manage heterogeneous data; thus, it is necessary to build a highly efficient algorithm to manage heterogeneous data effectively [16].

A. Security in Big Data is Essential

Many businesses use big data, yet they may need more security-related assets. If there is a security danger to big data, it may result in an even more significant problem. Companies utilise this technology to store petabyte-scale data on the firm, its business, and its customers. That has a significant impact on the categorisation of information. We must either encrypt it, log it, or use honeypot tactics to safeguard the data. The difficulty of identifying threats and malicious intruders must be resolved through big data analysis techniques [17].

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue IV Apr 2023- Available at www.ijraset.com

B. Extensive data Analysis and Computation

Speed is the most crucial factor when searching large datasets. However, the procedure may be time-consuming because it needs to explore all linked entries in the database quickly. While big data is becoming more complex, the indexes within big data target the most specific data types. The conventional serial technique could be more efficient for such a large data set [18].

VI. RISK AND CHALLENGES

Big data and cloud processing have many advantages, but they also have their share of dangers and difficulties. The following are some dangers and difficulties associated with large data in cloud computing:

1) *Data Security:* Data security is one of the major dangers associated with large data and cloud computing. Sensitive information is susceptible to hacking, data leaks, and cyber-attacks when it is kept in the cloud. To secure their data, organizations must make sure that the proper security measures are in place. Examples include encryption and multi-factor identification [19].
2) *Data Privacy:* When it comes to private and personal data, cloud computing can also be a danger to data privacy. Organizations must make sure they abide by data protection laws like the GDPR, CCPA, and HIPAA to prevent fines and other consequences.
3) *Data Governance:* Managing big databases can be difficult, and if it isn't done correctly, it can result in data mistakes, discrepancies, and faults. To guarantee that data is managed successfully, organizations must create clear data governance policies, methods, and protocols.
4) *Data Integration:* When working with big databases, integrating data from various sources can be difficult. To successfully combine data, organizations must make sure they have the appropriate platforms and tools in place [20].
5) *Scalability:* Businesses need to make sure their cloud systems can expand as needed to handle the growing amount of data. Failure to comply with this can result in subpar efficiency and system breakdowns.
6) *Provider lock-in:* Businesses that significantly rely on cloud services risk becoming reliant on just one provider, which results in vendor lock-in. Organizations may find it challenging to change cloud suppliers or sellers as a result.

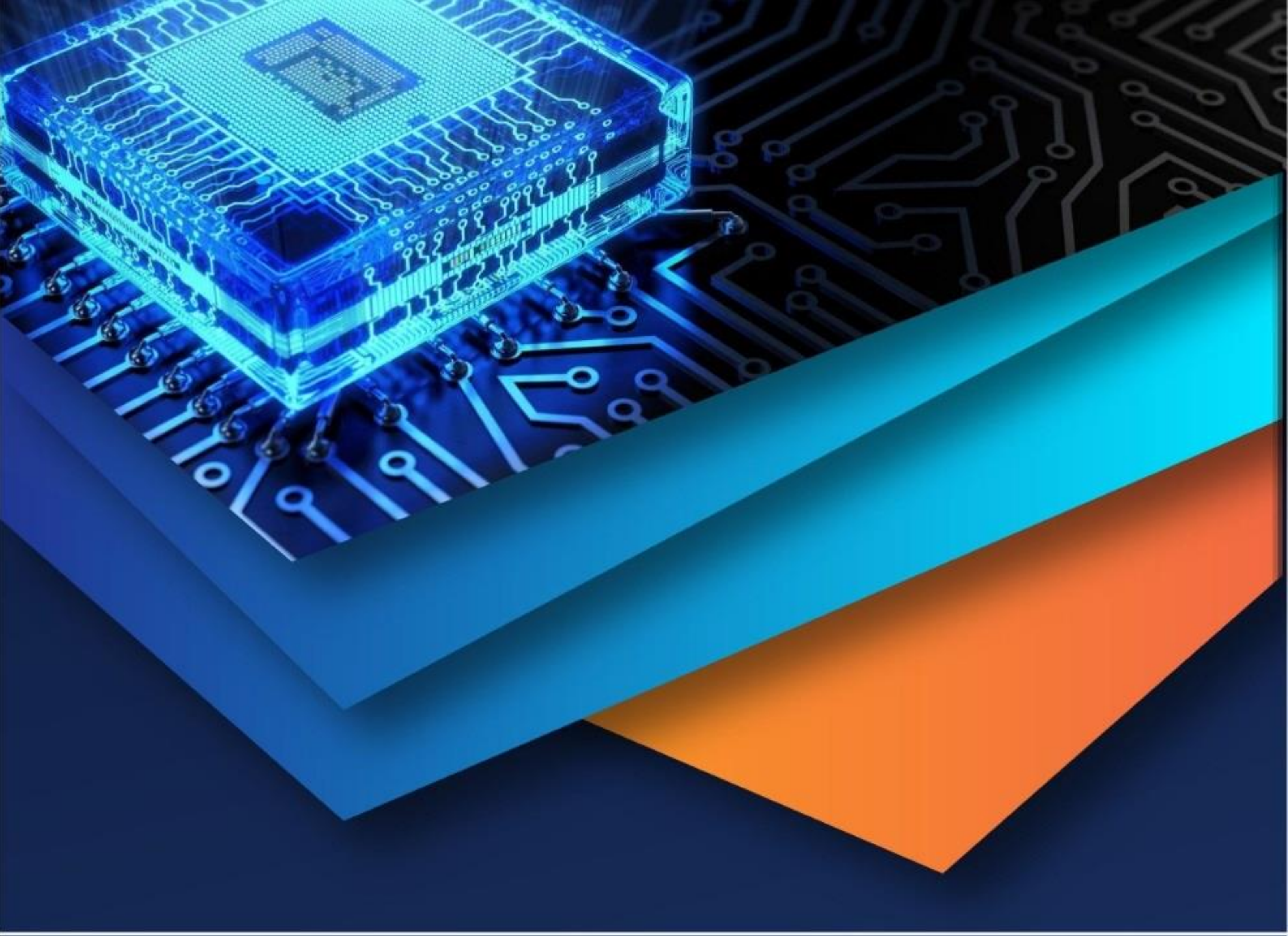
VII. CONCLUSION

Big Data is not a new concept, but it has recently come to the forefront due to the daily production of vast quantities of data from many sources. Our investigation revealed that big data is expanding rapidly, resulting in both advantages and concerns. Cloud computing is the ideal method for storing, processing, and analysing Big Data. The capacity to store vast volumes of data in a variety of formats and to analyse it at very high rates will provide data that can assist companies and educational institutions in their rapid development. The article provided an overview of Big Data and Cloud Computing, including its basic concepts and terminology, as well as the evolution of data management into cloud computing. As a bonus, it investigates the upsides and downsides of combining big data with cloud computing. Data storage and processing power are significant perks of cloud computing and extensive data integration; the cloud has access to a vast pool of resources and a variety of infrastructures that can accommodate this integration in the most suitable manner possible. The environment can be set up and managed with minimal effort to provide an excellent workspace for all extensive data requirements.

REFERENCES

[1] Neelay Jagani, Parthil Jagani, Suril Shah et al (2021) Big Data in Cloud Computing: A Literature Review. International Journal of Engineering Applied Sciences & Technology 5(11):185-191
[2] Samir A. El-Seoud, Hosam F. El-Sofany, Mohamed Abdelfattah, Reham Mohamed et al (2017) Big Data and Cloud Computing: Trends and Challenges. International Journal of Interactive Mobile Technologies 11(2):34
[3] Amanpreet Kaur Sandhu (2021) Big Data with Cloud Computing: Discussion and Challeneges. Big Data Mining and Analytics 5(1):32-40
[4] Venkatesh H, Shrivatsa D Perur, Nivedita Jalihal et al (2015) A Study On Use of Big Data in Cloud Computing Environment. International Journal of Computer Science and Information Technologies 6(3):2076-2078
[5] Pedro Caldeira Neves, Bradley Schmerl, Jorge Bernardino, Javier Camara et al (2016) Big Data in Cloud Computing: Features and Issues. International Conference on Internet of Things and Big Data 307-314
[6] T. Sri Harsha (2017) Big Data Analytics in Cloud Computing Environment. International Journal of Scientific & Engineering Research 8(8):393-398
[7] Ibrahim Abaker Targio Hashem, Ibrar Yaqoob, Nor Badrul Anuar, Salimah Mokhtar, Abdullah Gani, Samee Ullah khan et al (2015) The rise of “big data” on cloud computing: Review and open research issues. Information Systems 47:98-115
[8] Subia Saif, Samar Wazir (2018) Performance Analysis of Big Data and Cloud Computing Techniques: A Survey. International Conference on Computational Intelligence and Data Science 132:118-127
[9] Shahana PN (2022) Impact and Implications of Big Data Analytics in Cloud Computing Platforms. International Journal for Research in Applied Science and Engineering Technology 10(5)

[10] Md. Golam Morshed, Ling Yuan (2017) Big Data in Cloud Computing: An Analysis of Issues and Challenges. International Journal of Advanced Studies in Computer Science and Engineering 6(4):7-11
[11] Hassan Sohail, Zeenia Zameer, Hafiz Farhan Ahmed, Usama Iqbal, Pir Amad Ali Shah et al (2017) Challenges and Opportunities in Big Data and Cloud Computing. ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 175-181
[12] Chaowei Yang, Qunying Huang, Zhenlong Li, Kai Liu, Fei Hu et al (2017) Big Data and Cloud Computing: Innovation Opportunities and Challenges. International Journal of Digital Earth 10(1):13-53
[13] Venkata Narasimha Inukollu, Sailaja Arsi, Srinivasa Rao Ravuri et al (2014) Security Issues Associated with Big Data in Cloud Computing. International Journal of Network Security & its Applications 6(3):45-56
[14] Jinsong Zhang (2018) Applications and Challenges of Big Data and Cloud Computing in Power Industry. International Symposium on Communication Engineering & Computer Science 86:119-122
[15] Manoj Muniswamaiah, Dr. Tilak Agerwala, Dr. Charles Tappert et al (2019) Challenges of Big Data Applications in Cloud Computing. CS&IT-CSCP:221-232
[16] P. Mandana Mohan, B. Murali Manohar (2021) Challenges in Big Data Analytics & Cloud Computing. International Journal of Business and Management Research 9(2):156-161
[17] Bo Li (2022) Research Review of Cloud Computing Technology Based on Big Data. Conference on Image Processing , Electronics and Computers 198-201
[18] Blend Berisha, Endrit Meziu, Isak Shabani et al (2022) Big Data Analytics in Cloud Computing: An Overview. J Cloud Comput 11(1):24
[19] Jayaraj T, J. Abdul Samath (2020) Secure and Cost-Effective Big-Data Analysis in Cloud Computing. International Journal of Scientific & Technology Research 9(2):3717-3720
[20] Mythreyee S, Poornima Purohit, Apoorva D.R , Harshitha R, Lathashree P.V et al (2017) A Study On Use of Big Data in Cloud Computing Environment. International Journal of Advance Research , Ideas and Innovations in Technology 3(3):1312-1318



10.22214/IJRASET

INDEX COPERNICUS
45.98

IMPACT FACTOR:
7.129

TOGETHER WE REACH THE GOAL

IMPACT FACTOR:
7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY