BIG DATA MINING AND ANALYTICS
ISSN 2096-0654 03/06 pp32–40
Volume 5, Number 1, March 2022
DOI: 10.26599/BDMA.2021.9020016

Big Data with Cloud Computing: Discussions and Challenges

Amanpreet Kaur Sandhu*

Abstract: With the recent advancements in computer technologies, the amount of data available is increasing day by day. However, excessive amounts of data create great challenges for users. Meanwhile, cloud computing services provide a powerful environment to store large volumes of data. They eliminate various requirements, such as dedicated space and maintenance of expensive computer hardware and software. Handling big data is a time-consuming task that requires large computational clusters to ensure successful data storage and processing. In this work, the definition, classification, and characteristics of big data are discussed, along with various cloud services, such as Microsoft Azure, Google Cloud, Amazon Web Services, International Business Machine cloud, Hortonworks, and MapR. A comparative analysis of various cloud-based big data frameworks is also performed. Various research challenges are defined in terms of distributed database storage, data security, heterogeneity, and data visualization.

Key words: big data; data analysis; cloud computing; Hadoop

1 Introduction

With recent technological advancements, the amount of data available is increasing day by day. For example, sensor networks and social networking sites generate overwhelming flows of data. In other words, big data are produced from multiple sources in different formats at very high speeds[1]. At present, big data represent an important research area. Big data are rapidly produced and are thus difficult to store, process, or manage using traditional software. Big data technologies are tools that are capable of storing meaningful information in different types of formats. For the purpose of meeting users' requirements and analyzing and storing complex data, a number of analytical frameworks have been made available to aid users in analyzing complex structured and unstructured data[2]. Several programs, models, technologies, hardware, and software have been proposed and designed to access the information from big data. The main objective of these technologies is to store reliable and accurate results for big data[3]. In addition, big data require state-of-the-art technology to efficiently store and process large amounts of data within a limited run time.

Three different types of big data platforms are interactive analysis tools, stream processing tools, and batch processing tools[4]. Interactive analysis tools are used to process data in interactive environments and interact with real-time data. Apache Drill and Google's Dremel are the frameworks for storing real-time data. Stream processing tools are used to store information in continuous flow[5]. The main platforms for storing streaming information are S4 and Strom. Hadoop infrastructure is utilized to store information in batches. Big data techniques are involved in various disciplines, such as signal processing, statistics, visualization, social network analysis, neural networks, and data mining[6]. Mohajer et al.[7] designed an interactive gradient algorithm that receives controlled messages from neighboring nodes. The proposed method uses a self-optimization framework for big data.

2 Definitions of Big Data

Big data are huge in size and are difficult to manage and analyze relative to traditional data. Storing big data

• Amanpreet Kaur Sandhu is with University Institute of Computing, Chandigarh University, Mohali 140413, India. E-mail: aman123.brar@gmail.com.
* To whom correspondence should be addressed.
Manuscript received: 2021-06-11; revised: 2021-09-12; accepted: 2021-09-13

requires scalable architecture and efficient storage and manipulation. Table 1 presents the existing definitions of big data.

2.1 Characteristics of big data

Big data are characterized by three Vs: volume, velocity, and variety. These characteristics were introduced by Gartner to define the various challenges in big data[12]. With new-generation architecture, data are now stored in different types of formats; hence, the three Vs may be extended to five Vs, namely, volume, velocity, variety, value, and veracity[13].

(1) Volume: Data are generated by multiple sources (sensors, social networks, smartphones, etc.) and are continuously expanding. The Internet produces global data in large increments. In 2012, approximately 2.5 exabytes (EB) of data were produced every day. According to the report of International Data Cooperation, the volume of data in 2013 doubled, reaching 4.4 zettabytes (ZB). In 2020, the volume of data reached 40 ZB. Table 2 shows the names of the units of data that can be measured in bytes[14].

(2) Velocity: Data are exponentially growing at high speeds. Millions of connected devices are added on a daily basis, thereby leading to increases in not only volume but also velocity[15,16]. One relevant example is YouTube, which generates big data at high speeds[17,18]. Table 3 presents the number of users in India who had used social media networks by February 2021. Figure 1

Table 1 Definitions of big data.

Reference	Author's name	Definition
[8]	Batty	Big data are massive in size and cannot fit into Excel spreadsheets comprising approximately 16 000 columns and 1 million rows.
[9]	Havens et al.	Big data cannot be loaded into local storage devices (computer memory).
[10]	Fisher et al.	Big data cannot be easily processed and managed in a straightforward manner.
[11]	The State Council of People's Republic of China	Big data have several characteristics, such as high application value, fast access speed, large volume, and multiple types.
[12]	Bayer and Laney	Big data have large volume, variety, and velocity that demand cost effectiveness and are helpful in decision making.

Table 2 Units of data.

Name of unit	Equals	Size in bytes
Bit	1 or 0	1/8
Nibble	4 bits	1/2
Byte	8 bits	1
Kilobyte (KB)	1024 bytes	2^{10}
Megabyte (MB)	1024 KB	2^{20}
Gigabyte (GB)	1024 MB	2^{30}
Terabyte (TB)	1024 GB	2^{40}
Petabyte (PB)	1024 TB	2^{50}
Exabyte (EB)	1024 PB	2^{60}
Zettabyte (ZB)	1024 EB	2^{70}
Yottabyte (YB)	1024 ZB	2^{80}

Table 3 Users in India as of February 2021.

Application name	Count	Application name	Count
WhatsApp	53 Crore	Instagram	21 Crore
YouTube	44.8 Crore	Twitter	1.75 Crore
Facebook	41 Crore		

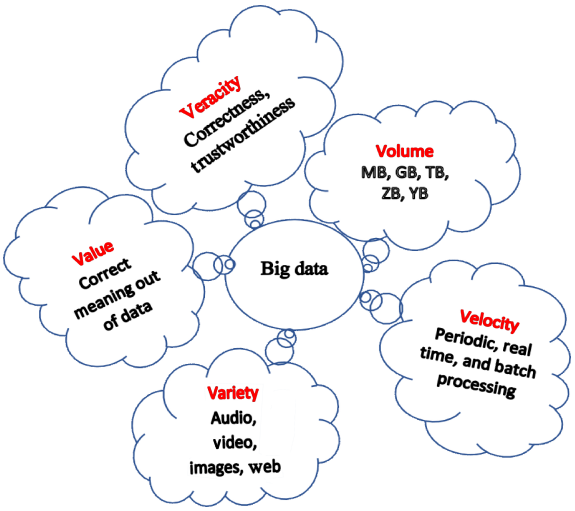

Fig. 1 Five Vs of big data.

shows the five Vs of big data.

(3) Variety: Data are generated in multiple formats via social networks, smartphones, or sensors. These tools produce data in the form of data logs, images, videos, audio, documents, and text. Data may also be structured, semistructured, and unstructured[19].

(4) Value: Value is an important characteristic of big data. It relates to how data can be dealt with and converted into meaningful information[20].

(5) Veracity: Veracity refers to the quality, correctness, and trustworthiness of data. Therefore, maintaining veracity in data is mandatory[21,22]. For example, data in huge amounts create confusion, whereas small amounts of data can convey incomplete or half information.

2.2 Types of big data

Data are produced at unprecedented rates from various sources, such as financial, government, health, and social networks. Such rapid growth of data can be attributed to smart devices, the Internet of Things, etc. In the last decades, companies have failed to store data efficiently and for long periods[23,24]. This drawback relates to traditional technologies that lack adequate storage capacity and are costly. Meanwhile, big data require new storage methods backed by powerful technologies[25,26]. Big data can be classified into several categories. Figure 2 depicts the classification of big data. Table 4 summarizes the definitions of various types of big data.

3 Big Data with Machine Learning

The main function of machine learning techniques is to discover knowledge and make intelligent decisions. Machine learning is used in various real-world applications, such as data mining, recognition systems, recommendation engines, and autonomous control systems. The machine learning domain can be divided into three areas, namely, supervised learning, unsupervised learning, and reinforcement learning[35].

3.1 Data streaming learning

Various real-time world technologies, such as stock management, network traffic, and credit card transactions, generate huge datasets. Data mining plays an important role in finding interesting patterns and fetching values from hidden streams and datasets. Traditional data mining techniques are related to clustering, association rule mining, accuracy, scalability, and classification, whereas big data are related to dynamic environments[36,37].

3.2 Deep learning

Deep learning is another important aspect in the field of machine learning and pattern recognition. It allows predictive analysis and involves natural language processing, speech recognition, and computer vision. The application of deep learning is to resolve the issues in data analysis and help extract complex datasets from huge amounts of data. Deep learning is called hierarchical learning because it extracts data from complex datasets at different levels. It is very helpful in the analysis of large volumes of data, information retrieval, data tagging, and discrimination tasks (e.g., prediction and classification)[38].

4 Cloud Computing

Cloud computing offers a cost-efficient and scalable solution to store big data. According to the National Institute for Standards and Technology, "Cloud Computing is based on pay-per-use services for enabling convenient, on-demand network access to a shared pool of configurable computing resources such as servers, networks, and services that can be rapidly provisioned and released with minimal management effort or service provider interaction". Cloud computing services can be

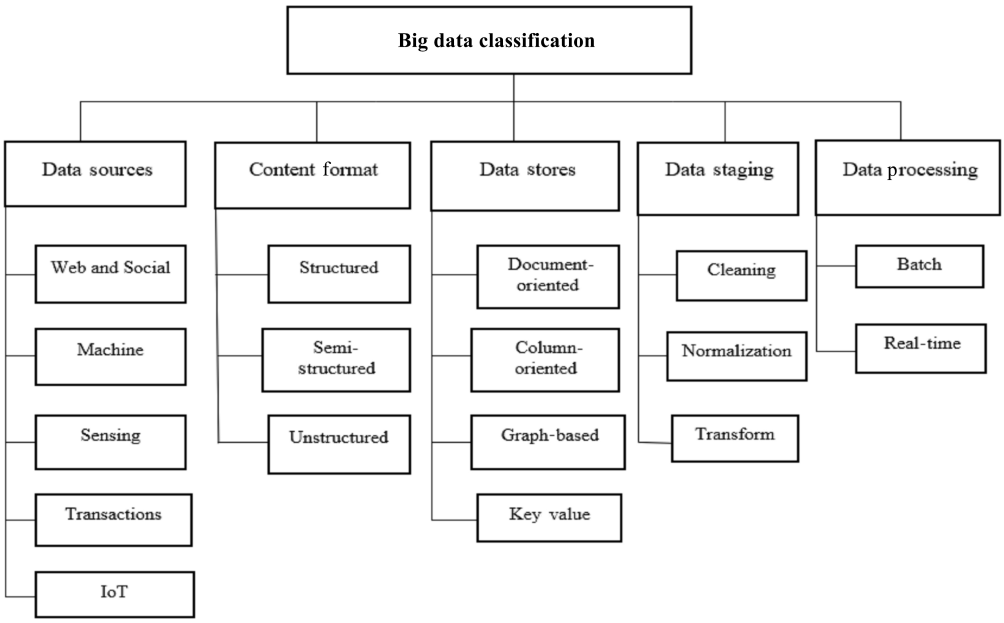

Fig. 2 Types of big data.

Table 4 Types of big data.

Type	Category	Explanation
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[27].
	Machine-generated data	Software and hardware, such as medical devices, computers, and other types of machines that generate data without human interferences.
	Sensing	Various types of sensing devices that generate data and convert them into signals[28].
	Transaction	Financial, business, and work data generate time-based dimensions that define data.
	IoT	Tablets, smartphones, and digital camera devices are connected over the Internet and thus generate huge amounts of data and information.
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. Structured-data are stored in the form of rows and columns; an example is a DataBase Management System (DBMS)[29].
	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[30].
	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.
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[31].
	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[32].
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[33].
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 design applications for processing continuous unbounded streams of data[34].

classified into the following three categories[39]:

(1) Infrastructure as a Service (IaaS): These services are basically based on the principle of "pay for what you need". It provides high-performance computing to customers. Amazon Web Services (AWS), Elastic Compute Cloud, and Simple Storage Services (S3) are examples of IaaS. AWS and S3 provide online storage services. At nominal charges, customers can easily access the world's largest data centers. At present, three companies provide IaaS landscape services: Google, Microsoft, and HP. Google provides Google Compute Engine to access IaaS services. Microsoft also provides a cloud platform through its Window Azure Platform. HP offers HP Cloud, which is designed by NASA and Rack Space.

(2) Software as a Service (SaaS): With the help of the Internet, all applications are run on remote cloud infrastructure in SaaS. To access SaaS services, users need an Internet connection and a web browser, such as Google Chrome or Internet Explorer[40]. Users connect to a desktop environment via a virtual machine, in which all software programs are installed. SaaS provides more facilities to users than IaaS.

(3) Platform as a Service (PaaS): It provides a runtime environment to users. It allows users to create, test, and run web applications. Users can easily access PaaS on the basis of the pay-per-use mode using an Internet connection. PaaS provides the infrastructure (networking, storage, and services) and platform (DBMS, business intelligence, middleware) for running a web application life cycle. Examples of PaaS include Microsoft Azure and Google Cloud[41].

The cloud computing environment has two important aspects: the frontend and the backend. From the frontend side, users access cloud services through an Internet connection; at the backend, all cloud services are run. Figure 3 shows the various types of cloud computing services[42].

Big data and cloud computing are closely associated. With technological changes, big data models provide distributed processing, parallel technologies, large storage capacity, and real-time analysis of heterogeneous databases. Data security and privacy are also considered in big data models. Big data require large amounts of storage space and thus entail the use of cloud computing. Cloud computing offers scalability and cost savings[43]. Moreover, it provides massive amounts of storage capacity and processing power.

Cloud computing works on different types of technologies, such as distributed storage and virtualization, and processes data for different types of tasks. It accesses distributed queries over multiple datasets and gives responses in a timely manner. Hadoop plays an important role in the storage of distributed datasets on the cloud[44]. The Hadoop Distributed File System (HDFS) stores large amounts of data in distributed form. The HDFS is a data storage management system in Hadoop. The advantage of the HDFS is that it is cost effective and capable of managing thousands of nodes in a cluster and massive amounts of unstructured data. It works as a batch processing system with latency operations.

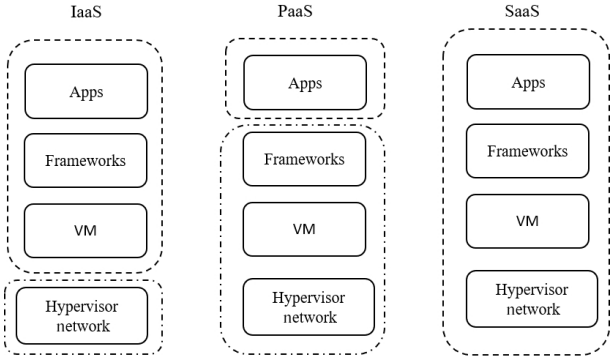

Fig. 3 Cloud computing services.

Moreover, Hadoop increases system performance and avoids network congestion. The HDFS is based on master-slave architecture. The HFDS comprises various types of daemons, such as DataNode, NameNode, Secondary NameNode, Resource Manager, and Node Manager. NameNode and Resource Manager work as master nodes, while DataNode and Node Manager work as slave nodes. In addition, it avoids fault tolerance with the help of data replication on various servers. The primary NameNode is used to solve computation problems and establishes coordination with DataNode. Secondary NameNode manages the availability and replication of data. The relationship between big data and cloud computing is shown in Fig. 4.

Cloud computing provides various types of facilities, such as processing, computation, and storage of big data. Moreover, the cloud computing infrastructure offers an efficient and effective platform to determine the storage requirements of big data analysis. It is also correlated with new patterns for the analysis of various types of resources that are available in the cloud. Several cloud-based technologies have been developed to deal with big data for parallel processing. MapReduce (MapR) is an

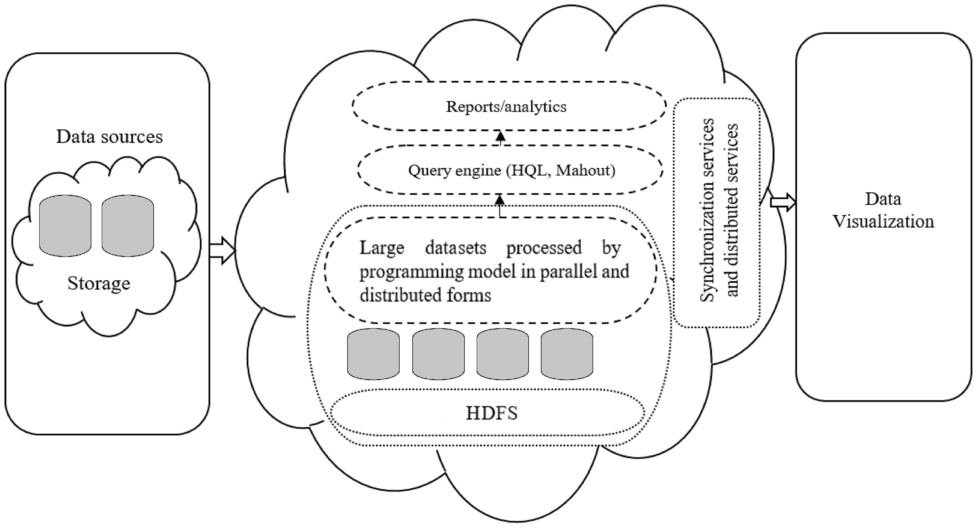

Fig. 4 Big data and cloud computing.

example of big data processing in a cloud environment that allows the storage of massive amounts of data in a cluster[45].

In other words, MapR is an efficient and cost effective model for processing big data. The MapR framework comprises the map and reduce functions for handling big data.

Cloud computing also plays an important role in distributed system environments by facilitating storage, boosting computing power, and aiding network communication. Big data technologies store data in cloud clusters rather than in local storage file systems. Several companies provide big data cloud platforms. Moreover, various cloud computing platforms are available to store big data. Table 5 shows a comparative analysis of big data cloud frameworks for storing massive amounts of data[46]. Cloud services such as Microsoft Azure, Google Cloud, AWS, IBM, Hortonworks, and MapR are compared on the basis of various parameters.

5 Research Issues in Big Data

As data are growing at exponential rates, a number of issues and problems emerge during the processing and storage of big data. Few tools are available to resolve these issues and problems in a cloud environment. Technologies, such as PigLatin, Dryad, MongoDB, Cassandra, and MapR, are not able to resolve these issues in big data processing. Even with the help of Hadoop and MapR, users cannot execute queries on databases, and they have low-level infrastructures for data processing and management. Some issues and problems in big data are summarized as follows[47]:

(1) Distributed database storage system: Numerous technologies are used to store and retrieve huge amounts of data. Cloud computing is an important aspect of big data. Big data are generated by multiple devices on a daily basis. At present, the main issue in distributed frameworks is the storage of data in a straightforward manner and the processing and migration of data between distributed servers.

(2) Data security: Security threats are an important issue in a cloud computing environment. Cloud computing has been transformed with modern information and communication technologies, and several types of unresolved security threats exist in big data. Data security threats are magnified by the variety, velocity, and volume of big data. Meanwhile, various issues and threats, such as the availability of data, confidentiality, real-time monitoring, identity and access authorization control, integrity, and privacy, exist in big data when used with cloud computing frameworks. Therefore, data security must be measured once data are outsourced to cloud service providers[48].

(3) Heterogeneity: Big data are heterogeneous in nature because data are gathered from multiple devices in different formats, such as images, videos, audio, and text. Before loading data into a warehouse, they need to be transformed and cleaned, and the processes present challenges in big data[49]. Combining all unstructured data and reconciling them for use in report creation are incredibly difficult to achieve in real-time.

(4) Data processing and cleaning: Data storage and acquisition require preprocessing and cleaning, which involves data merging, data filtering, data consistency, and data optimization. Thus, processing and cleaning data are difficult because of the wide variety of data sources[50]. Moreover, data sources may contain noise and errors, or they may be incomplete. The challenge is how to clean large amounts of data and how to determine whether such data are reliable.

(5) Data visualization: Data visualization is a technique to represent complex data in a graphical form for clear understanding. If the data are structured, then they can be easily represented in the traditional graphical way. If the data are unstructured or semistructured, then they are difficult to visualize with high diversity in real-time.

6 Conclusion

In the last decades, the size of data has grown, and it continues to increase day by day. Data are generated in different formats (variety) by multiple sources. Therefore, the variety of data is also expanding. Mobile devices and sensor networks that are connected generate data at very high speeds (velocity). Cloud computing services are used to process, analyze, and store data without the need for a dedicated space and maintenance of expensive computer hardware and software. This study reviews the relationship between big data and cloud computing. Furthermore, a comparative analysis of big data and cloud services is performed. Big data involve various issues and problems, such as distributed database storage, data privacy/security, and heterogeneity/data formats.

Table 5 Big data cloud frameworks.

	Microsoft Azure	Google Cloud	AWS	IBM	Hortonworks	MapR
Founding date	Oct. 2008	Oct. 2015	2006	2011	June 2011	2009
Big data analytics	Provides Azure HDInsight services	Provides Google Cloud Dataproc	Provides Amazon Elastic search services	Provides various IBM analysis engines	Provides Hortonworks Data Platform (HDP)	Provides MapR data analysis platform
Types of software	Open-source framework	Open-source framework	Open-source framework	Open-source framework	Open-source framework	Licensed
Content format	Only unstructured data format	Structured, semistructured, and unstructured	Structured, semistructured, and unstructured	Only unstructured data format	Structured, semistructured, and unstructured	Structured, semistructured, and unstructured
Types of OS supported	Windows Server, Ubuntu14	Debian 8	Linux, Ubuntu, CentOS	Only CentOS7	Linux, Ubuntu	CentOS, RedHat, Ubuntu
Various applications	Access batch and stream processes	Access machine learning, streaming, and batch process application	Real-time, logs analytics, and stream analytics	Data analytics	Real-time and stream analytics	Real-time, logs analytics, and stream analytics
Framework execution	Hadoop	Big Query	Elastic MapReduce	Elastic MapReduce	HDFS, YARN, MapReduce2	MapR
Big data storage framework	Microsoft Azure	Google Cloud Services	S3	IBM Cloud Object Storage	Hortonworks Data Platform	MapR Data Analysis Platform
Types of storage	Distributed	Distributed	Distributed	Distributed	Centralized	Distributed
Content format	XML	JSON, CSV	Any	Any	Any	Any
Storage size limit	Limited	Limited	Unlimited	Unlimited	Limited	Unlimited
Metadata	Yes	Yes	Yes	Yes	No	Yes
Relational database management system	SQL Azure	Cloud SQL	Oracle or MySQL	PostgreSQL, Oracle, MySQL	SQL	SQL
Big data warehouse	SQL Azure Data Warehouse	Big Query	Amazon Red Shift	Db2 warehouse	HIVE warehouse	Data Warehouse Optimization (DWO)
Content format	ORC, RC, Parquet	JSON, CSV	ORC, CSV, TSV	CSV	ORC	JSON
NoSQL (Not only SQL) database system	Stored in table format	AppEngine data store	DynamoDB	Apache Accumulo	MongoDB	MongoDB
Streaming processing	SteamInsight	Search application programming interface	Nothing prepackaged	Apache Spark	Storm, Spark Streaming, and Flink	Apache Spark
Machine learning	Mahout	Prediction application programming interface	Mahout	Hadoop	Hortonworks Data Platform	HPE Intelligent Data Platform

References

[1] I. A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, and S. U. Khan, The rise of 'big data' on cloud computing: Review and open research issues, *Inform. Syst.*, vol. 47, pp. 98–115, 2015.

[2] J. H. Yu and Z. M. Zhou, Components and development in big data system: A survey, *J. Electr. Sci. Technol.*, vol. 17, no. 1, pp. 51–72, 2019.

[3] S. Kumar and K. K. Mohbey, A review on big data based parallel and distributed approaches of pattern mining, *J. King Saud Univ. – Comput. Inform. Sci.*, doi: 10.1016/j.jksuci.2019.09.006.

[4] Y. N. Liu, N. Li, X. Zhu, and Y. Qi, How wide is the application of genetic big data in biomedicine, *Biomed. Pharmacother.*, vol. 133, p. 111074, 2021.

[5] V. Subramaniyaswamy, V. Vijayakumar, R. Logesh, and V. Indragandhi, Unstructured data analysis on big data using map reduce, *Procedia Comput. Sci.*, vol. 50, pp. 456–465, 2015.

[6] S. Maitrey and C. K. Jha, MapReduce: Simplified data analysis of big data, *Procedia Comput. Sci.*, vol. 57, pp. 563–571, 2015.

[7] A. Mohajer, M. Barari, and H. Zarrabi, Big data based self-optimization networking: A novel approach beyond cognition, *Intell. Automat. Soft Comput.*, doi: 10.1080/10798587.2017.1312893.

[8] M. Batty, Big data, smart cities and city planning, *Dialogues in Human Geography*, vol. 3, no. 3, pp. 274–279, 2013.

[9] T. C. Havens, J. C. Bezdek, C. Leckie, L. O. Hall, and M. Palaniswami, Fuzzy c-means algorithms for very large, *IEEE Trans. Fuzzy Syst.*, vol. 20, no. 6, pp. 1130–1146, 2012.

[10] D. Fisher, R. Deline, M. Czerwinski, and S. Drucker, Interactions with big data analytics, *Interactions*, vol. 19, no. 3, pp. 50–59, 2012.

[11] The State Council of the People's Republic of China, Action plan for promoting big data development, (in Chinese), http://www.gov.cn/zhengce/content/2015-09/05/content_10137.htm, 2015.

[12] M. A. Beyer and D. Laney, The importance of 'big data': A definition, Stamford, CT, USA: Gartner, G00235055, 2012.

[13] L. Rabhi, N. Falih, A. Afraites, and B. Bouikhalene, Big data approach and its applications in various fields: Review, *Procedia Comput. Sci.*, vol. 155, pp. 599–605, 2019.

[14] F. Ridzuan and W. M. N. Wan Zainon, A review on data cleansing methods for big data, *Procedia Comput. Sci.*, vol. 161, pp. 731–738, 2019.

[15] D. A. Shafiq, N. Z. Jhanjhi, and A. Abdullah, Load balancing techniques in cloud computing environment: A review, *J. King Saud Univ. – Comput. Inform. Sci.*, doi: 10.1016/j.jksuci.2021.02.007.

[16] S. Amamou, Z. Trifa, and M. Khmakhem, Data protection in cloud computing: A survey of the state-of-art, *Procedia Comput. Sci.*, vol. 159, pp. 155–161, 2019.

[17] P. J. Sun, Security and privacy protection in cloud computing: Discussions and challenges, *J. Netw. Comput. Appl.*, vol. 160, p. 102642, 2020.

[18] R. Nachiappan, B. Javadi, R. N. Calheiros, and K. M. Matawie, Cloud storage reliability for Big Data applications: A state of the art survey, *J. Netw. Comput. Appl.*, vol. 97, pp. 35–47, 2017.

[19] A. O'Driscoll, J. Daugelaite, and R. D. Sleator, 'Big data', Hadoop and cloud computing in genomics, *J. Biomed. Inform.*, vol. 46, no. 5, pp. 774–781, 2013.

[20] S. Karimian-Aliabadi, D. Ardagna, R. Entezari-Maleki, E. Gianniti, and A. Movaghar, Analytical composite performance models for Big Data applications, *J. Netw. Comput. Appl.*, vol. 142, pp. 63–75, 2019.

[21] H. F. Yu, A priori algorithm optimization based on Spark platform under big data, *Microprocess. Microsyst.*, vol. 80, p. 103528, 2021.

[22] M. Muniswamaiah, T. Agerwala, and C. Tappert, Big data in cloud computing review and opportunities, *Int. J. Comput. Sci. Inform. Technol.*, vol. 11, no. 4, pp. 43–57, 2019.

[23] T. Cherian and H. Bhadkamkar, A study and survey of big data using data mining techniques, *Int. J. Eng. Sci. Res. Technol.*, vol. 6, no. 10, pp. 169–174, 2017.

[24] A. Mehmood, I. Natgunanathan, Y. Xiang, G. Hua, and S. Guo, Protection of big data privacy, *IEEE Access*, vol. 4, pp. 1821–1834, 2016.

[25] S. Kumar and M. Singh, Big data analytics for healthcare industry: Impact, applications, and tools, *Big Data Mining Analytics*, vol. 2, no. 1, pp. 48–57, 2019.

[26] S. Kumar and M. Singh, A novel clustering technique for efficient clustering of big data in Hadoop Ecosystem, *Big Data Mining Analytics*, vol. 2, no. 4, pp. 240–247, 2019.

[27] C. K. Leung, Y. B. Chen, S. Y. Shang, and D. Y. Deng, Big data science on COVID-19 data, in *Proc. of 2020 IEEE 14th Int. Conf. Big Data Science and Engineering*, Guangzhou, China, 2020, pp. 14–21.

[28] M. S. Mahmud, J. Z. Huang, S. Salloum, T. Z. Emara, and K. Sadatdiynov, A survey of data partitioning and sampling methods to support big data analysis, *Big Data Mining Analytics*, vol. 3, no. 2, pp. 85–101, 2020.

[29] S. Aslam and M. A. Shah, Load balancing algorithms in cloud computing: A survey of modern techniques, in *Proc. of 2020 IEEE National Software Engineering Conference*, doi: 10.1109/NSEC.2015.7396341.

[30] D. A. Shafiq, N. Z. Jhanjhi, and A. Abdullah, vLoad balancing techniques in cloud computing environment: A review, *Journal of King Saud University-Computer and Information Sciences*, doi: 10.1016/j.jksuci.2021.02.007.

[31] A. Oussous, F. Z. Benjelloun, A. Ait Lahcen, and S. Belfkih, Big data technologies: A survey, *J. King Saud Univ. – Comput. Inform. Sci.*, vol. 30, no. 4, pp. 431–448, 2018.

[32] R. Misra, B. Panda, and M. Tiwary, Big data and ICT applications: A study, in *Proc. 2nd International Conference on Information and Communication Technology for Competitive Strategies*, https://doi.org/10.1145/2905055.2905099, 2016.

[33] B. Saraladevi, N. Pazhaniraja, P. V. Paul, M. S. S. Basha, and P. Dhavachelvan, Big data and Hadoop–A study in security perspective, *Procedia Computer Science*, vol. 50, pp. 596–601, 2015.

[34] A. Fahad, N. Alshatri, Z. Tari, A. Alamri, I. Khalil, A. Y. Zomaya, S. Foufou, and A. Bouras, A survey of clustering algorithms for Big Data: Taxonomy and empirical analysis, *IEEE Trans. Emerg. Top. Comput.*, vol. 2, no. 3, pp. 267–279, 2014.

[35] A. Katal, M. Wazid, and R. H. Goudar, Big data: Issues, challenges, tools and good practices, in *Proc. 6th Int. Conf. Contemporary Computing*, Noida, India, 2013, pp. 404–409.

[36] C. W. Tsai, C. F. Lai, H. C. Chao, and A. V. Vasilakos, Big data analytics: A survey, *J. Big Data*, vol. 2, no. 1, pp. 1–32, 2015.

[37] Z. Lv, H. Song, P. Basanta-Val, A. Steed, and M. Jo, Next-generation big data analytics: State of the art, challenges, and future research topics, *IEEE Trans. Ind. Informatics*, vol. 13, no. 4, pp. 1891–1899, 2017.

[38] K. S. Jadon, R. S. Bhadoria, and G. S. Tomar, A review on costing issues in big data analytics, in *Proc. 2015 International Conference on Computational Intelligence and Communication Networks* (*CICN*), Jabalpur, India, 2016, pp. 727–730.

[39] O. Y. Al-Jarrah, P. D. Yoo, S. Muhaidat, G. K. Karagiannidis, and K. Taha, Efficient machine learning for big data: A review, *Big Data Res.*, vol. 2, no. 3, pp. 87–93, 2015.

[40] G. S. Bhathal and A. Singh, Big Data: Hadoop framework vulnerabilities, security issues and attacks, *Array*, vols. 1&2, p. 100002, 2019.

[41] J. Hurwitz, A. Nugent, F. Halper, and M. Kaufman, *Big Data for Dummies*. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2013.

[42] V. P. Lalitha, M. Y. Sagar, S. Sharanappa, S. Hanji, and R. Swarup, Data security in cloud, in *Proc. of 2017 Int. Conf. Energy, Communication, Data Analytics and Soft Computing*, Chennai, India, pp. 3604–3608, 2017.

[43] C. L. Philip Chen and C. Y. Zhang, Data-intensive applications, challenges, techniques and technologies: A survey on Big Data, *Inform. Sci.*, vol. 275, pp. 314–347, 2014.

[44] S. Salloum, J. Z. Huang, and Y. He, Random sample partition: A distributed data model for big data analysis, *IEEE Trans. Ind. Inform.*, vol. 15, no. 11, pp. 5846–5854, 2019.

[45] L. Q. Kong, Z. F. Liu, and J. G. Wu, A systematic review of big data-based urban sustainability research: state-of-the-science and future directions, *J. Clean. Prod.*, vol. 273, p. 123142, 2020.

[46] P. Pääkkönen and D. Pakkala, Reference architecture and classification of technologies, products and services for big data systems, *Big Data Res.*, vol. 2, no. 4, pp. 166–186, 2015.

[47] M. Wook, N. A. Hasbullah, N. M. Zainudin, Z. Z. A. Jabar, S. Ramli, N. A. M. Razali, and N. M. M. Yusop, Exploring big data traits and data quality dimensions for big data analytics application using partial least squares structural equation modelling, *J. Big Data*, vol. 8, no. 1, pp. 1–15, 2021.

[48] S. Saif and S. Wazir, Performance analysis of big data and cloud computing techniques: A survey, *Procedia Comput. Sci.*, vol. 132, pp. 118–127, 2018.

[49] S. M. Shamsuddin and S. Hasan, Data science vs. big data @ UTM big data centre, in *Proc. of 2015 IEEE Int. Conf. Science in Information Technology*, Yogyakarta, Indonesia, 2015, pp. 1–4.

[50] T. Y. Yang and Y. Zhao, Application of cloud computing in biomedicine big data analysis cloud computing in big data, in *Proc. of the 2017 Int. Conf. Algorithms, Methodology, Models and Applications in Emerging Technologies* (*ICAMMAET*), Chennai, India, 2017, pp. 1–3.

Amanpreet Kaur Sandhu received the PhD degree from IKG Punjab Technical University, India in 2018. She is currently an assistant professor at University Institute of Computing, Chandigarh University, Mohali, India. Her research interests include image and video processing and Big Data analytics.