Journal of Global Research in Multidisciplinary Studies (JGRMS)
Volume 2 | Issue 7 | July 2026 | Pages 08–14
DOI: 10.5281/zenodo.21332231
e-ISSN: 3049-2963
REVIEW ARTICLE OPEN ACCESS
Available online: https://saanvipublications.com/journals/index.php/jgrms

A Survey of Tools and Techniques of Big Data Management in Cloud Environments

Mr. Himanshu Baraiya
Department of Master Computer Application, Lakshmi Narnia College of Technology, Bhopal, India
himanshub@lnct.ac.in
hemansu008@gmail.com

Received: 28 June 2026 | **Revised:** 3 July 2026 | **Accepted:** 9 July 2026

***Abstract*—Big data has emerged as a critical asset for organizations due to the rapid growth of data generated from social media, Internet of Things (IoT) devices, business transactions, and digital services. Managing such massive volumes of data requires scalable, efficient, and reliable computing infrastructures. Cloud computing provides a flexible platform for storing, processing, and analyzing big data through distributed resources and on-demand services. This survey paper presents an overview of big data management in cloud environments, focusing on the key tools and techniques used for data storage, processing, analytics, and security. The study discusses widely adopted tools such as Hadoop, MapReduce, Spark, HBase, and NoSQL databases, along with techniques including data mining, machine learning, predictive analytics, optimization methods, and text analytics. Furthermore, the paper highlights major challenges related to scalability, resource management, data privacy, and security in cloud-based big data systems. A comparative review of recent research works is also provided to identify current trends and future directions. The survey concludes that the integration of advanced analytics techniques with cloud computing can significantly improve the efficiency, reliability, and security of big data management.**

Keywords*—*Big Data, Cloud Computing, Hadoop, MapReduce, Spark, Cloud Security, Machine Learning, Data Mining.

I. Introduction

The rapid growth of data-intensive applications and digital transformation has significantly increased the demand for scalable, flexible, and cost-effective computing infrastructures to manage diverse workloads efficiently. Cloud environments promise several benefits such as reduced expenses and simplicity to service providers and service requester [1][2]. The definitions of cloud deployment and service delivery models, as well as of the essential characteristics of clouds, accepted by the community in the field, were discussed by the National Institute of Standards and Technology (NIST). The deployment models include public, private, hybrid, and community clouds, and Virtual Private Clouds (VPCs). The service delivery models include the infrastructure-as-a-Service (IaaS), the Platform-as-a-Service (PaaS), and the Software as-a-Service (SaaS). Finally, the distinguishing characteristics for this technology are broad network access, rapid elasticity, measured service, on-demand self-service, and resource pooling [3][4].

Big Data is a field related with the analysis and processing of large information that is available from various sources. Although Big Data may appear as a new discipline, it has been developing for years. The storage and analysis of large datasets have been a long-standing problem. Big Data is very large and complex that it is really tough to process and analyze the data. With traditional approaches it is difficult. Effective data management for Big Data sets is not possible with traditional RDBMS (Relational database management systems). Due to the size of Big Data, it is very difficult to extract the information in a proper and required manner. Bringing insights from the large amount of data is very much useful. In fact, it is the raw data that is processed into information. There are Different Types of Data processed by Big Data. Solutions can be human-generated (or) machine-generated, despite the fact that it is the machines that produce the systematic outcome Human-generated data is the result of human interaction with systems, such as online services and digital devices [5]. Examples of human-generated data are structured data, video, Textual data. Machine-generated data is generated by software programs and hardware devices in response to real-world events.

Previously to Big Data revolution, companies could not store all their archives for long periods nor efficiently manage huge data sets. Indeed, traditional technologies have limited storage capacity, rigid management tools and are expensive. They lack of scalability, flexibility and performance needed in Big Data context. In fact, Big Data management requires significant resources, new methods and powerful technologies. the different Big Data projects across the world, many Big Data models, frameworks and new technologies were created to provide more storage capacity, parallel processing and real-time analysis of different heterogeneous sources. In addition, new solutions have been developed to ensure data privacy and security. Compared to traditional technologies, such solutions offer more flexibility, scalability and performance. Furthermore, the cost of most hardware storage and processing solutions is continuously dropping due to sustainable technological advances [6].

A. *Structure of Paper*

This paper is organized as follows: Section II presents the fundamentals of big data and cloud environments, Section III discusses the major tools and techniques, Section IV describes the challenges and security techniques, Section V reviews recent literature, and Section VI concludes the paper with future research directions.

DOI: 10.5281/zenodo.21332231

II. BIG DATA MANAGEMENT IN CLOUD ENVIRONMENTS

Big data management in cloud environments focuses on the efficient storage, processing, integration, and analysis of massive datasets using scalable cloud computing resources. Cloud platforms provide flexible infrastructure, high availability [7], and cost-effective services that enable organizations to manage diverse data efficiently while supporting real-time analytics, security, and data-driven decision-making.

A. Characteristics of Big Data

Big data is defined by a set of key characteristics that determine its complexity and influence the methods used for its storage, processing, management, and analysis [8]. These characteristics, commonly referred to as the 5 Vs, provide a fundamental framework for understanding the nature of big data and its associated computational and analytical challenges, hence, the five Vs are as follows:

- **Volume:** The growing amount of data generated across various domains requires efficient storage and processing capabilities. Managing such large datasets demands scalability (vertical, horizontal, or both), high availability, and adequate bandwidth and performance to ensure timely data access.
- **Variety:** Modern applications generate data in multiple formats from diverse sources. Big Data frameworks address this heterogeneity by integrating structured, semi-structured, and unstructured data, including IoT data from internet traffic, smartphones, wearable devices, and other connected systems.
- **Velocity:** Data is produced and updated at different speeds depending on its source. While enterprise systems may update periodically, sensor networks generate continuous streams, requiring Big Data frameworks to efficiently process and integrate data arriving at varying rates.
- **Value:** Extracting meaningful insights is the primary objective of Big Data analytics. Large datasets become beneficial only when they provide valuable information that supports informed analysis and effective decision-making.
- **Veracity:** Ensuring data reliability is essential for accurate analysis and trustworthy outcomes. Organizations must maintain data confidentiality, integrity, availability, and accuracy so that both the information and the resulting analyses remain dependable.

B. Types of Big Data

Data are produced at unprecedented rates from various sources, such as financial, government, health, and social networks [9][10]. Such rapid growth of data can be attributed to smart devices, the IoT, etc. In recent decades, companies have failed to store data efficiently and for long periods [11]. This drawback relates to traditional technologies that lack adequate storage capacity and are costly. Meanwhile, big data requires new storage methods backed by powerful technologies [12]. Big data can be classified into several categories:

Figure 1 presents an efficient classification of big data based on data sources, content formats, storage models, staging processes, and processing approaches.

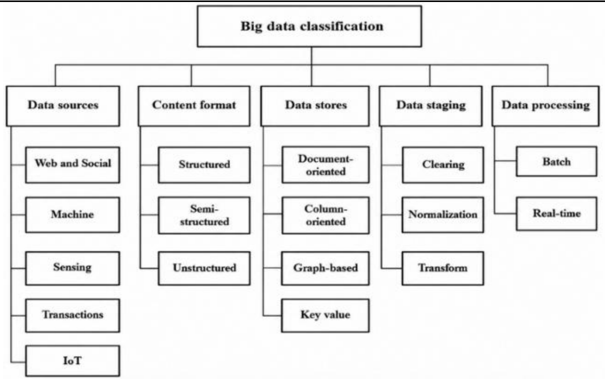

Fig. 1. Types of big data

1) Data source

- **Social media:** social media represents an important aspect of big data. Facebook, Twitter, emails, and microblogs are social media sources that generate massive amounts of data daily.
- **Machine generated:** data Software and hardware, such as medical devices, computers, and other types of machines that generate data without human interferences.
- **Sensing Various:** Various types of sensing devices that generate data and convert them into signals.
- Transaction Financial: business, and work data generate time-based dimensions that define data.
- **Io Tablets:** smartphones, and digital camera devices are connected over the Internet and thus generate huge amounts of data and information.

2) Content format source

- **Structured-data:** Structured-data are in a consistent order with a well-defined format. The advantage of structured-data is that they are easy to maintain, access, and store on computers [13]. Structured-data are stored in the form of rows and columns; an example is a Data Base Management System (DBMS).
- **Semi-structured data:** Semi-structured data can be considered as another form of structured-data. It inherits a few properties of structured-data that do not represent the data in database models. An example is Common Separated Value (CSV) files [14].
- **Unstructured data:** Unstructured data do not follow the formal structure rules of data models. Images, videos, text messages, and social media posts are examples of unstructured data.

3) Data store sources

- **Key value stores:** Key value stores are used to store and access data in key/value pairs. They are basically designed to store massive data and manage heavy loads. Apache HBase, Apache Cassandra, Redis, and Riak are examples of key value store databases.
- **Graph stores:** Graph stores are used to analyze data on the basis of the relationships between nodes, edges, and properties. Neo4j is an example of a graph store.
- **Column family stores:** Column family stores keep data and information within a column of a table at the same location on a disk in the same way a row store keeps row data together. Google Bigtable is an example of column family stores.

- **Document-oriented store:** Document-oriented stores offer complex data forms in multiple formats, such as XML, JSON, text, string, array, or binary forms. CouchDB and MongoDB are examples of document-oriented stores.

4) Data staging

- **Cleaning:** Cleaning is a process in which noisy data, outliers, and missing values are removed.
- **Transformation:** In data transformation, data are transformed in an appropriate format for analysis.
- **Normalization:** Normalization is a process used to reduce redundancies from data [15] .

5) Data processing

- **Batch data processing:** MapReduce-based systems are used to process data in the form of batches. Apache Hadoop, Apache Mahout, Skytree Server, and Dryad are examples of batch processing.
- **Real-time data processing:** Streaming systems, such as S4, are based on distributed frameworks that allow users to [16], design applications for processing continuous unbounded streams of data [17].

C. Cloud Service Models for Big Data Management

Cloud computing provides on-demand access to configurable computing resources [18], including applications, storage, servers, networks, and services, through scalable, virtualized, and multi-tenant infrastructure. Major providers such as Microsoft, Google, and Amazon offer services through Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS), enabling organizations to select suitable cloud models based on their computing and data management requirements.

1) Infrastructure as a service (IaaS)

IaaS delivers virtualized computing resources, including servers, storage, and networking, enabling scalable infrastructure for storing and processing large-scale datasets.

2) Platform as a Service (PaaS)

PaaS provides a development and deployment platform with built-in tools and services, allowing efficient creation, testing, and management of big data applications.

3) Software as a service (SaaS)

SaaS offers cloud-hosted software applications that enable users to access data management and analytics services through the internet without managing the underlying infrastructure.

Performance concerns in cloud environments are a serious obstacle for consumers. To avoid distrust, it is necessary to ensure a fair behavior of the system with respect to its different customers [19][20]. Due to sharing of resources, performance related issues are often caused by a minority of customers sending a high number of requests.

The owner of the cloud data centers can deploy the cloud. The atmosphere consists of either one Cloud or multiple Clouds. The deployment models include public, private, hybrid, and community clouds its shows in Figure 2 below:

D. Deployment Models In Cloud Enviroments

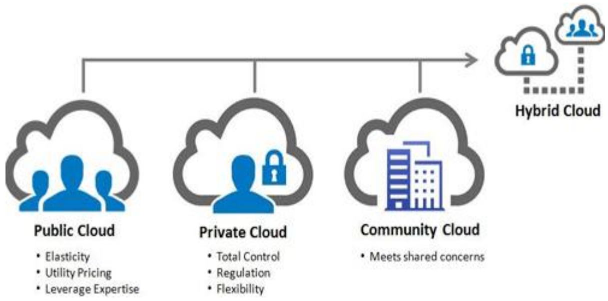

Fig. 2. Cloud deployment models

1) Public Cloud

The Cloud Infrastructure is available to the users over the Internet. The infrastructure is owned by a cloud provider. Resources such as applications and storage are made available to the public via the Internet. This setup is easy and inexpensive as the costs of the hardware, application and bandwidth costs are covered by the provider. Example: AWS, Microsoft Azure, IBM's Blue Cloud [21].

2) Private Cloud

The cloud infrastructure is operated by a single entity. It gives companies direct control over their data. The model offers services to a limited number of people behind firewalls. Hence, any security risk is minimal in this model. Examples: AWS, VMware.

3) Hybrid Cloud

This cloud computing environment uses a mix of on-premise, private cloud and third-party services, and public cloud services. Hence, it consists of functionalities of both private and public cloud. Example: Federal Agencies use private clouds when sensitive information is involved, whereas they utilize public clouds to share datasets with the general public or other government departments.

4) Community Cloud

This type of model follows a collaborative effort in which the cloud infrastructure is shared among various organizations belonging to a specific community with a common concern (such as security, jurisdiction, compliance, etc.) [22].

III. TOOLS FOR BIG DATA MANAGEMENT IN CLOUD

A wide range of cloud-based tools and frameworks have been developed to support efficient big data management. These technologies provide scalable storage, distributed processing, resource coordination, data querying, and high-performance analytics, enabling organizations to process and manage large, diverse, and rapidly growing datasets in cloud environments.

A. Apache Hadoop

Designed as an open-source distributed computing framework, this technology enables reliable storage and parallel processing of massive datasets across multiple nodes [23]. Its high scalability, fault tolerance, and cost-effectiveness make it a fundamental platform for big data management [24], with the Hadoop Distributed File System (HDFS) serving as its primary storage component.

DOI: 10.5281/zenodo.21332231

B. MapReduce

A distributed programming model that processes large datasets by dividing them into smaller tasks executed in parallel. The framework consists of a Map phase for data processing and a Reduce phase for aggregating intermediate results, ensuring efficient and fault-tolerant computation.

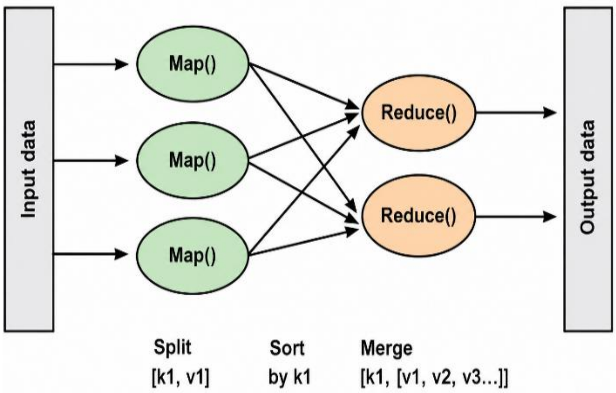

Fig. 3. MapReduce workflow

Figure 3 illustrates the MapReduce workflow, where input data is partitioned, processed in parallel, sorted, merged, and aggregated to generate output efficiently [25].

C. Hadoop Distributed File System (HDFS) (NameNode & DataNode)

Data Nodes are responsible for storing the blocks of files as determined by the Name Node. Data file to be stored is first split into one or more blocks internally [26]. Data Nodes serve the read/write requests from file system's client data. These are also responsible for creating, deleting and replicating blocks of file after being instructed by the Name Node.

D. Apache Spark

Apache Spark is an open-source distributed processing framework that was created at the UC Berkeley AMPLab. Spark is like HADOOP, but it is based on in-memory system to improve performance. It is a recognized analytics platform that ensures fast, easy-to-use and flexible computing. Spark handles complex analysis on large datasets. Indeed, Spark runs programs up to 100x faster than Hive and Apache Hadoop via MapReduce in memory system. Spark is based on the Apache Hive codebase [27]. Spark swaps out the physical execution engine of Hive. In addition, Spark offers Apis to support a fast application development in various languages, including Java, Python, and Scala. Spark [28][29]. is able to work with all file's storage systems that are supported by HADOOP.

E. Apache Hive

Hive is a data warehousing solution built on top of HADOOP. It provides SQL-like query language named HiveQL. The Apache Mahout free machine learning library's goal is to build scalable machine learning tools and data mining framework for use on analyzing Big Data on a distributed manner [30][31].

F. NoSQL Databases (HBase, Cassandra, MongoDB)

Schema-flexible NoSQL databases are designed to store and manage large volumes of structured, semi-structured, and unstructured data. HBase provides column-oriented storage for high-speed access to Hadoop data, Cassandra offers highly scalable and fault-tolerant distributed storage for continuous data availability, while MongoDB stores data in flexible document-based formats, making it suitable for dynamic and rapidly evolving applications.

G. ZooKeeper

Zookeeper maintains, configures, and names large amounts of data. It also provides distributed synchronization and group services. This instance enables distributed processes to manage and contribute to one another through a name space of data registers (*z*-nodes) that is shared and hierarchical, such as a file system [32]. Alone, Zookeeper is a distributed service that contains master and slave nodes and stores configuration information.

IV. Techniques For Big Data Management

Big data management employs a variety of techniques to support efficient data collection, storage, integration, processing, analysis, and security. These techniques enable scalable, reliable, and high-performance management of large, diverse, and rapidly growing datasets in cloud environments [33].

A. Optimization Techniques

Optimization Methods have been applied to solve quantitative problems in a lot of fields, such as physics, biology, engineering, and economics. Stochastic optimization, including genetic programming, evolutionary programming, and particle swarm optimization are useful and specific optimization techniques inspired by the process of nature. Real-time optimization is also required in many Big Data application, such as WSNs and ITSs [34]. Data reduction and parallelization are also alternative approaches in optimization problems.

B. Statistical Techniques

Statistics is the science to collect, organize, and interpret data. Statistical techniques are used to exploit correlation ships and causal relationships between different objectives [35]. Numerical descriptions are also provided by statistics. However, standard statistical techniques are usually not well suited to manage Big Data, and many researchers have proposed extensions of classical techniques or completely new methods.

C. Data Mining Techniques

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information that can be used to increase revenue, cuts costs, or both. It including clustering analysis, classification, regression and association rule learning [36]. It including C4.5, k-means, SVM, Apriori, EM, Naive Bayes, and Cart, etc. These ten algorithms cover classification, clustering, regression, statistical learning, association analysis, and linking mining, all of which are the most important problems in data mining research.

D. Predictive Analytics

Predictive Analytics (PA) is consisting of a variety of techniques that predict future outcomes based on historical data and current data. Predictive analysis is used to capture the relationships of data and discover the patterns [37]. PA which is primarily based on statistical methods, is highly applicable on many disciplines.

E. *Text Analytics*

Text Analytics is used for information retrieval from data. It is referred to techniques that extract information from textual data. Social network feeds, E-mails, blogs, online forums, and news and call center records are all examples of text data. Text analytics involve machine learning, statistical analysis and computational linguistics [38]. Text analytics enable to extract meaningful summaries from large scale data. Information Extraction, Text Summarization, Question Answering and Sentiment Analysis are some of the techniques used in text analytics [39][40].

F. *Security Techniques for Big Data Management in Cloud*

There are security measures that can be used to improve the cloud computing environment.

- **Encryption:** Since the data in any system present in a cluster, a hacker can easily steal the data from the system. This may become a serious issue for any company or organization to safeguard their data. To avoid this, may go for encrypting the data. Different encryption mechanisms can be used own different systems and the keys generated should be stored secretly behind firewalls. By choosing this method the data of the user may be kept securely.
- **Nodes authentication:** The node must be authenticated whenever it joins the cluster. If the node turns out to be a malicious cluster, then such nodes must not be authenticated.
- **Honeypot nodes:** The honeypot nodes appear to be like a regular node but are a trap. It automatically traps the hackers and does not allow any damage to happen to the data.
- **Access control:** The differential privacy and access control in the distributed environment are good measures of security. To prevent the information from leaking, use SELinux. Security-Enhanced Linux is a feature that provides a mechanism for enforcing an access control security policy through the use of Linux Security Modules in the Linux kernel.

V. LITERATURE REVIEW

The literature review highlights recent advances in big data management for cloud environments. Table I summarizes the reviewed studies by comparing their approaches, findings, advantages, challenges, and recommendations, providing a concise overview of current research trends and gaps.

B. Zhang, P. He, and N. Jin (2026) Presents a multi-scale big data analytics and interactive visualization system for IoT smart home energy management. High-resolution sensor data is cleaned, prepared, and analyzed using time-series analysis, correlation modeling, dimensionality reduction, anomaly detection, and causal contribution analysis. The results improve energy data interpretability and support data-driven energy-saving decisions by identifying key power consumers and peak usage patterns [41].

P. N. Deshmukh and V. H. Deshmukh (2025) an intelligent cloud security framework integrating AI-driven sensitivity classification, adaptive encryption, blockchain-based key management, and multi-cloud redundancy. Homomorphic encryption enables privacy-preserving computation, while AI-based encrypted search improves secure data retrieval. Experimental results demonstrate improved encryption speed, storage efficiency, threat mitigation, and retrieval performance [42].

S. Yang, Q. Xia, and B. Zhu (2024) Aims to enhance data interoperability and reuse by developing a comprehensive metadata scheme for standardized scientific data description. A data management cloud platform is established to support data collection, processing, storage, retrieval, monitoring, and sharing, ensuring effective data governance and seamless research management [43].

Y. Yang (2024) a decomposition synthesis load prediction algorithm with error compensation using statistical analysis and Kalman filtering. The approach improves load prediction accuracy and resource utilization by reducing scheduling uncertainty. Additionally, a multidimensional resource scheduling mechanism combining virtual machines and containers is introduced to address resource utilization imbalance [44].

S. Wu, B. Cui, and Y. Zhang (2023) present the current state of application of the complex generic environment big data platform, power grid engineering management technology and power grid engineering management big data analysis technology. Regarding RTBDT technology, this paper describes the framework, data cleaning technology, data synchronization technology, multi-stream data alignment technology, and batch stream integration data analysis technology [45].

Miao (2022) Improves information management efficiency in small and medium-sized enterprises by developing a cloud-based SaaS management system. The system uses Alibaba Cloud, Hadoop-based distributed storage and computing, and J2EE with the SSH framework. A virtual cluster and storage resource pool enable efficient enterprise data storage and resource sharing [46].

TABLE I. SUMMERY OF THE RECENT STUDY ON BIG DATA MANAGEMENT IN CLOUD ENVIRONMENTS

Authors	Approach	Findings	Advantages	Challenges	Recommendations
B. Zhang, P. He, and N. Jin (2026)	Multi-scale analytics and visualization for IoT energy data	Improved energy data interpretation and consumption analysis	Better decision support and energy optimization	Limited scalability across heterogeneous IoT environments	Develop scalable real-time analytics for diverse IoT ecosystems
P. N. Deshmukh and V. H. Deshmukh (2025)	AI-driven cloud security with encryption and blockchain	Enhanced secure storage and data retrieval	Strong privacy and threat protection	High computational overhead and deployment complexity	Design lightweight and scalable security frameworks
S. Yang, Q. Xia, and B. Zhu (2024)	Metadata-based cloud data management platform	Improved data interoperability and governance	Efficient data sharing and lifecycle management	Lack of unified metadata standards across platforms	Develop universal metadata standards for multi-cloud environments

Y. Yang (2024)	Load prediction with resource scheduling	Improved prediction accuracy and resource utilization	Reduced scheduling uncertainty	Dynamic workload adaptation remains limited	Integrate AI-based adaptive resource scheduling techniques
S. Wu, B. Cui, and Y. Zhang (2023)	Big data platform for power grid management	Enhanced real-time data processing and analysis	Improved engineering data management	Managing large-scale heterogeneous streaming data	Develop efficient real-time stream processing and integration methods
Miao (2022)	Cloud-based SaaS enterprise management system	Improved enterprise information management efficiency	Cost-effective distributed storage and resource sharing	Limited scalability and interoperability	Enhance scalability and support hybrid and multi-cloud deployments

Research Gap: Although numerous studies have explored big data management and cloud computing, several research gaps remain. Most existing research focuses on specific tools, techniques, or application domains rather than providing a comprehensive comparison of big data management approaches in cloud environments. Issues such as efficient resource allocation, real-time processing, interoperability across cloud platforms, and scalable data management remain largely unaddressed. Furthermore, challenges related to data security, privacy, and performance optimization continue to limit the effectiveness of cloud-based big data systems. Emerging technologies such as AI-driven management, edge computing, blockchain, and federated learning have shown potential, but their integration into big data management frameworks remains limited. Therefore, there is a need for comprehensive, intelligent cloud-based big data management solutions that enhance scalability, security, resource utilization, and real-time decision-making.

VI. Conclusion And Future Work

Big data management in cloud environments has emerged as a fundamental research area for supporting the storage, processing, analysis, and management of rapidly growing datasets. This survey reviewed the major tools and techniques used for cloud-based big data management by examining its fundamental concepts, cloud service and deployment models, widely adopted frameworks, data management techniques, security mechanisms, and recent research developments. The survey highlights that cloud computing provides scalable, flexible, and cost-effective infrastructure for managing heterogeneous data, while tools such as Hadoop, Spark, Hive, HDFS, NoSQL databases, and ZooKeeper enable efficient distributed storage and processing. Furthermore, techniques including optimization, statistical analysis, data mining, predictive analytics, and text analytics enhance data processing efficiency and support intelligent decision-making. The reviewed literature demonstrates continuous progress toward secure, scalable, and high-performance cloud-based big data management while identifying interoperability, resource optimization, real-time processing, and privacy as key research challenges.

Future research should focus on developing intelligent, integrated big data management frameworks that combine artificial intelligence, edge computing, blockchain, and federated learning. Further studies should improve real-time analytics, resource optimization, interoperability, security, privacy preservation, and energy-efficient cloud infrastructures to support scalable, reliable, and next-generation data-driven applications.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest: The author declares no conflict of interest.

How to cite this article:

H. Baraiya, "A Survey of Tools and Techniques of Big Data Management in Cloud Environments," *Journal of Global Research in Multidisciplinary Studies*, vol. 2, no. 7, pp. 08-14, Jul. 2026. DOI: 10.5281/zenodo.21332231

References

[1] B. Sotomayor, R. S. Montero, I. M. Llorente, and I. Foster, "Virtual Infrastructure Management in Private and Hybrid Clouds," *IEEE Internet Comput.*, vol. 13, no. 5, pp. 14–22, Sep. 2009, doi: 10.1109/MIC.2009.119.

[2] T. H. Noor, Q. Z. Sheng, S. Zeadally, and J. Yu, "Trust management of services in cloud environments," *ACM Comput. Surv.*, vol. 46, no. 1, pp. 1–30, Oct. 2013, doi: 10.1145/2522968.2522980.

[3] A. Joon, B. K. R. Janumpally, A. Gogineni, and P. Chatterjee, "Efficient Large-Scale Intrusion Identification and Prevention in Distributed Cloud Networks Using Artificial Intelligence," in *2025 5th International Conference on Intelligent Technologies (CONIT)*, HUBBALI, India: IEEE, Jun. 2025, pp. 1–8. doi: 10.1109/CONIT65521.2025.11167760.

[4] D. A. B. Fernandes, L. F. B. Soares, J. V. Gomes, M. M. Freire, and P. R. M. Inácio, "Security issues in cloud environments: a survey," *Int. J. Inf. Secur.*, vol. 13, no. 2, pp. 113–170, Apr. 2014, doi: 10.1007/s10207-013-0208-7.

[5] B. Suvarnamukhi and M. Seshashayee, "Big Data Concepts and Techniques in Data Processing," *Int. J. Comput. Sci. Eng.*, vol. 6, no. 10, pp. 712–714, Oct. 2018, doi: 10.26438/ijcse/v6i10.712714.

[6] A. Oussous, F.-Z. Benjelloun, A. Ait Lahcen, and S. Belfkih, "Big Data technologies: A survey," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 30, no. 4, pp. 431–448, Oct. 2018, doi: 10.1016/j.jksuci.2017.06.001.

[7] A. Siddiqa *et al.*, "A survey of big data management: Taxonomy and state-of-the-art," *J. Netw. Comput. Appl.*, vol. 71, pp. 151–166, Aug. 2016, doi: 10.1016/j.jnca.2016.04.008.

[8] P. Neves and J. Bernardino, "Big Data in the Cloud: A Survey," *Open J. Big Data*, vol. 1, pp. 3–20, 2015.

[9] A. Joon, S. A. Pahune, S. S. S. Neeli, and B. Boddu, "A Robust Artificial Intelligence-Based Financial Forecasting System to Enhance Accuracy with Machine Learning Algorithms," in *Proceedings of the Second International Conference on Advanced Computing and Systems*, W. Ding, A. Chakrabarti, M. Chakraborty, and S. Chakraborty, Eds., Cham: Springer Nature Switzerland, 2026, pp. 522–533.

[10] S. K. Madishetty and G. N. M. V. Kumar, "Intelligent Surgical Tables: Real-Time Patient Monitoring and Autonomous Adjustment Using IoT, Cloud Computing, and Cybersecurity," *J. Comput. Anal. Appl.*, vol. 31, no. 02, pp. 803–809, Feb, Feb. 2023, doi:

10.48047/jocaaa.2023.31.02.50.

[11] T. Cherian, "A Study And Survey Of Big Data Using Data Mining Techniques," *Int. J. Eng. Sci. Res. Technol.*, 2019.

[12] S. Kumar and M. Singh, "Big data analytics for healthcare industry: impact, applications, and tools," *Big Data Min. Anal.*, vol. 2, no. 1, pp. 48–57, Mar. 2019, doi: 10.26599/BDMA.2018.9020031.

[13] S. Aslam and M. A. Shah, "Load balancing algorithms in cloud computing: A survey of modern techniques," in *2015 National Software Engineering Conference (NSEC)*, IEEE, Dec. 2015, pp. 30–35. doi: 10.1109/NSEC.2015.7396341.

[14] D. A. Shafiq, N. Z. Jhanjhi, and A. Abdullah, "Load balancing techniques in cloud computing environment: A review," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 7, pp. 3910–3933, Jul. 2022, doi: 10.1016/j.jksuci.2021.02.007.

[15] B. Saraladevi, N. Pazhaniraja, P. V. Paul, M. S. S. Basha, and P. Dhavachelvan, "Big Data and Hadoop-a Study in Security Perspective," *Procedia Comput. Sci.*, vol. 50, pp. 596–601, 2015, doi: 10.1016/j.procs.2015.04.091.

[16] A. Fahad *et al.*, "A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis," *IEEE Trans. Emerg. Top. Comput.*, vol. 2, no. 3, pp. 267–279, Sep. 2014, doi: 10.1109/TETC.2014.2330519.

[17] A. K. Sandhu, "Big data with cloud computing: Discussions and challenges," *Big Data Min. Anal.*, vol. 5, no. 1, pp. 32–40, Mar. 2022, doi: 10.26599/BDMA.2021.9020016.

[18] V. K. Sharma, "Cloud Computing & IoT: 5G Focused IoT with Cloud Solutions," *Int. J. AI, BigData, Comput. Manag. Stud.*, vol. 6, no. 3, pp. 21–25, July, 2025, doi: 10.63282/3050-9416.IJAIBDCMS-V6I3P103.

[19] R. Krebs, C. Momm, and S. Kounev, "Metrics and techniques for quantifying performance isolation in cloud environments," *Sci. Comput. Program.*, vol. 90, pp. 116–134, Sep. 2014, doi: 10.1016/j.scico.2013.08.003.

[20] R. K. Kanneganti, "Security and Compliance Issues in Cloud-Based Deployments of Content and Workflow Management Systems," *Eastasouth J. Inf. Syst. Comput. Sci.*, vol. 2, no. 01, pp. 131–138, Aug. 2024, doi: 10.58812/esiscs.v2i01.1122.

[21] Tanmaykumar Shah, "Cloud-Based Data Warehousing for Marketing Agility: Lessons from FinTech Migrations to Snowflake and AWS," *Int. J. Adv. Res. Sci. Commun. Technol.*, vol. 4, no. 4, March, pp. 642–652, Mar. 2024, doi: 10.48175/IJARSCT-16000B.

[22] S. Verma, "Survey Paper of Cloud Computing," vol. 11, no. 01, pp. 431–433, 2022.

[23] M. R. C. Mukkolakkal, "1.5 Million Messages Per Second on 3 Machines: Benchmarking and Latency Optimization of Apache Pulsar at Enterprise Scale," in *22nd International Conference on Network and Service Management*, 2026, pp. 1–4, Mar. doi: https://doi.org/10.48550/arXiv.2603.29113.

[24] C. Xie, J. Gao, and C. Tao, "Big Data Validation Case Study," in *2017 IEEE Third International Conference on Big Data Computing Service and Applications (BigDataService)*, IEEE, Apr. 2017, pp. 281–286. doi: 10.1109/BigDataService.2017.44.

[25] H. Venkatesh, S. D. Perur, and N. Jalihal, "A Study on Use of Big Data in Cloud Computing Environment," *Int. J. Comput. Sci. Inf. Technol.*, 2015.

[26] K. Singh and R. Kaur, "Hadoop: Addressing challenges of Big Data," in *2014 IEEE International Advance Computing Conference (IACC)*, IEEE, Feb. 2014, pp. 686–689. doi: 10.1109/IAdCC.2014.6779407.

[27] K. Skracic and I. Bodrusic, "A Big Data solution for troubleshooting mobile network performance problems," in *2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, IEEE, May 2017, pp. 472–477. doi: 10.23919/MIPRO.2017.7973471.

[28] I. Soric, D. Dinjar, M. Stajcer, and D. Orescanin, "Efficient social network analysis in big data architectures," in *2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, IEEE, May 2017, pp. 1397–1400. doi: 10.23919/MIPRO.2017.7973640.

[29] M. Kari, "A comparative performance study of hashing algorithms: MD5,SHA-256, and SHA-512 using Java : 2 NEEDED," *Int. J. Curr. Eng. Technol.*, vol. 13, no. 06, pp. 619–626, Nov/Dec, 2023, doi: https://doi.org/10.14741/ijcet/v.13.6.14.

[30] J. P. Verma, B. Patel, and A. Patel, "Big Data Analysis: Recommendation System with Hadoop Framework," in *2015 IEEE International Conference on Computational Intelligence & Communication Technology*, IEEE, Feb. 2015, pp. 92–97. doi: 10.1109/CICT.2015.86.

[31] W. Almeida, L. Monteiro, A. Lima, R. Rodrigues, and F. Escobar, "Survey on Trends in Big Data: Data Management, Integration and Cloud Computing Environment," 2019.

[32] N. Khan *et al.*, "Big Data: Survey, Technologies, Opportunities, and Challenges," *Sci. World J.*, vol. 2014, pp. 1–18, 2014, doi: 10.1155/2014/712826.

[33] Mr. Jay Bharat Mehta, "Ai-Driven Test Engineering For Cloud-Native Systems," *Int. J. Data Sci. IoT Manag. Syst.*, vol. 5, no. 1, Jan. 2026, doi: 10.64751/ijdim.2026.v5i1.297.

[34] S. Zhang, "Research on the Construction of Information System of Enterprise Financial Management in the Big Data Environment," in *2025 International Conference on Information Technology, Communication Ecosystem and Management (ITCEM)*, IEEE, Dec. 2025, pp. 189–193. doi: 10.1109/ITCEM68692.2025.00043.

[35] J. Fan, F. Han, and H. Liu, "Challenges of Big Data analysis," *Natl. Sci. Rev.*, vol. 1, no. 2, pp. 293–314, Jun. 2014, doi: 10.1093/nsr/nwt032.

[36] M. Chen, S. Mao, and Y. Liu, "Big data: A survey. Mobile Networks and Applications," *Mob. Networks Appl.*, 2014.

[37] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," *Int. J. Inf. Manage.*, vol. 35, no. 2, pp. 137–144, Apr. 2015, doi: 10.1016/j.ijinfomgt.2014.10.007.

[38] S. Devarakonda, R. LINGAM, and V. Challa, "Confidence-Gated RAG for Adaptive Retrieval in Sequential Agents," in *ICLR 2026 Workshop on Logical Reasoning of Large Language Models*, ICLR, 2026, pp. 01–10, Apr.

[39] R. D. Patil and O. S. Jadhav, "Some Contribution of Statistical Techniques in Big Data : A Review," vol. 4, no. April, pp. 293–303, 2016.

[40] M. N. Paralath, "Bridging Astronomical Imagery and Text with Advanced Multi-Modal AI," in *2025 IEEE 7th International Conference on Computing, Communication and Automation (ICCCA)*, Greater Noida, India: IEEE, Nov. 2025, pp. 1–8. doi: 10.1109/ICCCA66364.2025.11325649.

[41] B. Zhang, P. He, and N. Jin, "A Multi-scale Big Data Analytics and Interactive Visualization Framework for IoT Smart Home Energy Efficiency Using Unity," in *2026 6th International Symposium on Computer Technology and Information Science (ISCTIS)*, IEEE, May 2026, pp. 1359–1365. doi: 10.1109/ISCTIS70043.2026.11572251.

[42] P. N. Deshmukh and V. H. Deshmukh, "Securing Big Data Storage in a Cloud Environment," in *2025 3rd DMIHER International Conference on Artificial Intelligence in Healthcare, Education and Industry (IDICAIHEI)*, IEEE, Nov. 2025, pp. 1–5. doi: 10.1109/IDICAIHEI65991.2025.11378052.

[43] S. Yang, Q. Xia, and B. Zhu, "Exploration of Data Reuse Patterns Based on Scientific Data Lifecycle in Big Data Environment," in *2024 9th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA)*, IEEE, Apr. 2024, pp. 280–284. doi: 10.1109/ICCCBDA61447.2024.10569511.

[44] Y. Yang, "Practice and Exploration of Multidimensional Resource Scheduling and Optimization in Cloud Computing Big Data Environment," in *2024 International Conference on Interactive Intelligent Systems and Techniques (IIST)*, IEEE, Mar. 2024, pp. 398–402. doi: 10.1109/IIST62526.2024.00113.

[45] S. Wu, B. Cui, and Y. Zhang, "Application of Real-Time Big Data Technology in Complex Power Engineering Management," in *2023 8th International Conference on Data Science in Cyberspace (DSC)*, IEEE, Aug. 2023, pp. 526–532. doi: 10.1109/DSC59305.2023.00082.

[46] X. Miao, "Application of Cloud Computing in Information Management System in the Era of Big Data," in *2022 2nd International Conference on Computational Modeling, Simulation and Data Analysis (CMSDA)*, IEEE, Dec. 2022, pp. 307–310. doi: 10.1109/CMSDA58069.2022.00063.