# Big Data and Hadoop: A Review

LDSB Weerasinghe[1,#], B Hettige[1]

[1] Department of Computer Science, General Sir John Kotelawala Defence University, Sri Lanka

[#]sidath.weerasinghe@gmail.com

**Abstract—** *Most of the people use electronic data stored systems which are powered by new technology and share data across the other users. Therefore the data become unstructured and larger in size. This paper reviews on big data and Hadoop by considering security, data analysis, data storage methods and speed. Then figure what are the problems in big data and how these problems overcome using Hadoop with its architecture. Critically review the limitations of the Hadoop and find the solutions using various technologies such as multi-agent concept and machine learning concept. Big Data is mean that about a terabyte or the Zettabyte size of the file. Big data define by using four parameters, such as the scale of the data as volume, a different form of data as variance, analysis of streaming data as velocity and uncertainty of data as veracity. The people use Relational database management schemas (RDBMS) to store big data and faced a lot with difficulties such as very costly, only have fixed schema, difficult to save huge files, difficult to access files and take a lot of time to perform analytics. Hadoop is a framework which helps people to save that kind of data and do an analysis of that data. Hadoop using a distributed file system to save the big size of data and implementation of google map reduces algorithm for analysis big data on Hadoop. Relational databases deal with structured data, but the Hadoop deal with unstructured data. Hadoop is an open source data management system with no cost associated with it. This can design from the single server computer and can scale up to millions of computer servers using parallelism with a high degree of fault tolerance.*

***Keywords***— **Big Data, Hadoop, HDFS, Map Reduce**

## I. INTRODUCTION

In our day-to-day life most of the people use databases in their daily activities, nevertheless people are not aware on it. In terms of the database means a group of data or collection of data which has meaningfulness that stored in a computer related system or physical phenomena. Early day's people use log books to keep and store data. Papers, log books, files are a database for them. It is very difficult to manage. Then File-based systems come into the world. This file-based system has lots of disadvantages. Uncontrolled redundancy, data inconsistency, inflexibility, limited data sharing and poor enforcement of standards are some disadvantages of the file-based system. So people try to find an easier way to store and retrieve data. In this part discuss the computer related databases. People design a software related system to create or manage the databases is defined as a Database Management System (DBMS). The main functions of a DBMS is to provide efficient methods of data retrieval, reliable methods of data retrieval, make new databases, or add, delete or modify data and etc. MySQL, Oracle, Microsoft Access, MS-SQL Server are some of the examples for the DBMS. Data independence and efficient access, reduced application development time, data integrity and security, concurrent access, recovery from crashes and etc. are some advantages of the databases.

There are several components in DBMS. Such as physical database, file manager, database manager, query processor, DDL (Data Definition Language) and the data dictionary (Thompson et al., 2013). There are several types of DBMS build according to the three-schema architecture. Internal schema on the internal level to describe physical storage structures and access paths. Typically uses a physical data model. Conceptual schema at the conceptual level to describe the structure and constraints of the whole database for a community of users. External schemas at the external level to describe the various user views. Hierarchical DBMS, relational DBMS, network DBMS, and the object-oriented DBMS are the types of DBMS. In the world, most of the people use relational database management systems (RDBMS) (Robbins, 1995). RDBMS means a database that treats all of its data as a collection of relations. Attributes, entities and the relations are the key parts on RDBMS. Modeling ER-Diagrams and normalization technique used to develop databases. Structured Query Language (SQL) is a database language allow the user to create the database, relation structures perform data management tasks and perform queries.

To store different kind of media sources, such as photos, audio data, video data, graphs people use Object-oriented DBMS. This kind of DBMS are needed more cost to develop. This is the main disadvantage of Object-oriented DBMS. In databases data stored as a recodes and each record are connected to the other records. These are called hierarchical database. Network DBMS is making a relationship between the data in the databases via the network. Network DBMS contain the distributed databases. Most of the banking system uses this kind of databases. There are two types of distributed databases. They are homogeneous distributed databases and the

heterogeneous distributed databases. Data spread over multiple machines and network interconnects these machines so users shared data over multiple machines.

When designing the database system DBA should satisfy the ACID (Atomicity, Consistency, Isolation and Durability) properties (Korth and Sudarshan, 2001). Loss of confidentiality of data, loss of integrity on data, loss of availability in stored data and the privacy issues of the users are the problem faced by the DBA when maintaining a database system. Another most important factor is the security of the database. Modern database tool provides the security facilities for the user. Some of them are backup and the recovery, data encryption, RAID technology, authorization, access controls, views and integrity. Using sessions, proxy servers, firewalls, digital certificate and VPN can provide security for web based databases.

The rest of the paper is organized as follows. Section 2 describes big data, some related works on big data, problems of big data processing and solution for that. Section 3 how Hadoop works according to its architecture, limitations of Hadoop and how to overcome the limitations. After that last section summarize the review work.

## II. BIG DATA

Nowadays, nearly 3 billion people are connected to the internet and the amount of time they spend online about 35 billion hours a month in 2015 ("Why NoSQL?,") . And also Google, Amazon, Facebook, twitter and etc. like website sends millions of unstructured data to others at a time. These millions of people do transactions at the same time on the database. So that RDMS cannot handle such kind of data transaction because RDMS are structured databases. Such variety of information is called big data that depend on volume (size), variability (complexity), and velocity (rate of growth) of data set (Bhosale and Gadekar, 2014). Figure 1 shows the overview of big data.

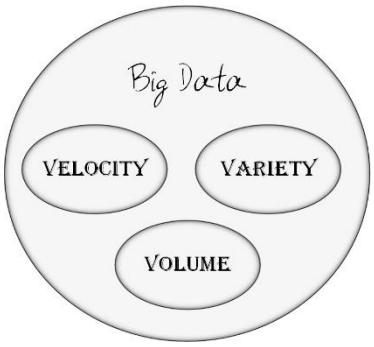


Figure 1 Big Data Overview

- Size of data: Volume (size) discusses to amount of data. A mass of information stored in enterprise databases has got from modest size to petabytes in size.
- Complexity of data: Various types of data and sources of data stored in the databases. Enterprise databases store structured, semi- structured and unstructured data. (Like video, pictures, accounts, files and etc.).
- Velocity of data: Speed of the data analyzing and processing refers to the data processing.

### A. *Problem With Big Data Processing*

In the world, most of the people are using a computer and people store and manage their data on it. When size of data and data become unstructured people cannot handle it correctly and also they cannot analyze data efficiently. For that computer engineers develop a powerful process to analyze such kind of data. But rapidly increasing volumes of data has been a challenge for them. Nowadays, data volumes are increasing rapidly than process developments (Rathi and Lohiya, 2014). So that person faces a lot of problems to handle data. If people handle data, then it takes a lot of time. Then people try to use multiple computers and make distributed environment to overcome this problem. In this approach the probability of failures rises is high. Always failures are an expected and they are common to that approach. People use natural languages to store data. So computers cannot identify these kind of data and that data is called unstructured data or incomplete data. The privacy is the biggest problem faced when handling big data. Managing privacy is effectively both a technical and a sociologically. The privacy level differs from user to user and date set to data set. It's really complex when petabyte of size data have on it.

### B. *Solution For Big Data Processing*

The solution for the big data processing is Hadoop. Apache Hadoop is an open source development framework. Hadoop is a comprehensive distributed batch processing infrastructure designed to efficiently distribute large amounts of jobs across a set of machines and developed by using Java language. Hadoop has two main components namely, google map reduce algorithm and Hadoop distributed file system. Other components such as Apache Hive, Base, Redis, Cassandra, Zookeeper and etc *(Shilpa and Kaur)*.

### C. *Some Related Works*

Big data analytic has generated a large volume of research publication within the last decade. According to research by Gupta and Dixti this field still suffers from many problems (Gupta and Dixit, 2015). Among other problems, Gupta and Dixti have specifically identified several problems including storing, analyzing, sharing, capturing, transferring, visualizing and searching on big data. Privacy violations is the big problem in data handling. With using

HDFS and Map Reduce algorithm researchers use Hadoop technology to overcome above problems. HDFS stores large volume of data and Map Reduce processes the data in a parallel manner. Finally discuss the advantages and disadvantages of Hadoop with its components which are HBase, Hive, Pig, Sqoop, Zookeeper, Avro, Oozie and etc.

In their research Dhomse, and others show that the Facebook and twitter like social web applications randomly generate Zettabytes of data in a day (Dhomse et al., 2015). Hadoop framework is used to handle such kind of huge amount of unstructured data. There are several modules based on the Apache Hadoop framework. They are Hadoop Common, Hadoop Distributed File System, Hadoop YARN and Hadoop Map Reduce. Hive is a data warehousing infrastructure which is used on top of the Hadoop for query optimization and analysis. The massive amount of data cannot be processed efficiently with the help of RDMS so people move to the techniques like Hadoop.

Nicolas and Carrera did a Systematic Study of Hadoop Deployment Variables and their called ALOJO (Poggi et al., 2014). This study presents about the mechanisms for an automated characterization of Hadoop deployments with low cost. IaaS and PaaS are included for the infrastructures deployed in both on cloud servers and premise physical. The performance of Hadoop depends on various hardware and software configuration. The best configuration Hadoop is the number of mappers to run parallel according to the available CPU cores and job types. ALOJA at present shows value to the Hadoop community by producing more knowledge and understanding of the underlying Hadoop runtime while it is executing.

*1) Hadoop Distributed File System (HDFS)*

A research is done by Karun and Chitharanjan, mainly discuss the Hadoop Distributed File System (HDFS) and how it's implemented (Karun and K, 2013). It is an open-source implementation of the Google File System and also it provides high throughput and full fault tolerance. Researcher figures out the advantages and disadvantages of Hadoop DB, Hadoop++, Co-Hadoop, Hail, Elastic replica management for HDFS, Trojan HDFS, Cheetah, RCFile, adaptive data replication mechanism (DARE) and Clydesdale (Mouliswaran and Sathyan). Several factors like fault tolerance, scalability, data locality, load balancing, performance, load time interface change to Map/Reduce, changes to Hadoop framework, indexing/layout and data compression are compared with above Hadoop infrastructure extensions (Shah et al., 2014). Hadoop increases the performances because of the layouts of the framework, strategic data partitioning/processing, replication and placement of data blocks.

Borkar and Surtakar did a critical review about Hadoop Distributed File System (Borkar and Surtakar, 2014). HDFS is one of the major components of Hadoop use to store bulk amount of data. Also, it provides the reliable and availability of data on the client side. Researchers discuss how the replicas are managed in HDFS for providing high availability and high computational of data. In Hadoop architecture sometimes HDFS is not utilized because due to scheduling delays. Hadoop develops using Java so that performance enhancing features in the native file system are not available. To overcome this limitation Hadoop bypass the cache and transfer data directly into the buffers. Hadoop is a power frame for large companies and it's dedicated to scalable, distributed, data –intensive and computing.

*2) Hadoop Map Reduce*

Kalra and lamba did a review about DDOS attack and defense methods for Hadoop, Hadoop cluster architecture, Hadoop distributed file system and job aware scheduling algorithms for map reduce framework (Kalra and lamba, 2014). Job Tracker, Task Tracker, Name Node and Data Node may include in the master node on the small Hadoop cluster. In large Hadoop cluster, this architecture is different. The difference is the cluster use servers for each node. The Job Tracker server can manage job scheduling like vice all nodes have a server and do special tasks. Map Reduce is designed for fast loading. Reviews compare the RDBMS and Hadoop approaches. There are four main ways to protect against DDoS attacks. They are attacking reaction, attack source identification, attack detection and attack prevention.

*3) Security*

Gupta and Jyoti did a research related to the Hadoop and attacks on enterprise data (Gupta and Jyoti, 2014). In that research discussed the big data analyzing technique of Hadoop and security of data for enterprises. The procedure of analyzing and mining large data is called big data analytics. The analysis is the process of analyzing huge data to discover hidden patterns, unknown relationships and uses extracted to make better decisions in the data set. Analysis of data become very difficult because of sophisticated targeted threats and fast development in data. Hadoop restrictions the volume of communication done by the individual processes. Records are processed in separately by isolated tasks on using map reduce algorithm. In big data security analytics reviews talk about problems with threat detection, Moore's law, open source and tons of activity on the supply side. Data security in the enterprises is an inspiring task to implement and calls for robust support in terms of security policy formulation and mechanisms.

*4) Hadoop with Agent-based technology*

The Sindhuja and others present about analyzed and compared service composition methodology using Hadoop framework and agents(Sindhuja et al., 2015). Researchers create two agents. Ones are consumer agent. This agent

process the user requirements. Another agent is service provider agent that clusters the output from consumer agent. Lastly, each user of the system is mapped to each producer. This system totally on the cloud and developed using JADE framework. Operations on the system are very fast, errors of agents do not affect the whole system, the system is reliable, and the greater response speed is some advantages of agent-based systems. Agents have set of features like autonomy, pro-activity, communication, cooperation, learning and negotiation. Agents are better than the Hadoop because of the less execution time for processing user request in agents.

Twardowski and Ry˙zko developed a Multi-agent architecture for real-time Big Data processing (Twardowski and Ryzko, 2014). In this paper discuss the big data processing using multi-agents' architecture. Consistency, availability and the partition tolerance properties are used to real-time processing big data systems instead of ACID properties. Very large sets of online and offline data independently handling on the real-time done by Lambda Architecture. This architecture consists of 3 layers. The batch layer can implement using Hadoop with map reduce algorithm. The batch layer is generally developed using Hadoop ecosystem. Serving layer used to compute the batch layer and facilitated by extra indexing of the data in order to speed up the reads of the system. Another layer is speed layer is used to compute in real-time the data. This research shows that self-governing agents can increase the architecture and offer capabilities for robust processing of data in real-time.

Essa, Attiya and Ayman introduced a new framework using a mobile agent for improving big data analysis (Essa et al., 2013). A new framework MRAM is developed using map reduce, a mobile agent and JADE and researcher are discussed about that in this paper. When increasing data productivity, people want to analyze these kind of data. Proposed framework is used to improve analysis technique of big data and to overcome the limitation of the Hadoop.

## III. HADOOP

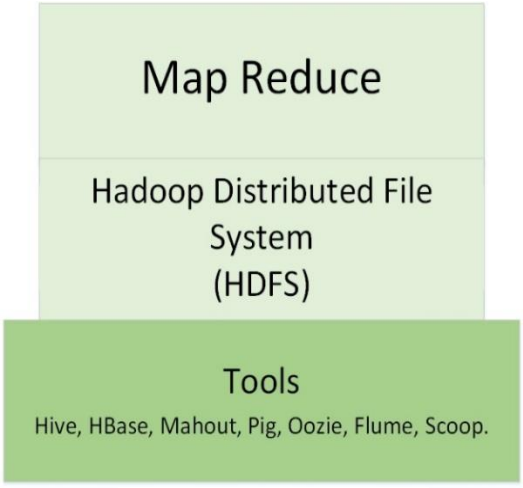


Figure 2. Hadoop Architecture

Hadoop proceeds a very different approach than the enterprise approach. It breaks up data into small pieces. That's why it's able to deal with big data. Breaking data into smaller pieces is a good method, but then how are going to perform the computation. For that its break the computation into small pieces and it sends each pieces computation each piece of data. So data break in down into equal pieces so each computation finishes with an equal amount of time. Ones all computations are finished, then the results are combined together and send back to the connected application.

At a very high-level Hadoop has a simple architecture to break data and computations (Sheshasayee and Lankshmi, 2014). Hadoop use map reduce algorithm and special kind of file system call Hadoop distributed file system (HDFS). Hadoop is a set of tools. These tools handle and manage by apache and objective of this tool is to provide an assistance in a task that is related to Hadoop. One importance characteristics of Hadoop are that its work on a distributed model. Hadoop hasn't used very powerful full computers and it uses low-cost computers called as commodity hardware. Hadoop is Linux-based set of tools. All low-cost computers have Task tracker and Data node. The job of task tracker is to process the smaller piece of the task that has been given to a particular node. The job of data nodes to manage the piece of data that has been given to a particular node. Figure 3 shows the Hadoop master-slave architecture.

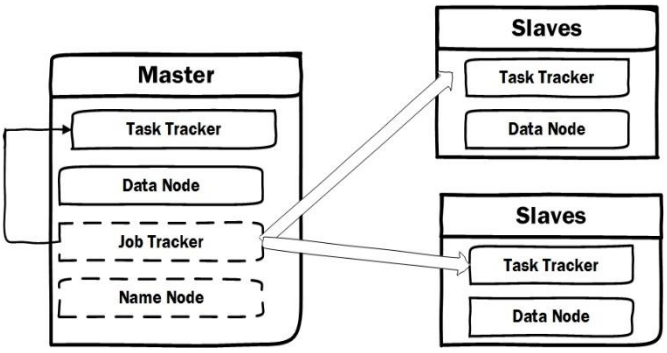


Figure 3. Hadoop Master-Slave Architecture

These task trackers and data node computers are called slaves, according to the architecture. The difference between master and slaves is the master has task tracker, data node, job tracker and name node. Job tracker and Task tracker are a part of Map-reduce and Data Node and Name Node is a part of HDFS (Kumar and Rathore, 2014). The applications that are running on Hadoop interact the master node. The special attribute of Hadoop is using batch processing set of tools. Thus applications would assign or provide a task for Hadoop to perform and its process according to the queue (Kulkarni and Khandewal, 2014) .

Job tracker is a module running on the master node and decomposes a bigger task into smaller pieces and to send each small piece of computation to task tracker. Then the task tracker executes these pieces and results sent to the job tracker. After that job tracker combines all the result and sent it to the task tracker. Name Node is in authority for keeping all indexes which data is residing on which data node. When an application context name node and it tells the application to go to a particular computer to get data (Katal et al., 2013). So application gets data directly from data node. When the hardware failures happen uncertainty, but is not matter for Hadoop because it has a built-in fault tolerant. Fault tolerance is achieved by redundancy. Hadoop maintains three copies of each file in different computers. One of a very important characteristic of Hadoop is that it is highly saleable and consist of one computer to thousands of computers. According to business needs developers can add computers so it is cost saving methods to a business.

### A. *Tools of Hadoop*

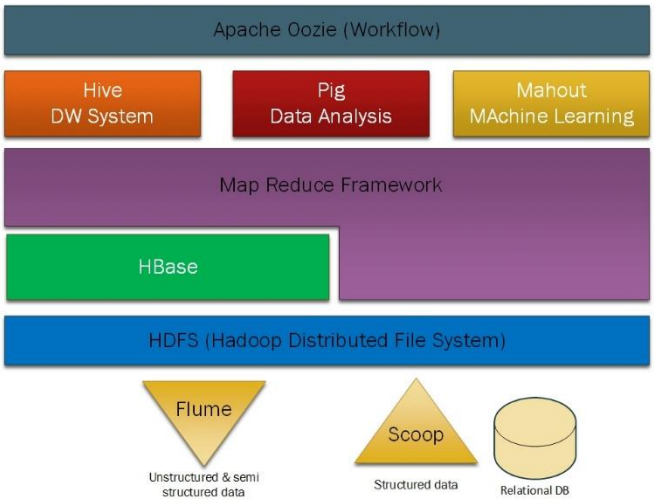


Figure 4. Hadoop Tool Structure

Hadoop consists of a set of tools which is managing by apache and fall under the umbrella of Hadoop. These tools provide core functionality at multiple levels. Namely Hive, HBase, Mahout, Pig, Oozie, Flume and Scoop. Figure 4 show that how tools interact with the Hadoop architecture.

Flume- A distributed service for gathering data, aggregating, and moving large volumes of log data. Flume has a simple and flexible architecture based on streaming data flow so that it is robust and fault tolerance. This tool allows for online analytic because of it use a simple extensible data model *(Rathee, 2013)*.

Hive - A data warehousing infrastructure built on top of Hadoop for providing data summarization, query, and analysis *(Taylor, 2010)*.

HBase – This is an open source, non-relational distributed database written on Java and runs top of the Hadoop *(Patil and Phursule)*.

Mahout – Apache mahout is distributed and scales up with machine learning algorithms on the Hadoop platform. Its work in some progressive manner and provide recommendations to users' taste based on the data which is provided by the user previously *(Barrachina and O'Driscoll, 2014)*.

Oozie – This is a Java-based application that runs in a Java servlet container. Oozie is responsible for scheduling workflow in the system to manage Hadoop jobs *(Mohammad Islam, 2012)*.

Pig – Actual competition is done by Map reduce component but in order to make the programming easy so higher-level language was created called Pig. This is similar to SQL in RDMS system. If you want to retrieve data from the certain table you don't have to write a detail level program you simply write an SQL command select column from the table and provide a condition. Then this high-level commands translated into a detail level instructions *(Loebman et al., 2009)*.

Scoop - A tool which is used to commute bulk data competently, between Apache Hadoop and structured datastores such as relational databases *(Aravinth et al., 2015)*.

### B. *Limitations of Hadoop*

This review paper presents a solution for big data processing using Hadoop. As well discuss the Hadoop architecture and how it works. Also, the Network File System (NFS) is one of the oldest and most commonly used distributed file system (Yu et al., 2014). It is designed to serve general purpose applications, however in the case of Hadoop only a specific kind of applications can make use of this technology. Hadoop gives solution for distributed file system problems (Shvachko et al., 2010). Also, Hadoop has some limitations such as random reads, updates, loss of performance, security concerns, not suitable for all kinds of applications, some vulnerabilities, stability issues and etc. (Gupta and Haryana, 2015). Hadoop is designed for the applications that require a sequential reach. The application has N number of parts and would like to read all the parts one by one. In random seek when you want to go to the specific location or part it's much compromised. Hadoop is designed for non-real-time batch processing of data. Hadoop is designed for streaming reads, but the caching of data is not provided ("5 Big Disadvantages of Hadoop for Big Data," n.d.). The reason caching is not available because you get faster access to the data directly by doing the sequential read fast to get access to the data to caching so caching is not available in Hadoop. But some

application using caching a get random data faster than Hadoop. The second characteristic in the application is written and read the data. But read data will not be updated. When the file is closed, it is not possible to update it but can read it. So update data are very hard (Karwande et al., 2015). System performance is lost in proportion to the number of nodes failed. Hadoop security model is disabled due to the sheer complexity and it is also missing encryption at the storage and network levels. So that data could in risk. Hadoop is Linux based and written in Java. Java and Linux have been heavily exploited by cyber criminals. So Hadoop became vulnerable by its nature. Hadoop developed on an open source platform. Because of open source Hadoop is devolved by the contributions of the many developers. While developments are continuously made, like all open source software, Hadoop has had stability issues.

### C. Overcome Limitations Using Other Technologies

Multi-agent technology is the most popular research field of the world. To get a solution for the real-world problem, a large number of agents involved and these agents deal with decision making and communications *(El Fazziki et al., n.d.)*. If millions of agents worked then the running time for each cycle can be several seconds. It's very speed and accurate. Agents typically include a set of features. They are autonomy, Pro-activity, Re-activity, negotiation and the learning. Some machine fails during the run time, then the whole process wants to restart. If we can dynamically rebalance the workload on remaining a number of computers and maintain the system integrity and consistency without performance degrading, so the system can continue the process from the point of failure. One of the problems in Hadoop is loss performance during failure, but the system can work with the failure so partially fault tolerance *(Sethia and Karlapalem, 2011)*. Hadoop requires a full fault tolerance and full failure- resilient framework which is can easily extensible to execute on a large number of processors and agents. For that only needs to develop a layer using agent-based simulation on top of the Hadoop. After implementing agents to of the Hadoop process never stop and not degrade performance when failure occurrence. It automatically balances the workload by communication with agents. This new framework can develop using based on the Java Agent Development Framework (JADE). It supports generic services such as communication with agents, resource discovery, content delivery and data encoding. Hadoop develops using Java and JADE framework also develop in Java so these two frameworks can easily connect without any interference. JADE provide data encoding so its solve some of the security issues on Hadoop.

Machine Learning (ML) is a part of an intelligent system that delivers computers with the ability to learn without being pre-programming. The developers use ML for their computer programs because of after applying ML can teach themselves to grow and change according to the new data. The difference between data mining and the ML is the programs developed using ML can detect patterns in data and adjust program behaviour according to the data. The purpose of integrating ML algorithms with Hadoop is to improve the timing efficiency of Hadoop system *(Asha et al., 2013)*. Hadoop consists with map reduce and HDFS *(Hans et al., n.d.)*. ML algorithms combine to map reduce paradigms. Hadoop using K-Nearest Neighbours algorithm which is a non-parametric method used in ML for classification data. KNN classifier is an instance-based learning algorithm. It based on Euclidean distance and Manhattan distance by observations. In the KNN classification workflow, K adjacent neighbours of a test sample are retrieved first. After that, the correspondences among the test sample and the k adjacent neighbours are aggregated according to the class of the neighbours. Then the test sample is assigned to the most similar class. The K is depending on data. The KNN algorithm takes data from Hadoop distributed file system and KNN classifies into several tasks *(Anchalia and Roy, n.d.)*. The task one is separate the lost values in the data set and the task two produces a separate sequence file for each of the missing rows. Task three is updating the lost value by calculating the distance. The final task is integrating the lost and non-missing rows and produces the output with no absent values.

## IV. Conclusion

This paper reviews on big data and Hadoop by considering security, data analysis, data storage methods and speed. Starting with RDBMS and describe the big data concepts. Furthermore, this theme concentrates on troubles of big data processing and solution for that. Open source Hadoop is a solution for big data processing problems. Later on that, paper talking about the how Hadoop works, architecture, how big data tools id useful for different aspects and its restrictions. Finally, found how to overcome Hadoop limitations using new technology such as multi-agents technology and machine learning.

## References

5 Big Disadvantages of Hadoop for Big Data [WWW Document], n.d. . Big Data Co. URL http://www.bigdatacompanies.com/5-big-disadvantages-of-hadoop-for-big-data/ (accessed 9.19.15).

Anchalia, P.P., Roy, K., n.d. The k-Nearest Neighbor Algorithm Using MapReduce Paradigm.

Aravinth, M.S., Shanmugapriyaa, M.S., Sowmya, M.S., Arun, M.E., others, 2015. An Efficient HADOOP Frameworks SQOOP and Ambari for Big Data Processing. Int. J. Innov. Res. Sci. Technol. 1, 252–255.

Asha, T., Shravanthi, U.M., Nagashree, N., Monika, M., 2013. Building Machine Learning Algorithms on Hadoop for Bigdata. Int. J. Eng. Technol. 3.

Barrachina, A.D., O'Driscoll, A., 2014. A big data methodology for categorising technical support requests using Hadoop and Mahout. J. Big Data 1, 1. doi:10.1186/2196-1115-1-1

Bhosale, H.S., Gadekar, D.P., 2014. A Review Paper on Big Data and Hadoop. Int. J. Sci. Res. Publ. 4.

Borkar, S.D., Surtakar, C.S., 2014. A REVIEW PAPER ON THE HADOOP DISTRIBUTED FILE SYSTEM. Int. J. Res. Sci. Eng. 1.

Dhomse, G., Komal, K., Manali, L., Latika, A., 2015. A Review Approach for Big Data and Hadoop Technology. Int. J. Mod. Trends Eng. Res.

El Fazziki, A., Sadiq, A., Ouarzazi, J., Sadgal, M., n.d. A Multi-Agent Framework for a Hadoop Based Air Quality Decision Support System.

Essa, Y.M., Attiya, G., El-Sayed, A., 2013. Mobile Agent Based New Framework for Improving Big Data Analysis. IEEE, pp. 381–386. doi:10.1109/CLOUDCOM-ASIA.2013.75

Gupta, B., Jyoti, K., 2014. Big Data Analytics with Hadoop to analyze Targeted Attacks on Enterprise Data. Int. J. Comput. Sci. Inf. Technol. 5.

Gupta, P., Haryana, K., 2015. Big Data: Problems, Challenges and Techniques. J. Environ. Sci. Comput. Sci. Eng. Technol. 4.

Gupta, T., Dixit, S., 2015. A Brief Outline on Bigdata Hadoop. Int. J. Emerg. Technol. Adv. Eng. 5.

Hans, N., Mahajan, S., Omkar, S.N., n.d. Big Data Clustering Using Genetic Algorithm On Hadoop Mapreduce.

Kalra, S., lamba, A., 2014. A Review on HADOOP MAPREDUCE-A Job Aware Scheduling Technology. Int. J. Comput. Eng. Res. IJCER 4.

Karun, K., K, C., 2013. A Review on Hadoop – HDFS Infrastructure Extensions. Presented at the Information and Communication Technologies (ICT 2013), IEEE.

Karwande, V.., Lomte, S.., Auti, R.., 2015. The Data Recovery File System for Hadoop Cluster -Review Paper. Int. J. Comput. Sci. Inf. Technol. 6.

Katal, A., Wazid, M., Goudar, R.H., 2013. Big data: Issues, challenges, tools and Good practices, in: Contemporary Computing (IC3), 2013 Sixth International Conference on. IEEE, pp. 404–409.

Korth, Sudarshan, 2001. Database System Concepts.

Kulkarni, A.P., Khandewal, M., 2014. Survey on Hadoop and Introduction to YARN. Int. J. Emerg. Technol. Adv. Eng. 4.

Kumar, R., Rathore, D.V.S., 2014. Efficient Capabilities of Processing of Big data using Hadoop Map Reduce. Int. J. Adv. Res. Comput. Commun. Eng. 3, 7123–7126.

Loebman, S., Nunley, D., Kwon, Y., Howe, B., Balazinska, M., Gardner, J.P., 2009. Analyzing massive astrophysical datasets: Can Pig/Hadoop or a relational DBMS help?, in: Cluster Computing and Workshops, 2009. CLUSTER'09. IEEE International Conference on. IEEE, pp. 1–10.

Mohammad Islam, A.K.H., 2012. Oozie: towards a scalable workflow management system for Hadoop. doi:10.1145/2443416.2443420

Mouliswaran, S.C., Sathyan, S., n.d. STUDY ON REPLICA MANAGEMENT AND HIGH AVAILABILITY IN HADOOP DISTRIBUTED FILE SYSTEM (HDFS). J. Sci. Inf. Technol. 2.

Patil, P.S., Phursule, R.N., n.d. Survey Paper on Big Data Processing and Hadoop Components.

Poggi, N., Carrera, D., Call, A., Mendoza, S., Becerra, Y., Reinaure, R., Vujic, N., Green, D., Blakeley, J., 2014. ALOJA: a Systematic Study of Hadoop Deployment Variables to Enable Automated Characterization of Cost-Effectiveness. Presented at the IEEE International Conference on Big Data, IEEE.

Rathee, S., 2013. Big Data and Hadoop with components like Flume, Pig, Hive and Jaql, in: International Conference on Cloud, Big Data and Trust. pp. 13–15.

Rathi, R., Lohiya, S., 2014. Big Data and Hadoop. Int. J. Adv. Res. Comput. Sci. Technol. 2.

Robbins, R., 1995. Database Fundamentals.

Sethia, P., Karlapalem, K., 2011. A multi-agent simulation framework on small Hadoop cluster. Eng. Appl. Artif. Intell. 24, 1120–1127. doi:10.1016/j.engappai.2011.06.009

Shah, G., Annappa, A., Shet, K.C., 2014. Design an Efficient Big Data Analytic Architecture for Retrieval of Data Based on Web Server in Cloud Environment. Int. J. Cloud Comput. Serv. Archit. 4, 1–10. doi:10.5121/ijccsa.2014.4201

Sheshasayee, A., Lankshmi, J., 2014. A STUDY ON HADOOP ARCHITECTURE FOR BIG DATA ANALYTICS. Int. J. Adv. Technol. Eng. Sci. 2.

Shilpa, Kaur, M., n.d. BIG Data and Methodology-A review. Int. J. Adv. Res. Comput. Sci. Softw. Eng. 3.

Shvachko, K., Kuang, H., Radia, S., Chansler, R., 2010. The hadoop distributed file system, in: Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium on. IEEE

Sindhuja, K., Monisha, A.V., Padmavathi, S., 2015. Performance Analysis of Agent Based Framework. Procedia Comput. Sci. 47, 37–44. doi:10.1016/j.procs.2015.03.181

Taylor, R.C., 2010. An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics. BMC Bioinformatics 11, S1. doi:10.1186/1471-2105-11-S12-S1

Thompson, C.A., A. Latchman, H., Angelacos, N., Kumar Pareek, B., 2013. A Distributed IP-Based Telecommunication System using SIP. Int. J. Comput. Netw. Commun. 5, 121–136. doi:10.5121/ijcnc.2013.5607

Twardowski, B., Ryzko, D., 2014. Multi-agent Architecture for Real-Time Big Data Processing. IEEE, pp. 333–337. doi:10.1109/WI-IAT.2014.185

Why NoSQL? [WWW Document], n.d. URL http://www.couchbase.com/nosql-resources/what-is-no-sql

Yu, W., Wang, Y., Que, X., 2014. Design and evaluation of network-levitated merge for hadoop acceleration. Parallel Distrib. Syst. IEEE Trans. On 25, 602–611.