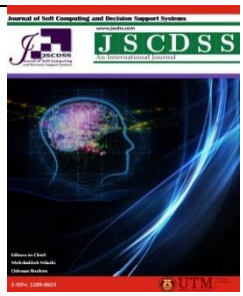
An International Journal
JSCDSS
http://www.jscdss.com
Vol.6 No.6 December 2019: 14-20
Article history:
Accepted 13 November 2019
Published online 13 November 2019

Journal of Soft Computing and Decision Support Systems



E-ISSN: 2289-8603

Big Data Tools: Advantages and Disadvantages

Maria Ijaz Baig [a,*], Liyana Shuib [a,*], Elaheh Yadegaridehkordi [a,*]

[a] Department of Information Systems, Faculty of Computer Science & Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia

*** Corresponding authors email address**: mariaijazbaig@hotmail.com, liyanashuib@um.edu.my, yellahe@gmail.com

Abstract

Big data tools have increasingly become crucial requirements of managing the complex and voluminous data. The selection of right tool requires an in-depth knowledge of existing big data tools and their potentiality. This paper provides a review of big data tools by selecting 34 articles from 2011 to 2018. This study provides pertinent information about most popular big data tools. The various big data tools related advantages and disadvantages are also discussed in detail. The findings of this study categorized the big data tools according to their potentiality. This study is beneficial for researchers to explore the big data sets according to its potentiality. It also provides deep insight of big data tools applicability to apply in real-time environment. This research is also helpful for practitioners to select the right big data tool according to requirement.

Keywords: Big data tools, Big data advantages and disadvantages, Big data potentiality

1. Introduction

The modern world enormously generated big amount of data. This big data is producing from variety of sources. The size of data is increasing day by day and recently reached to several xenottabyte. Big data is known by three characteristics, referred as the 3V's (Volume, Velocity and Variety). Volume refers to the size of the data. Velocity refers to the speed at which data is accessible (streaming of data). Variety refers to the type of the data (Zhang et al., 2018). Meanwhile, big data has different forms of structured (e.g., database and oracle), unstructured (e.g., videos, audio, images), and semi-structured data (e.g., commerce) (Shukla, Radadiya and Akotiya, 2015). The big data characteristics are significant to understand the nature of data and use the available tools accordingly. The existing tools have potential to handle the characteristics of big data (Landset et al., 2015; Zhang et al., 2018). Now, many tools such as Hadoop, HDFS, MapReduce, YARN, Hbase, Hive, Mahout, and Zookeeper have appeared out to manage the large data (Simovic, 2018). However, these tools application datasets matched with the characteristics of big data (Oussous et al., 2018). Conversely, each of the tool has advantages and disadvantages. Currently, a lot of research in big data has been on revolution, significance, analyzing the factors related to acceptance of technology (Mayer and Cukier, 2013; Jin et al., 2015; Marz and Warren, 2015; Zhang et al., 2018). However, none of the study focuses on big data potentiality and analyzes the advantages and disadvantages of big data tools accordingly. Therefore, this paper aims to present a short review of recent research on big data with a specific focus on big data potentiality, tools, and related advantages and disadvantages. The main research questions of this study are:

1) What are the most popular big data potentiality and tools?
2) What are the advantages and disadvantages of big data tools?

The rest of the paper is structured as follows. The research methodology is presented in Section 2. Section 3 discusses the findings. Conclusion is given in the Section 4.

2. Methodology

The aim of this study is to find out the big data tools and its advantages and disadvantages. For this purpose, initially various databases were explored to find out the relevant studies. During first search stage (S1), it has been observed that majority of studies were exist in IEEE Xplore, Emerald Insight, Scopus Elsevier, Springer Link and ACM. Additionally, keywords were searched to find as many articles. The key strings of 'big data tools', combine with advantages and disadvantages were explored. The inclusion and exclusion criteria were also considered for final selection of studies. This study considered the articles that published in English language from January 2011 to

December 2018. Only complete articles were included while summaries and discussions were excluded. Table 1 shows a summary of inclusion and exclusion criteria.

Table 1
Inclusion and exclusion criteria

Inclusion	Exclusion
Full text	Full text not available to download
Relevant published between January 2011 and December 2018	Not relevant to research questions
Written in English language	Not in English language
Relevant with topic	Not related with topic

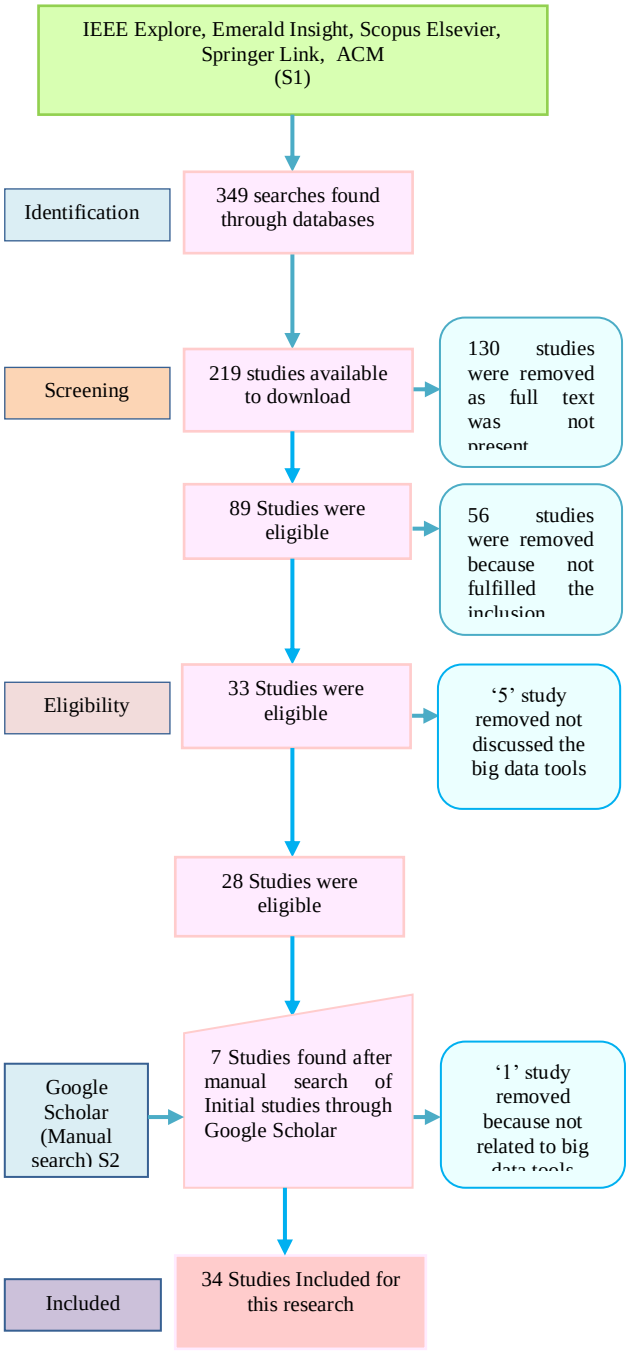


Fig. 1. Selection strategy

The article selection strategy is shown in the Fig. 1. After initial search, 349 studies were found. 219 studies were available to download. However, 130 studies were removed as full text was not present. All 89 studies were downloaded and studied. The inclusion and exclusion criteria were used for final selection of studies. 56 studies removed as these were not fulfilled the selection criteria. In total, 33 eligible studies were rechecked, and 5 studies removed as it was not describing the big data tools. In second phase of search (S2), Snowball technique was applied on all initially searched studies references. The 7 studies were found through Google Scholar. However, 1 study removed because it was not related to big data tools. Therefore, finally 34 studies were included for this study. The selection of studies was from: IEEE Explore (13), Emerald Insight (1), Scopus Elsevier (5), Springer Link (4), ACM (5), and Google scholar (6). Table 2 presents the selection of studies.

Table 2
Selection of studies

Databases	Initial	Final
IEEE Xplore	104	13
Emerald Insight	19	1
Scopus Elsevier	43	5
Springer Link	86	4
ACM	97	5
Google scholar (second search stage)	-	6
Total	349 (Initial searched studies found through databases)	34 (Final selected studies)

3. Findings

Big data potentiality is an important to organize the various operations accurately. In order to handle the power of big data, storage, processing and accessibility is required to manage the large volume of data. Data storage refers to collection of large data sets. HDFS and Hbase tools are more flexible and consistent for storage of data. However, HDFS is mostly used with relational databases. Data processing is a paradigm that sends data to system where the data can reside. Map Reduce, Hadoop and YARN are the tools that can process the big data more efficiently. Data access refers to easy accessibility of stored data. The huge amount of data is easily accessible through Pig, Hive, Cassandra, Mahout and Jaql. Data Management refers to maintain the workflow and organize the resources to control the system. Zookeeper and Oozie are the most reliable tools to manage big data.

3.1 Hadoop

Hadoop was developed by Doug and mike in 2005. It is the most important Apache open-source distributed tool for big data. Primarily, it was developed for simple functions such as web search indexes. NoSQL, CouchDB and MongoDB also belong to same class. Hadoop is used to process

An International Journal
JSCDSS
E-ISSN: 2289-8603

enormously large amount of data with varying or no structure. It is also suitable to conclude results of large problems having different parts or many servers' nodes. Hadoop has potential to bind the data that is difficult to analyze. It can be used as a tool or a data organizer. Most of the social networking websites like Google, Facebook and Yahoo are using Hadoop. The major limitations of Hadoop are (1) it is difficult to install, organize, and administer, and (2) it needs skillful staff to adopt Hadoop completely. Hadoop is further consisting of two modules:

3.1.1 The Hadoop Distributed File System (HDFS)

It is a consistent and fault tolerant system based on java based distributed file structure. It is basically designed to store large Volume of data and to ensure data storage for the Hadoop if host failure is expected. Its main structure consists of blocks called master and slave. Master is a name node which stores Meta data (place of data) while, slave is a data node. Block main function is to read/write and its default size is 128 Mb. HDFS store files and distribute these files into several blocks known as chunks. These chunks can work independently. For instance, if file size is 90 MB then blocks only utilize 90 MB memory out of 128 Mb. Each Hadoop cluster has several data nodes but single name node. The major advantage of HDFS is high bandwidth, cost effectiveness and write once and read data infinite times (Table 1 in Appendix A). The down sides of HDFS are cluster managing is difficult and join operation is time consuming process.

3.1.2 MapReduce

Map Reduce consists of two Functions Map and Reduce. Mapper function takes aggregate data and distribute into various keys and values. Whereas, Reducer function merges the data after getting input from Map function. Both functions work together to produce final output. It is basically developed by Google which supports Hadoop platform and Java language. Map Reduce plays a role of bridge for the allocation of subtasks and assembling of aggregated output. It basically consists of task tracker and job tracker node. The job tracker plays a role of Master and distributes resources. However, task tracker act as a slave and execute the scheduled job assigned by Master (job tracker). The status of job completed is send by slave. If mission of execution accomplished by task tracker then job tracker sends another task. The major disadvantages of map reduce is that: it's only workable for batch-oriented processes. To manage the complexity of Hadoop tool, many applications developed and incorporated with it. For instance, Jaql (query language for JSON), Hive (developed by face book), and Pig (developed by yahoo) etc. Apache Zookeeper and Cassandra is an open-source project used to maintain the large cluster and provide distributed synchronization.

3.2 Hbase (NoSQL Database)

It is an open source database. It has several versions and can be used on HDFS or on local file systems. Hbase structure is based on Bigtable and it was originally developed by Google. Its basic purpose is to provide more easiness in storing and organizing the big data. It provides support to structured data. Hbase play a role of storage layer in Hadoop. Storage wise it also supports HDFS by incorporating MapReduce tool to execute the data. The main advantages are: It is consistent, highly flexible and have proper backup system. Through Hbase Insert, delete, update operations can be performed easily. The disadvantages of Hbase are: It cannot be suitable for complicated data for example, transactional data.

3.3 Pig (Programming Tool)

Pig was introduced by Yahoo in 2006. It was developed for Big data analysis, evaluation and assimilation of structured, unstructured and semi- structured data types. It consists of two modules Pig Latin, and the runtime model. Runtime version carried out the Pig Latin code. Initially Pig loads the data into HDFS. After that mapper and reducer tasks transform the data. Finally, program display output on the screen or it stored at a location. The major advantage of pig language is that it reduces the volume of the data. It reduces the total code size. For instance, 20 lines of pigLatin code is equivalent to 400 lines of java code. Pig is flexible, cost effective, easy for read/write operations, and saves development time. It is very much resembling with SQL. Therefore, it is easy to learn for SQL programmers.

3.4 Hive

Pig is much easier and flexible to use but still required learning effort. Therefore, programmers need to become skillful before using this language effectively. To handle this issue Jeff Hammerbacher introduced Hive under Facebook working platform. Facebook team realized that large amount of data is produced through their social site. This data should be extracted and analyzed in a proper way. Hive support queries which is called HiveQL. It supports all data types and compiled the HQL queries into map reduce task. The key advantage of Hive is that it easily extracts, store, retrieve and converts data. SQL programmer can write HiveQL without any difficulty. HiveQL needs coding lines as compared to Map reduce based programs. The major disadvantages are it cannot process unstructured data. By using Hive difficult tasks cannot be performed.

3.5 Hadoop YARN (Yet another Resource Negotiator)

YARN is the central resource that was by used Hadoop. It monitors the cluster nodes and all related processing operations. YARN allows Hadoop to perform operational activities without waiting batch tasks to finish. It has four core components (1) ResourceManager, (2) NodeManager,

An International Journal
JSCDSS
E-ISSN: 2289-8603

(3) ApplicationMaster and (4) Container. Resource Manager consists of scheduler and application manager. Scheduler handles assigning resources and scheduler manages the resource allocation. NodeManager analyze total usage, check the resources and execute the container application. It is also responsible for reporting all information to the scheduler. Application Master can replace all containers in case of error to maintain the continuity of the process. Container helps to improve the efficiency. Application master can select any number of containers.

3.6 Cassandra

It is an Apache project, built on a distributed database system. Cassandra has NoSQL system and it is developed by face book to support inbox search. It has 2 million columns in a single row. It is highly robust, accessible, scalable, and reliable and has zero point of failure due to its dynamo model. It is also suitable for transactions because of its ACID property. It consists of several interconnected servers. If one node stuck, then another node will complete the task. Cassandra is used by Amazon, Face book, Twitter, and rack space. For Amazon, it helps to track customer shopping carts history. Cassandra is not workable where the use of sub query, join operation, and data aggregation operations are needed.

3.7 Oozie

It's an Apache open source project used to manage the workflow and synchronization between the tasks. It allows developers to describe job and their relationships. Once the relationship standard has been defined, it automatically plans the execution of job. Oozie have two types of jobs workflow and coordination. Workflow managed the hadoop jobs followed by Directed Acyclical Graphs (DAGs) of actions. Coordinator jobs are activated by time and availability of data. Oozie is highly scalable, reliable and flexible. But it does not work for off grid type scheduling.

3.8 Mahout

Mahout is another Apache open source project. Its major aim is to support the big data tool while Hadoop is used. It is still under processing project and various algorithms are including in it to enhance overall efficiency. It is widely used for collaborative filtering, clustering and classification applications. It provides free applications for data mining and machine learning methods. Mahout is used by Twitter, Face book and LinkedIn for data mining purpose.

3.9 Jaql

It Jaql is a query language for JSON (Java Script Object Notation). It was introduced by IBM to handle both structured and unstructured data. Jaql can perform read, write, create and run queries operations easily. It can select, join, group, and delete the data situated in HDFS.

3.10 Zookeeper

Zookeeper is Apache open source project. It is simple, reliable offers atomicity, synchronization, ensures the availability of data and owned powerful features. It synchronizes with multiple nodes of cluster and sends pattern traits to nodes. It handles the requests in queue and guarantees the release of message. However, multiple stack maintenance is needed.

4. Conclusion

Big data tools tools radically changed the data analysis platforms. This study provide overview of big data tools and their associated advantages and disadvantages. The studies were extracted through IEEE Explore, Emerald Insight, Scopus Elsevier, Springer Link, ACM and Google Scholar. Finally, 34 articles published between the year 2011 to 2018 (December) were utilized to extort germane information to address the research questions. Finding of this study concluded that Hadoop Distributed File System (HDFS), Hbase, MapReduce Tool, Hadoop, YARN, Pig, Hive, Cassandra, Mahout, Jaql, Zookeeper, and Oozie are the most popular big data tools. Furthermore, Hadoop distributed file system and Hbase were the most scable and flexible for data storage. But, it was not feasible for complex join operations. Mapreduce, Hadoop and YARN were highly capable to process large volume of data. However, it was difficult to administered. Pig, Hive, Cassandra, Mahout and Jaql provided smooth accessibility of large data. These were suitable for structured data. Zookeeper and Oozie were the most reliable to manage data. However, it lacked the multiple stack maintenance feature.

An International Journal
JSCDSS
E-ISSN: 2289-8603

Appendix A

Table 1
Big data potentiality, tools, advantages, and disadvantages

Big Data Potentiality	Tools	Advantages	Disadvantages	References
Data Storage	Hadoop Distributed File System (HDFS)	High bandwidth to support other tools. Highly scalable and cost effective. Write once and read data many times.	Cluster managing is difficult. Join operation is slow.	Lee, Kang and Lee (2011); Patel, Birla and Nair (2012); Katal, Wazid and Goudar (2013); Liu, Iftikhar and Xie (2014) ; Bende and Shedge (2016); Zaharia et al. (2016)
	Hbase	Highly flexible, consistent, and fault tolerant.	Not suitable for complicated operations like joins.	Bakshi (2012); Chandarana and Vijayalakshmi (2014); Hashem et al. (2015)
Data Processing	MapReduce	Support Java language. Process Independently.	Use only for batch-oriented processes.	Moon, Lee and Kee (2014); Fontugne, Mazel and Fukuda (2014); Simovic (2018)
	Hadoop	Can process the huge volume of data easily.	Difficult to install, organize, and administer. Organizations lack skillful staff to handle. Hadoop completely.	Mukherjee et al. (2013); Odriscoll, Daugelaite and Sleator (2013) ; Oancea and Dragoescu (2014)
	YARN	Efficiently maintain the resources, continuity and scalability of the process.		Ranjan (2014); Manogaran et al. (2017)
Data Access	Pig	Ensures the originality of data by decreasing replication and coding lines. Easy for read/write operations.	Lack web interface. JDBC and ODBC network connectivity is absent.	Herodotou et al. (2011); Shoro and Soomro (2015)
	Hive	Data accessibility, transformation, loading, querying, and extraction are much easier. Directly extract the data instead of writing jobs into Map reduce program. Can be incorporated with Hbase.	Not support unstructured data and complicated jobs.	Dhyani and Barthwal (2014); Marchal et al. (2014); Bhardwaj et al. (2015)
	Cassandra	High throughput and efficient response time. Support ACID property.	Not supportive for join operation and sub queries. Limited storage space.	Abramova and Berbardino (2013); Chebotko, Kashlev and Lu (2015)
	Mahout	Supports different data mining patterns and huge volume of data.	Decision tree algorithm is absent.	Condie et al. (2013); Singh et al. (2014); Verma, Patel and Patel (2015)
	Jaql	Support semi structured data and physical transparency.	Need consistent format while using select statement and transform operators.	Rathee (2013); Chen, Mao and Liu (2014); Gupta, Gupta and Mohania (2012)
Data Management	Zookeeper	Highly reliability offers atomicity, synchronization and ensures the availability of data.	Multiple stacks maintenance is needed.	Fan and Bifet (2013); Sharma and Navdeti (2014)
	Oozie	It supports execution of workflow in case of error or failure it can be restarted. Web service API is present.	Inappropriate for off grid development.	Islam et al. (2012); Loganathan et al. (2014); Oussous et al. (2018)

An International Journal JSCDSS E-ISSN: 2289-8603

References

Abramova, V., & Bernardino, J. (2013). NoSQL databases: MongoDB vs cassandra. In Proceedings of the international conference on computer science and software engineering (pp. 14-22). ACM.

Bakshi, K. (2012). Considerations for big data: Architecture and approach. In 2012 Aerospace Conference (pp. 1-7). IEEE.

Bende, S., & Shedge, R. (2016). Dealing with small files problem in hadoop distributed files system. Procedia ComputerScience, 79(1), 1001-1012.

Bhardwaj, A., Kumar, A., Narayan, Y., & Kumar, P. (2015). Big data emerging technologies: A Case Study with analyzing twitter data using apache hive. In 2015 2nd International Conference on Recent Advances in Engineering & Computational Sciences (RAECS) (pp. 1-6). IEEE.

Chandarana, P., & Vijayalakshmi, M. (2014). Big data analytics tools. In 2014 International Conference on Circuits,Systems, Communication and Information Technology Applications (CSCITA) (pp. 430-434). IEEE.

Chebotko, A., Kashlev, A., & Lu, S. (2015). A big data modeling methodology for Apache Cassandra. In International Congress on Big Data (pp. 238-245). IEEE.

Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. Mobile networks and applications, 19(2), 171-209.

Condie, T., Mineiro, P., Polyzotis, N., & Weimer, M. (2013). Machine learning on big data. In 29th International Conference on Data Engineering (ICDE) (pp. 1242-1244). IEEE.

Dhyani, B., & Barthwal, A. (2014). Big data analytics using Hadoop. International Journal of Computer Applications, 108(12), 265-270.

Fan, W., & Bifet, A. (2013). Mining big data: current status and forecast to the future. ACM sIGKDD Explorations Newsletter, 14(2), 1-5.

Fontugne, R., Mazel, J., & Fukuda, K. (2014). Hashdoop: A MapReduce tool for network anomaly detection. In conference on computer communications workshops (INFOCOM WKSHPS) (pp. 494-499). IEEE.

Gupta, R., Gupta, H., & Mohania, M. (2012). Cloud computing and big data analytics: what is new from databases perspective. In International Conference on Big Data Analytics (pp. 42-61). Springer, Berlin, Heidelberg.

Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (2015). The rise of "big data" on cloud computing: Review and open research issues. Information systems Elsevier, 47, 98-115.

Herodotou, H., Lim, H., Luo, G., Borisov, N., Dong, L., Cetin, F. B., & Babu, S. (2011). Starfish: A Self-tuning System for Big Data Analytics. In Cidr, 11(2), 261-272.

Islam, M., Huang, A. K., Battisha, M., Chiang, M., Srinivasan, S., Peters, C., & Abdelnur, A.(2012). Oozie: Towards a scalable workflow management system for hadoop. In Proceedings of the 1st ACM SIGMOD Workshop on Scalable Workflow Execution Engines and Technologies (p. 4). ACM.

Jin, X., Wah, B. W., Cheng, X., & Wang, Y. (2015). Significance and challenges of big data research. Big Data Research, 2(2), 59-64.

Katal, A., Wazid, M., & Goudar, R. H. (2013). Big data: issues, challenges, tools and good practices. In Sixth international conference on contemporary computing (IC3) (pp. 404-409). IEEE.

Landset, S., Khoshgoftaar, T. M., Richter, A. N., & Hasanin, T. (2015). A survey of open source tools for machine learning with big data in the Hadoop ecosystem. Journal of Big Data, 2(1), 24.

Lee, Y., Kang, W., & Lee, Y. (2011). A Hadoop-based packet trace processing tool. In International Workshop on Traffic Monitoring and Analysis (pp. 51-63). Springer, Berlin, Heidelberg.

Liu, X., Iftikhar, N., & Xie, X. (2014). Survey of real-time processing systems for big data. In Proceedings of the 18th International Database Engineering & Applications Symposium (pp. 356-361). ACM.

Loganathan, A., Sinha, A., Muthuramakrishnan, V., & Natarajan, S. (2014). A systematic approach to Big Data. International Journal of Information & Computation Technology, 4(09), 869-878.

Manogaran, G., Thota, C., Lopez, D., Vijayakumar, V., Abbas, K. M., & Sundarsekar, R. (2017). Big data knowledge system in healthcare. In Internet of things and big data technologies for next generation healthcare (pp. 133-157). SpringerLink.

Marchal, S., Jiang, X., State, R., & Engel, T. (2014). A big data architecture for large scale security monitoring. In International Congress on Big Data (pp. 56-63). IEEE.

Marz, N., & Warren, J., (2015). Big Data: Principles and best practices of scalable real-time data systems, 1st ed., (pp. 59-79). Manning Publications, America.

Mayer-Schönberger, V., & Cukier, K., (2013). Big data: A revolution that will transform how we live, work, and think, 1st ed., Houghton Mifflin, New Zealand.

Moon, S., Lee, J., & Kee, Y. S. (2014). Introducing ssds to the hadoop mapreduce tool. In 7th International Conference on Cloud Computing (pp. 272-279). IEEE.

Mukherjee, A., Datta, J., Jorapur, R., Singhvi, R., Haloi, S., & Akram, W. (2012). Shared disk big data analytics with apache hadoop. In 19th International Conference on High Performance Computing (pp. 1-6). IEEE.

Odriscoll, A., Daugelaite, J., & Sleator, R. D. (2013). 'Big data', Hadoop and cloud computing in genomics. Journal of biomedical informatics, 46(5), 774-781.

Oancea, B., & Dragoescu, R. M. (2014). Integrating R and hadoop for big data analysis. arXiv preprint arXiv:1407.4908.

Oussous, A., Benjelloun, F. Z., Lahcen, A. A., & Belfkih, S. (2018). Big Data technologies: A survey. Journal of

An International Journal
JSCDSS
E-ISSN: 2289-8603

King Saud University-Computer and Information Sciences, 30(4), 431-448.

Patel, A. B., Birla, M., & Nair, U. (2012). Addressing big data problem using Hadoop and Map Reduce. In Nirma University International Conference on Engineering (NUiCONE) (pp. 1-5). IEEE.

Ranjan, R. (2014). Streaming big data processing in datacenter clouds. IEEE Cloud Computing, 21(1), 78-83.

Rathee, S. (2013). Big data and Hadoop with components like Flume, Pig, Hive and Jaql. In International conference on cloud, big data and trust (Vol. 15).

Sharma, P. P., & Navdeti, C. P. (2014). Securing big data hadoop: a review of security issues, threats and solution. Int. J. Comput. Sci. Inf. Technol, 5(2), 2126-2131.

Shoro, A. G., & Soomro, T. R. (2015). Big data analysis: Apache spark perspective. Global Journal of Computer Science and Technology, 14(4), 47-59.

Shukla, P., Radadiya, B., & Akotiya, K. (2015). An emerging trend of big data for high volume and varieties of data to search of agricultural data. Oriental journal of computer science & technology, 8(2), 121-129.

Simovic, A. (2018). A Big Data smart library recommender system for an educational institution. Library Hi Tech, 36(3), 498-523.

Singh, K., Guntuku, S. C., Thakur, A., & Hota, C. (2014). Big data analytics tool for peer-to-peer botnet detection using random forests. Information Sciences, 278(2), 488-497.

An International Journal
JSCDSS
E-ISSN: 2289-8603