Discussion on Big Data: TDFS Vs HDFS

C Bagath Basha, S Rajaprakash, S Karthick, Preesha louise, Amritha.S

Abstract**:** ***In recent years, big data is huge amount of data to uncover hidden attributes. Today's technologies has possible to analyze the data and get data is almost immediately. Why big data is very important? Because cost reduction, faster, and better decision making using Hadoop. For example a large warehouse of terabytes of data is generated daily from social media's like Twitter, LinkedIn and Facebook are case of organization in the people to people communication area for big data. Big data has 3 most important challenges of Volume, Variety, and Velocity. In this paper we have studied about the performance of Traditional Distributed File System (TDFS) and Hadoop Distributed File System (HDFS). Benefits of HDFS has support for flume tool in Hadoop comparing with TDFS. Memory block size data retrieving time and security are used as metrics in evaluating the performance of TDFS and HDFS. Result shows HDFC performs better than TDFS in the above metrics and HDFS is more suitable for big data analysis comparing of TDFS.***

Keywords **:** ***Big data, Hadoop, HDFS, TDFS.***

I. INTRODUCTION

Data generated online in last decade is very high compared to previous decade. This is due increase in internet users. Traditional technologies are not effective in process of huge amount of data due to following characteristics such as Volume, Velocity, and Variety. Volume: A huge amount of data generated continuously from social media's, sensors, government, etc [4]. Velocity is data generated rapidly and it could be process immediately to extract useful information data [4][13]. The social medias has every one hour data could be generated more than 2.5 Petabytes from customer transactions. Example of velocity is Youtube [13].Variety of characteristic is structured, semi structured and unstructured datas. Data has multiple sources such as text, videos, twitter data, etc...[4][13].

Manuscript published on November 30, 2019.

* Correspondence Author

Mr.C.Bagath Basha*, Research Scholar, CSE, Vinayaka Mission's Research Foundation, Salem ,India. Email: chan.bagath@gmail.com

Dr.S.Rajaprakash, CSE, Aarupadai Veedu Institute of Technology,Vinayaka Mission's Research Foundation, Chennai, India. Email: srajaprakash_04@yahoo.com

Mr. K Karthick, CSE Aarupadai Veedu Institute of Technology,Vinayaka Mission's Research Foundation, Chennai, India. Email: kkarthik@avit.ac.in

.Preesha louise Final year CSE, Aarupadai Veedu Institute of Technology, Vinayaka Missions Research Foundation Chennai, India

Amritha.S Final year CSE, Aarupadai Veedu Institute of Technology, Vinayaka Missions Research Foundation Chennai, India

II. BIG DATA HISTORY

Figure 1 shows the history of big data. In 1960's storage capacity in a database is high but proceesing speed is low [2]. In 1970's storage capacity has increase from megabytes to gigabytes and processing speed also increased compared to previous decade [2]. In 1980's storage capacity in a parallel database has increase from gigabytes to terabytes and data processing is high compared to last year's [2]. In 1990's storage capacity in google file system is increase from terabytes to petabytes and data processing speed also high compared to last decades [2]. In 2011 storage capacity has rapidly increased in a database from petabytes to exabytes and data processing speed is very high compared to previous years [2]. In 2018, Currently we have very large storage capacity in a database from exabytes to zetabytes and data processing is very high compared to last

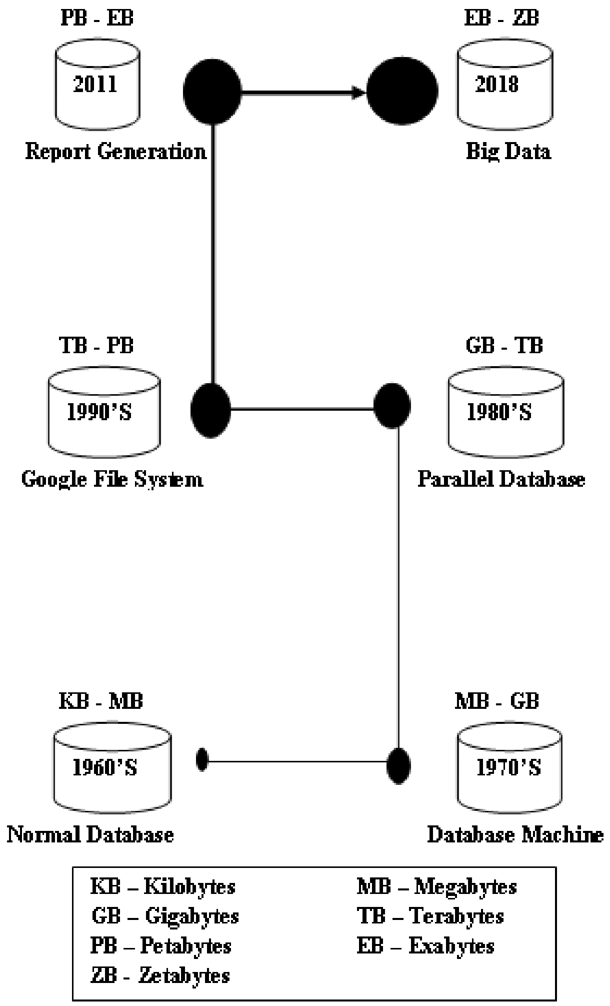


Fig. 1. History of big data

decades.

Lydia et al., (sep – 2016) " A Literature Inspection on Bigf Data Analytics". Author Ahmed Oussous said 2020 the data is generated by internet will 400 times by now. The data generated reaches zeta bytes [13].

III. LITERATURE SURVEY

Chun-Wei Tsai et al.., (2015) "Big data analytics: a survey". In this paper discuss about Knowledge Discovery in Database (KDD) process are used three parts i.e. input, analysis and output. Authors find the useful thinks in big data, and then they have to develop and design the appropriate mining algorithms. Data mining is used to improve performance and result oriented by data analytics of KDD. Finally, they show the quality of input and output data, security & privacy of data in KDD.

Lydia et al., (2016) proposed "A Literature Inspection on Big Data Analytics". In this paper discuss as a four stages, first stage is history of big data, second stage is how to store a big data using Hadoop Distributed File System (HDFS) in Hadoop and how to process a big data using MapReduce in hadoop, Third stage is security of big data in cloud computing and its creating a personal computer security, system security, data security and information security, Fourth stage is big data value chain such as generation of data, acquisition of data, storage of data, and data analysis. Finally authors have several big data technologies and techniques used to solve analyse big data.

Lydia et al., (2016) proposed "Processing Image Files using Sequence File in Hadoop". In this work a big amount of data is divided into separate files used MapReduce algorithm in hadoop. Here the proposed algorithm removed unwanted and repeated files from database. Data's are compressed and, so it cannot be split a data. A single file can be takes as an input file to a MapReduce job, and mapper can be used to process a file. MD 5 algorithm is used to improve the image quality by removing unwanted and repeated files.

Acharjya et al..,(2016) "A Survey on Big data Analytics: Challenges, Open Research Issues and Tools". In this paper first understand the stages of big data for separately and also specific functionality. Authors mainly used tools for processing of big data analysis such as statistical analysis, intelligent analysis, cloud computing, quantum computing, data stream process, machine learning and data mining.

Sara Landse et.al.,(2015) proposed "A survey of open source tools for machine learning with big data in the Hadoop ecosystem". Author discussed as machine learning tools with distributed and real time process. Authors proposed algorithms on classification, collaborative filtering, deep learning & regression. Classification and Regressions algorithms are based on decision tree, Logistic regression, Navie Bayes, Support vector machine, Random forest, and Linear regression. Clustering algorithms are based on K-means, Fuzzy K-means, Streaming K-means, Power iteration, and Spectral clustering. Collaborative filtering (cf) algorithms are User based cf, Item based cf, and alternative least squares. A processing engine has MapReduce, Spark, Storm, Flink, H2O and machine learning frameworks. The frameworks are evaluate using memory processing, low latency, fault tolerance as metrics.

Dilpreet sing et. al., (2014) proposed "A survey on platforms for big data analytics". In this work author discussed about performance of big data analytics and discussed software frameworks and hardware platforms along with advantage and disadvantages. Authors surveyed the different hardware platforms using several metrics like scalability, data input and output rate, fault tolerance, real time processing , data size supported and iterative task support in big data analytics. They used some characteristic of K-means clustering algorithm to understand various big data platforms. They compared the horizontal and vertical scaling of advantage and disadvantage.

Priyank jain at.al.,(2016) proposed "Big data privacy: a technological perspective and review". In this paper manly discussed privacy and security of data in big data. Authors take some existing methods to implement the business purpose like HYbrEx, K-anonymits, T-closenes & L-diversity and big data life cycle. They refer healthcare for privacy and security of data in big data. User information should be poor privacy and good security practices. Data could be confidential, integrity and availability of security data.

Zaheer khan at.al., (2015) proposed "Towards cloud based big data analytics for smart future cities". Author proposed a cloud based analytics service for big data analytics & management and used some big data analytics techniques for analytical engine such as data pre-processing and integration, supervised & unsupervised learning, visualisation, reduction data and finding association rules. They implement and compare results hadoop and spark through machine learning or statistical analytical techniques in data mining. They mainly focused on quality of life such as crime & safety and economy over the year to asses positive and negative trends.

Joseph issa (2016) proposed "Performance characterization and analysis for Hadoop K-means iteration". In this paper author discussed performance of software and hardware. Authors find computing processing time is very high or low and compare performance using Intel and AMD processor for hadoop K-means by modelling different processor. They predict the performance with less than 5% error margin relative to a measured baseline.

Faiz et. al., (2017) proposed "Source camera identification: a distributed computing approach using Hadoop". In this work an increase performance of process by implement the distributed computing environment. They collect the 6000 images from various mobiles phones for processing by image classification used Apache Mahout's Random Forest Classifier. Authors build a prediction model used Mahout's Random Forest algorithm is scalable machine learning tool.

Michael Crawford et. al., (2015) proposed "Survey of review spam detection using machine learning techniques". In this paper mainly discuss and focussed to solve the problem of spam detection using Machine Learning Techniques. Machine learning techniques are supervised learning, unsupervised learning and semi-supervised learning. Supervised learning method have attribute of learning from a set of labelled data and require training data

Retrieval Number: D9497118419/2019©BEIESP
DOI:10.35940/ijrte.D9497.118419
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

. Unsupervised learning method have attributes are set of unlabeled data, relationship of data independence and form is clustering. Semi-supervised learning method have attributes are set of labelled & unlabeled data, require small set of labelled data and huge amount of unlabeled data.

They spammers manipulate or poison reviews such as making fake review, untruth review and deceptive reviews for profit or gain because all online reviews are not truthful.

IV. HADOOP

Hadoop is an open source framework which has been apache foundation. Hadoop is a storing and processing of huge amount of data with cluster of cheap hardware. Hadoop has two main components HDFS and MapReduce [2][14].

A. Hadoop Distributed File System (HDFS)

HDFS is a specially designed file system for storaging large amount a data with cluster of commodity hardware and with streaming access pattern. HDFS has default minimum block size is 64MB and maximum block size is 128MB in hadoop. HDFS has been given by default 3 replications of data such as original copy, extra copy, and duplicate copy [14].

HDFS has two services: Master Service and slave service. Master Service has three parts: Name node, Secondary Name node, and Job Tracker. Slave service has two parts: Data node and Task tracker [14]. Every master service communicate to each other and slave service communicate to each other. If name node is master node and its corresponding slave node is data node. If Job tracker is a master node and its corresponding slave node is Test tracker [14][2].

B. MapReduce

MapReduce is process on huge amount of dataset and it should be distributed into several machines and parallel process of data [12]. MapReduce has two computations: Map and Reduce [11]. Figure 2 shows the map reduce process. The mapper takes as input data for raw data. The raw data can be divided into several files. Each files should be mapping individual and shuffle the all files, merge the several files as Reduce output [1][2][14].

MapReduce algorithms are

1. Map
2. Merge
3. Reduce

1. Map (MK1, MV1) – MK2, MV2

2. Reduce (MK2, list(MV2)) – MV3

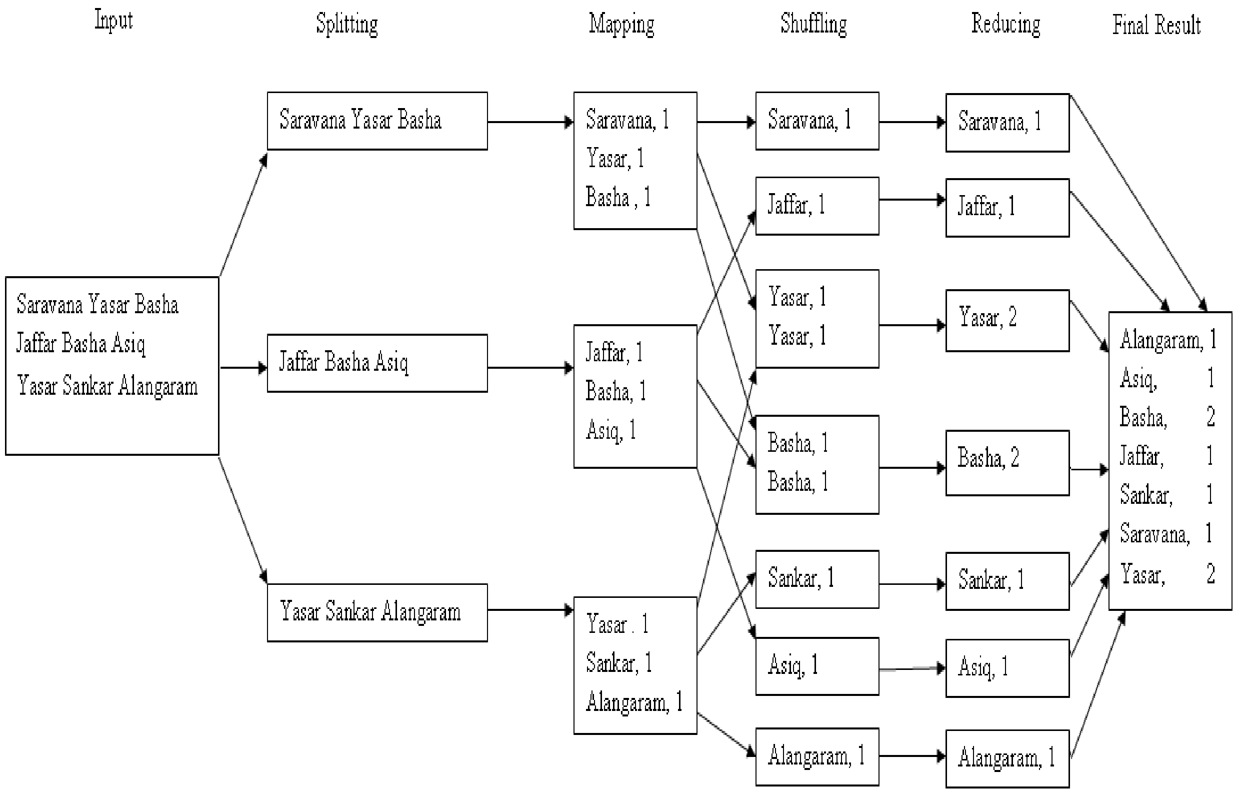


Fig. 2. Mapreduce

V. COMPARISON OF TDFS AND HDFS

HDFS client request to Name Node for storing a data and Name Node response to the client these systems are free for storing a data then, the client send a file to Data Node (DN) 1. Each Data Node could be copy and store multiple disks and

Data Node informed to Name Node. Every 10 seconds Data Node should be send the block report to the Name Node otherwise the data should be failed or damaged. If data damaged or loss we have back up of data as shown in below Figure 3 [14].

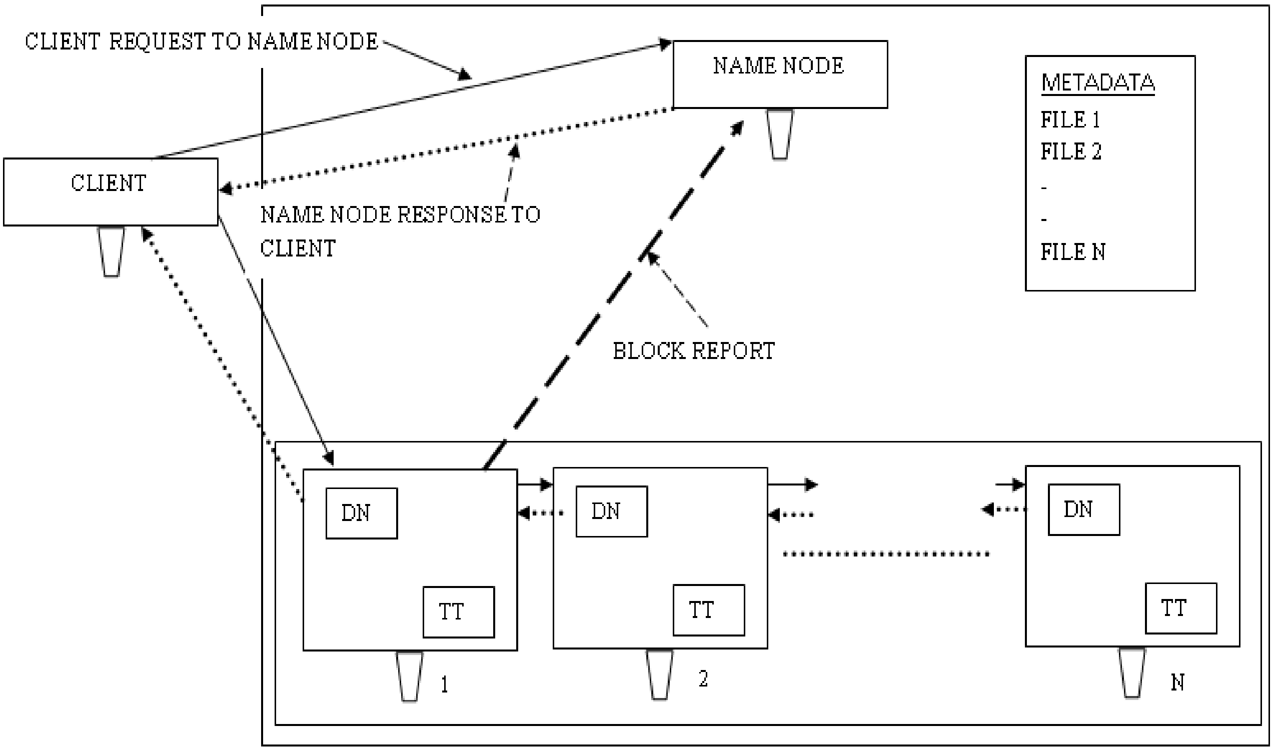


Fig. 3. HDFS Workflow

TDFS client request to server for storing a data and server response to the client these systems are free for storing a data then, the client send a file to Disk 1, Disk 2, ... Disk N. If data could be damaged or loss we did not have back up of data as shown in below Figure 4.

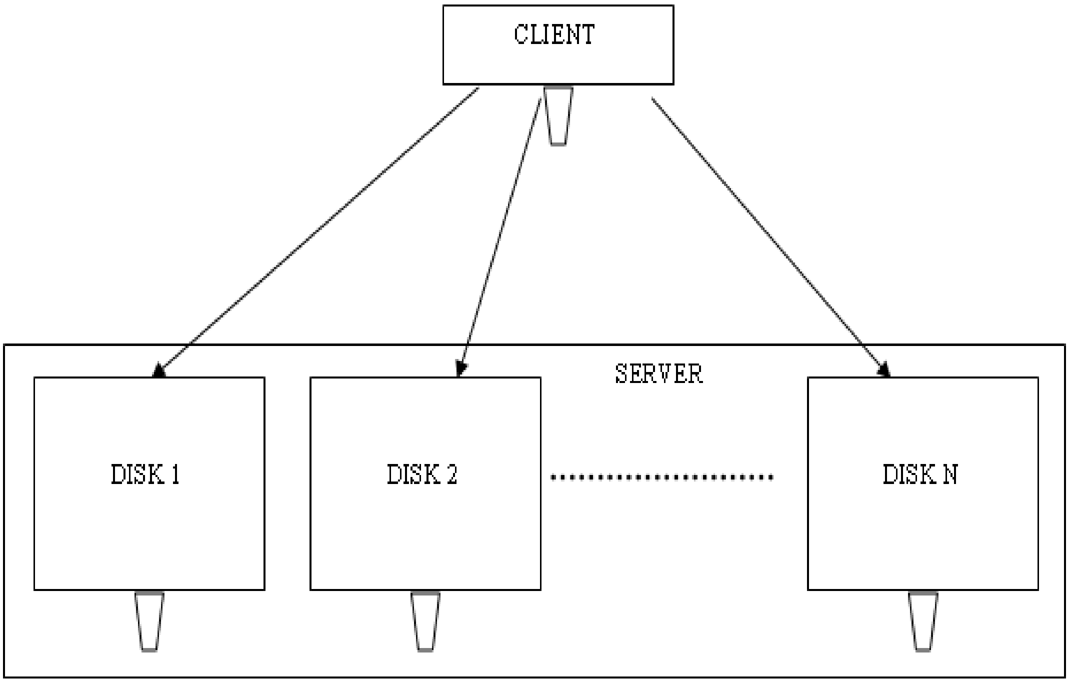
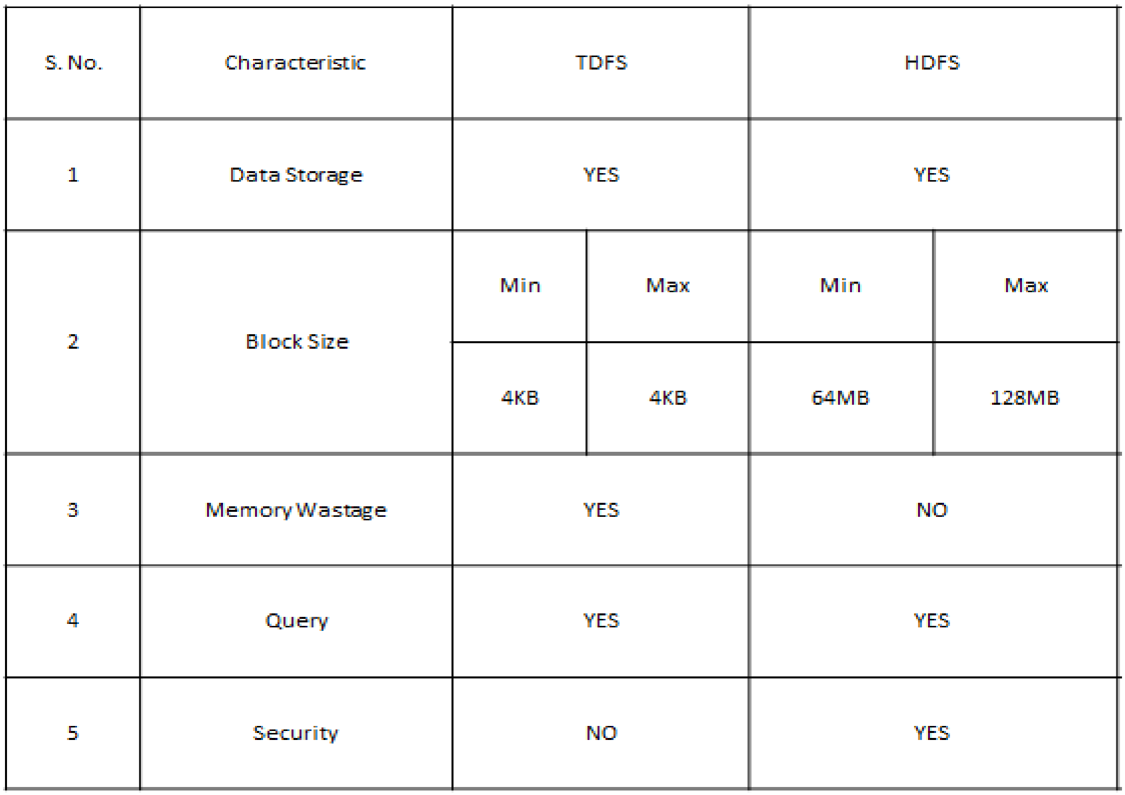


Fig. 4. TDFS Workflow

Table 1 show the comparison of TDFS and HDFS have multiple characteristics of hadoop. TDFS has possible to store a data and retrieve a data but no security of data. The minimum and maximum storing each block size is 4KB but wastage of memory. HDFS also has possible to store a data and retrieve a data with security of data. The minimum of each block size is 64 MB and maximum of each block size is 128 MB with memory reduced while comparing TDFS.

Table-1: Comparison of TDFS and HDFS



S. No.	Characteristic	TDFS		HDFS	
1	Data Storage	YES		YES	
2	Block Size	Min	Max	Min	Max
		4KB	4KB	64MB	128MB
3	Memory Wastage	YES		NO	
4	Query	YES		YES	
5	Security	NO		YES	

VI. CONCLUSION

In this paper we have analyzed performance of TDFS and HDFS. HDFS has flume tool in Hadoop for converting unstructured data to structured data which is not supported in TDFS. Memory block size, data retrieved time and security are used as metrics in evaluated the performance of TDFS and HDFS. Result shows HDFC performed better than TDFS in the above metrics. Therefore HDFS is more suitable for big data analysis compared of TDFS. Future work is to find supporting methods in Hadoop and compare to SPARK database.

REFERENCES

1. Chun-Wei Tsai, Chin-Feng Lai, Han-Chieh Chao, and Athanasios V. Vasilakos, "Big data analytics: a survey", Journal of Big Data (2015), pp.1-32.
2. Dr.E.Laxmi Lydia, M. Vijay Laxmi, Dr. M.Ben Swarup, "A Literature Inspection on Big Data Analytics", International Journal of Innovative Research in Engineering & Management (2016), ISSN: 2350-0557, Volume-3, Issue-5, pp. 422 – 430.
3. Dr. E. Laxmi Lydia, Dr. A. Krishna Mohan, Dr. M. Ben Swarup, "Processing Image Files Using Sequence File in Hadoop**",** International Journal of Engineering Sciences & Research (2016), ISSN: 2277-9655, pp.521-528.
4. D. P. Acharjya, Kauser Ahmed P, "A Survey on Big Data Analytics: Challenges, Open Research Issues and Tools", International Journal of Advanced Computer Science and Applications (2016), Vol. 7, No. 2, pp. 511-518.
5. Sara Landset, Taghi M. Khoshgoftaar, Aaron N. Richter and Tawfiq Hasanin, "A survey of open source tools for machine learning with big data in the Hadoop ecosystem", Journal of Big Data (2015), pp. 1–36.
6. Dilpreet Singh and Chandan K Reddy, "A survey on platforms for big data analytics", Journal of Big Data (2014), pp. 1–20.
7. Priyank Jain, Manasi Gyanchandani and Nilay Khare, "Big data privacy: a technological perspective and review", Journal of Big Data (2016), pp. 1–25.
8. Zaheer Khan, Ashiq Anjum, Kamran Soomro and Muhammad Atif Tahir, "Towards cloud based big data analytics for smart future cities", Journal of Cloud Computing: Advances, Systems and Applications (2015), pp. 1-11.
9. Joseph Issa "Performance characterization and analysis for Hadoop K-means iteration", Journal of Cloud Computing: Advances, Systems and Applications (2016), pp. 1-15.
10. Muhammad Faiz, Nor Badrul Anuar, Ainuddin Wahid Abdul Wahab, Shahaboddin Shamshirband and Anthony T. Chronopoulos, "Source camera identification: a distributed computing approach using Hadoop", Journal of Cloud Computing: Advances, Systems and Applications (2017), pp. 1-11.
11. Michael Crawford, Taghi M. Khoshgoftaar, Joseph D. Prusa, Aaron N. Richter and Hamzah Al Najada , "Survey of review spam detection using machine learning techniques", Journal of Big Data (2015), pp. 1–24.
12. E. Laxmi Lydia,Dr. M.Ben Swarup, "Big Data Analysis using Hadoop components like Flume, MapReduce, Pig and Hive",ISSN" 2231-0711, Vol 5, pp.390-394, November 2015.
13. Ahmed Oussous, Fatima-Zahra Benjelloun, Ayoub Ait Lahcen, Samir Belfkih, "Big Data technologies: A survey", Journal of King Saud University – Computer and Information Sciences (Production and hosting by Elsevier - 2017), pp. 1- 18.
14. https://www.youtube.com/watch?v=DLutRT6K2rM.

AUTHORS PROFILE

C. Bagath Basha is having teaching experience about 6 years and 6 months. He served in various positions in Teaching. He is currently doing as Research Scholar, Department of Computer Science and Engineering, Vinayaka Mission's Research Foundation, Salem, Tamil Nadu, India. His area of interest includes Big Data and Data Analytics, Security.

Dr.S.Rajaprakash M.E Ph.D. currently working as Associate professor of CSE in Aarupadai Veedu Institute of Technology an ambit institution of Vinayaka Missions Research Foundation (Deemed to be University), Tamil Nadu, India. He has 18 years of experience in academics, research, and development activities. Published 19 research papers in referred Journals and Conferences. His area of Interest Artificial Intelligence, Computational Intelligence, Discrete Mathematics and Automata theory. .Received grants from Tamil Nadu State Council for Science and Technology .He has peer Reviewed Manuscripts in reputed international Journals and Conferences. He is a member in following professional societies: CSI and ISTE and Ramanujam Mathematical Society.

Mr. K.Karthik ME (Ph.D) currently working as Assistant professor Aarupadai Veedu Institute of Technology an ambit institution of Vinayaka Missions Research Foundation (Deemed to be University), Tamil Nadu, India published more than 8 national and international journal and conference and organizing committee for 4 international conference,2 national conference and 15 years of teaching experience with 4 years of research experience. He is a member in following professional societies: CSI and ISTE.

Preesha louise Final year CSE, Aarupadai Veedu Institute of Technology, Vinayaka Missions Research Foundation Chennai, India

Amritha.S Final year CSE, Aarupadai Veedu Institute of Technology, Vinayaka Missions Research Foundation Chennai, India