Journal of Scientific Research & Reports

26(9): 58-84, 2020; Article no.JSRR.62638
ISSN: 2320-0227

Using Hadoop Technology to Overcome Big Data Problems by Choosing Proposed Cost-efficient Scheduler Algorithm for Heterogeneous Hadoop System (BD3)

Abou_el_ela Abdou Hussein[1*]

[1]Department Computer Science, Modern Academy-Maddi, ARE, Egypt.

Author's contribution

The sole author designed, analysed, interpreted and prepared the manuscript.

Article Information

DOI: 10.9734/JSRR/2020/v26i930310
Editor(s):
(1) Dr. Kleopatra Nikolopoulou, University of Athens, Greece.
Reviewers:
(1) Mohamed Doheir, Universiti Teknikal Malaysia Melaka (UTeM), Malaysia.
(2) A. S. Manekar, Sant Gadge Baba Amravati University (SGBAU), India.
(3) Vijayaraj J, Pondicherry University, India.
Complete Peer review History: http://www.sdiarticle4.com/review-history/62638

Original Research Article

Received 28 August 2020
Accepted 04 November 2020
Published 28 November 2020

ABSTRACT

Day by day advanced web technologies have led to tremendous growth amount of daily data generated volumes. This mountain of huge and spread data sets leads to phenomenon that called big data which is a collection of massive, heterogeneous, unstructured, enormous and complex data sets. Big Data life cycle could be represented as, Collecting (capture), storing, distribute, manipulating, interpreting, analyzing, investigate and visualizing big data. Traditional techniques as Relational Database Management System (RDBMS) couldn't handle big data because it has its own limitations, so Advancement in computing architecture is required to handle both the data storage requisites and the weighty processing needed to analyze huge volumes and variety of data economically. There are many technologies manipulating a big data, one of them is hadoop. Hadoop could be understand as an open source spread data processing that is one of the prominent and well known solutions to overcome handling big data problem. Apache Hadoop was based on Google File System and Map Reduce programming paradigm. Through this paper we dived to search for all big data characteristics starting from first three V's that have been extended during time through researches to be more than fifty six V's and making comparisons between researchers to reach to best representation and the precise clarification of all big data V's

**Corresponding author: E-mail: abo_el_ela_2004@yahoo.com;*

characteristics. We highlight the challenges that face big data processing and how to overcome these challenges using Hadoop and its use in processing big data sets as a solution for resolving various problems in a distributed cloud based environment. This paper mainly focuses on different components of hadoop like Hive, Pig, and Hbase, etc. Also we institutes absolute description of Hadoop Pros and cons and improvements to face hadoop problems by choosing proposed Cost-efficient Scheduler Algorithm for heterogeneous Hadoop system.

Keywords: Big data; traditional techniques; hadoop; hadoop distributed file system; MapReduce; improvements; Scheduler.

1. INTRODUCTION

With the use of digitization of last decades, the capacity to create and interchange data, information, messages, audios, and videos over networks unimaginably or even believed has grew and expanded. This enormous data revolution phenomenon was called big data. Big data as a term used for such a collection of massive datasets, have typical characteristics as, fast-moving, multi-source origin, tremendously large and unstructured [1,2,3] and also has been one of the greatest trend that attracted many researchers to do research introducing how to utilize and benefit from this technology. Also big data refers to the enormous amounts of digital information that collected or captured, stored, distributed, manipulated, interpreted, analyzed, investigated and visualized between people all over the world. With the appearance of big data technologies like Hadoop and existing models as MapReduce and Clustering algorithm, as explained latter in this paper, there is more predicts for productively analyzing for big data sets. The paper is structured as follows: in section 2 we introduce back ground that includes grid, data grid, grid computing, computational grid, and cloud computing, Sources of Big Data, big data 56 V's characteristics, Big Data analytics, and Big Data Processing Problems. In section 3 we introduce related topics includes Hadoop as a solution to overcome BD problems, Problems with Hadoop, Proposed Improvements to Overcome Hadoop problems, Hadoop Tools for Handling BD, and Hadoop Applications. In section 4 we highlights Hadoop Improvements using Hadoop Scheduling technique, and chosen proposed Cost-efficient Scheduler Algorithm for heterogeneous Hadoop system. Section 5 contains Conclusion.

2. RELATED STUDY

The web makes it easy to collect and share knowledge as well as data in raw form. Big Data is about how these data can be stored, processed, and comprehended with an aim of using it in expecting the action in future with a reasonable accuracy and allowable time delay. The current and emerging focus of big data analytics is to explore traditional techniques as pattern mining, decision trees rule-based systems, and other data mining techniques to progress business rules even on the large data sets efficiently. It can be accomplish by either progress algorithms that uses spread data storage, in-memory computation or by using cluster computing for mutual computation. Earlier these processes were implemented using grid computing, which was passed by cloud computing in recent days.

2.1 Data Grid

Data grids supply a base to support data storage, data recognition, data processing, data spread, and data manipulation of huge volumes of data really saved in various contrasting databases and file systems [4]. Also can be defined as a system made up from several servers that process simultaneously to manipulate information and associated operations such as computations in a spread environment. A data grid could also introduced as an architecture or set of services that gives users the ability to access, process and transport extremely huge amounts of spread data for research purposes.

2.2 Grid Computing

Grid computing is a form of spread computing that uses geographically and managerial different resources [4]. Also Grid computing, is a means of customize the computing power in spread means to solve great problems and that requires lots of operating time and power [5]. Grid computing could be constituted by many connected servers using a high rate network; each server participates on one or numerous roles. Grid computing main benefits are the high repository capability and the processing energy by offers

the opportunity to hardness unutilized computing power. Grid computing gives the opportunity for beneficiaries to share resources as processing and storage capabilities to be used by different people. Grid computing goal is to reach computers only when demanded and widening the range of problems in which even small computers can make a contribution to the grid. We can consider every device connected to the Internet in a grid computing as a node in a hugely large computing machine. In the grid domain, the problem is divided and spread to enormous number of computers to obtain a solution in a cost effective way. There are several applications using this technology as astronomy, weather, medicine, multi-player gaming, etc. Typically grid computing works on two dominant models: Commercial Model and Social Model [5].

- Commercial Model: It works on the principle that this technology can be used for the commercial purpose by setting up large processing centers, selling their capabilities to hourly users and collecting money [5]. The advantage of this model is Quality of Service (QoS) maintenance and is a reliable method of computation.

- Social Model: It works with the understanding that these resources must be harnessed for the benefit of society [5]. Grid computing technique is implemented through programs that follow Open Grid Service Architecture (OGSA). Globus toolkit is popular software program that implements OGSA and is used in a grid computing environment.

2.3 Computational Grid

A computational grid is a hardware and software framework that provides dependable, fixed, full, and cheap access to advanced computing capabilities. This grid provides protected access to huge pool of shared processing power appropriate for high throughput applications. The computational grids provide a convenient way to connect multiple devices, helping to reduce the power consumption and also increases system speed. Computational grids are used by many organizations for example, health maintenance, material science collaboratory, computational market economy, government etc [4,6].

2.4 Cloud Computing

Recently cloud computing is considered very important concept. Its roots comes from grid computing and other related concepts like distributing systems, utility computing, and cluster computing. Cloud computing indicates to the concept of computing at a faraway location with manage at the users' end (computer or even mobile phones) through a thin client system [5]. Processing, memory, and storage will be done at the service providers' infrastructure as explained in Fig. 1 [5].Users need to connect to the virtual system occupy at some faraway location, that run different virtual operating systems on physical servers with the aid of virtualization. Cloud computing assist all types of fault tolerant features like live migration, scalable storage, and load balancing [7].

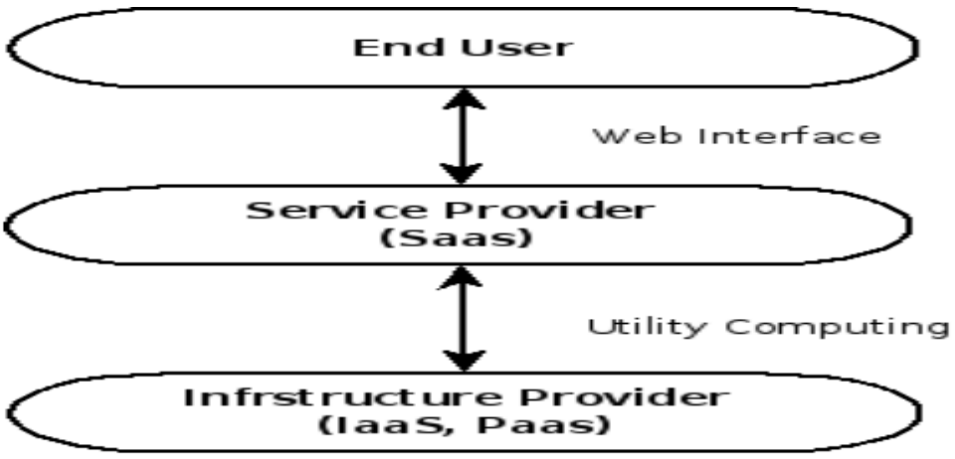

Fig. 1. Basic model of cloud computing

Cloud computing works on three dissimilar levels, namely, Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS) [8,9] as explained in Fig. 2,[10]. Also Cloud computing agonize from similar negatives of grid computing as data replication, data location, security threats, data segregation, regulatory compliances, recovery issues, long-term viability, high dependency on the Internet for accessing the remote virtual machine, investigative support, different laws of different countries, etc [5]. Cloud computing also suffers from similar defects in grid computing like data location, data replication, data segregation, security threats, regulatory compliances, recovery issues, long-term viability, high reliance on the Internet for remote virtual machine access, different laws from different countries, the concept of investigative support, etc.

2.5 Sources of Big Data

We present here the different sources of Big Data. Digitization of content by industries is the new source of data [5] .Advances in technology are also leading to a higher rate of data generation. For example, one of the largest surveys in astronomy, the Sloan Digital Sky Survey (SDSS) recorded a total of 25 terabytes of data during the first (2000-2005) and second (2005-2008) surveys combined. With advances in the telescope accuracy, the amount of data collected reached hundreds Terabytes. The use of "smart" devices is another source of big data. Smart meters in the energy sector record electricity use every15minutes compared to previous monthly readings. In addition to social media, the Internet of Things (IoT) is now the new source of data. Data can be obtained from agriculture, industry, medical care and other smart cities developed on the basis of the Internet of Things [5]. Table 1 summarizes the various sources of data produced in different sectors [5],

2.6 Fifty Six V's Characteristics of Big Data

Recently, the term "Big Data" has become a very well-known term, although it isn't expresses accurately because it points only to the size of data ignoring obtainable properties. The concept of big data returns back to the year 2001, with a 3Vs model [11,12]. Theses 3Vs, also known as the dimensions of big data, represent the increasing Volume, Variety, and Velocity of data. After the first big data 3 V's characteristics, began getting a lot of attention for many people to add some more V's to the characterization of big data.as explained in our paper as in Fig. 3 [11,12], these 3 Vs exceeds and become more than fifty six V's characteristics, two of them were added by the author as explained in [11,12]. Different declarations for each "V" characteristic (dimension) are explained as follows as in Table 2 [11,12]:

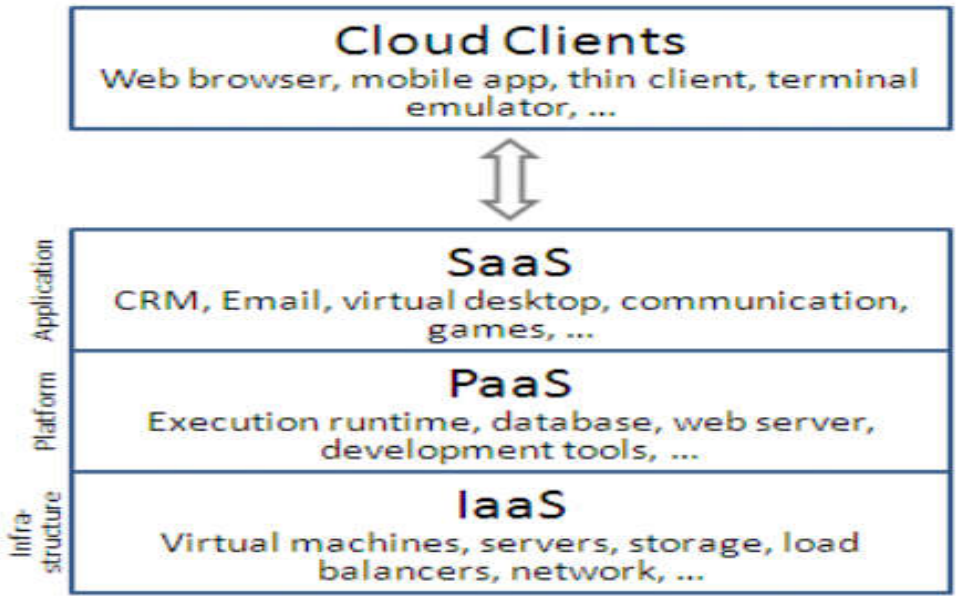

Fig. 2. Cloud computing infra-structure, platform, and application

Table 1. Different sources of data

Sector	Data Produced	Use
Astronomy	Movement of stars, satellites, etc.	To monitor the activities of asteroid bodies and satellites
Financial	News content via video, audio, twitter and news report	To make trading decisions
Healthcare	Electronic medical records and images	To aid in short-term public health monitoring and long-term epidemiological research programs
Internet of Things (IoT)	Sensor data	To monitor various activities in smart cities
Life Sciences	Gene sequences	To analyze genetic variations and potential treatment effectiveness
Media/Entertainment	Content and user viewing behavior	To capture more viewers
Social Media	Blog posts, tweets, social networking sites, log details	To analyze the customer behavior pattern
Telecommunications	Call Detail Records (CDR)	Customer churn management
Transportation, Logistics, Retail, Utilities	Sensor data generated from fleet transceivers, RFID tag readers and smart meters	To optimize operations
Video Surveillance	Recordings from CCTV to IPTV cameras and recording system	To analyze behavioral patterns for service enhancement and security

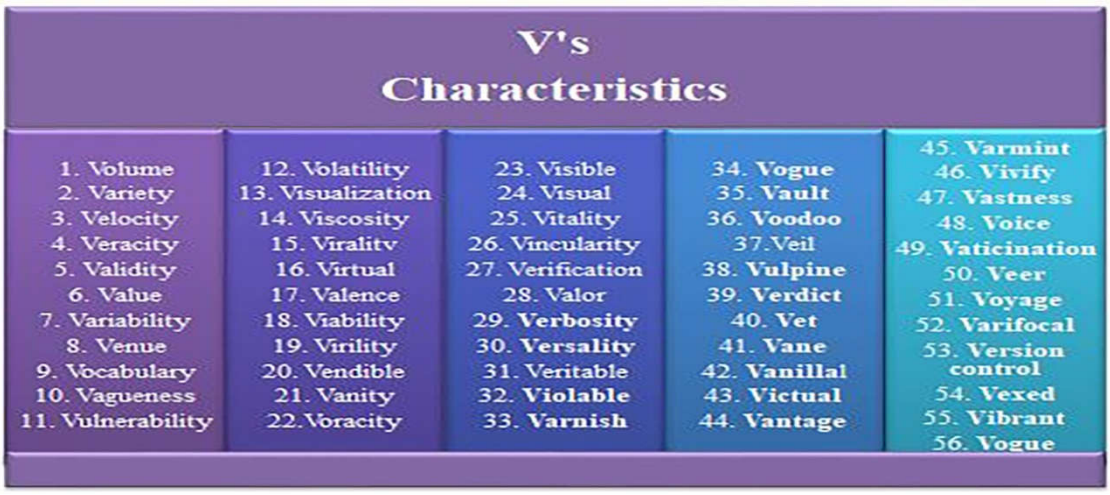

Fig. 3. Big data fifty six V's characteristics

2.7 Big Data Analytics

Information storage is an ancient process that people do when they try to write texts or stamped symbols onto a rock and clay disc to simulate some knowledge [27]. The recent computer advancements is reflected in easy and cheap processing that makes storing and representing data a trivial task. Datafication could be defined as the process of collecting, storing, and analyzing data related to a specific activity or process and turn it into assess to make it more clear and help to make certain decisions [27]. Datafication is combined with sensors industry that is reproduce due to the low cost of sensors. Datafication plays a big role as the backbone of Big Data invasion. So, data can be easily generated, gathered, stored, modified, analyzed, implemented, and visualized. The term big data analytics could be subdivided into two expressions. First Big data can be considered as huge data sets that may be analyzed to reveal patterns, trends, and associations, especially relating to human behavior and interactions.

Second, the analytics is a mix of different type of analytical tools as statistical analysis, data mining, Natural Language Processing (NPL), etc [28]. Gathering first and second expressions we get big Data analytics that is the use of advanced analytic techniques that are used to analyze the data, extract hidden pattern and explore insights out of data against very large, diverse data sets that include structured, semi-structured and unstructured data, from different sources, and in different sizes from terabytes to zettabytes.

Table 2. Characteristics of big data: The 56 V's

V's Characteristics	Explanation
1. Volume	Many companies have already amount of archived "Ocean of data" in the form data or of information that can came from every possible sensor, logs, Hundreds hours of YouTube uploaded videos, billions gigabytes from global mobile traffic [11,12].
2. Variety	Big Data is represented by different formats and varied types of data between structured, semi-structured, multi-structured and mostly unstructured data as well that came from many types of data resources, so it is heterogeneous in both size and type, consequently cannot be putted together into a relational database [11,12,13].
3. Velocity	Could be defined as the speed of data traveling from one side to another or moves around and the speed of processing it with high rate of receiving data and information [11,12].
4. Veracity	We need clear and definite answer for a very important question, does data comes from a reliable source [11,12].
5. Validity	How quality consistence, preciseness, reasonableness and correctness the data for its intentional use [11,12].
6. Value	Unless turning the enormous amount of data in big data into value, it could be useless and unusable [11,12].
7. Variability	Variability in big data's circumstances means variability in the data, which required to be found by deviation and aberration detection methods leading for any relevant analytics to occur [11,12,14].
8. Venue	Big data is distinguished by its distributed heterogeneous data from various platforms, from numerous owners' systems, with different formatting and access needs, private or popular [11,12,15].
9. Vocabulary	All metadata shapes like data models, schema, semantics, ontologies, taxonomies, and other contents that describe the data's structure, syntax, content, and origin [11,12,15].
10. Vagueness	The meaning of found data is often very unclear, not only has how much data been available but also how much it is not obscure [11,12,15].
11. Vulnerability	This means that no system is perfect, which means it's probable there is a way for its hardware or software to be agreement, successively meaning that any associated data can be tacked or manipulated [11,12,14].
12. Volatility	What time does remain data valid and should be stored. How old dose data need to be before it is considered irrelevant [11,12,14].
13. Visualization	Refers to the application of more recent visualization techniques to explain the relationships between data and can display real-time changes and more illustrative graphics, thus going beyond pie, bar and other charts [11,12,14].
14. Viscosity	It is occasionally used to express the delay, latency or lost time in the data relative to the phenomenon being described [11,12,16].
15. Virality	Measures the rate at which data can propagates through a network [11,12,16].
16. Virtual	Enterprises and other groups can benefit from big data virtualization because it authorize them to use all the data assets they gather to accomplish various goals and objectives [11,12].
17. Valences	It is a measure indicating how dense the data is [11,12].

V's Characteristics	Explanation
18. Viability	Viability could be seen as carefully choosing those attributes in the data that are most likely to forecast outcomes that matter most to organizations [11,12].
19. Virility	With big Data it means that it creates itself. The more Big Data you have, the more Big Data gets strength and forceful [11,12,17,18,19,20].
20. Vendible	The very existence of client's for Big Data shows crucially that it is appreciable – this is evident from the communication of some known means of trading with subscribers data [11,12,17,21,22].
21. Vanity	Vain of data means that it is glad with the effect it produces on other individuals [11,12,17,21,23].
22. Voracity	Big Data is potentially so insatiable that it may achieve the influence, manage and the possibility to consume itself [12,17,21,23].
23. Visible	Not only pertinent information should exist, but also should be evident to the intended person at the proper time [11,12,22].
24. Visual	We currently live in a world of seeing, watching, and exchanging photos and videos, whether they are personal or product pictures or weather photos through the Internet [11,12,17,22].
25. Vitality	Vitality of the data is an important perception that is vital and is included in the concept of Value [11,12,17,22].
26. Vincularity	It implies in its exact meaning connectivity or linkage. This idea is very pertinent in today's interconnected world through the internet [11,12,17,24].
27. Verification	The process of initiate the fact, precision, or validity of data [11,12,17,22].
28. Valor	The specific data that has the possibility to produce value and guiding how this can be accomplished [11,12,17,25].
29. Verbosity	Understanding how to quickly separate the meaning you keep about from its repetition is important for efficiency of processing [11,12,17,22].
30. Versality	Versatility of data shows to what extent the data is useful, in different scenarios [11,12,17,22].
31. Veritable	Data being in fact the thing named and not false, unreal, or imaginary [11,12]
32. Violable	Violable data capable of being or likely to be violated [11,12].
33. Varnish	Interaction of end-users with our work matters, and polish counts. [11,12].
34. Vogue	Artificial intelligence are become? [11,12,26].
35. Vault	Importance of data security [11,12,26].
36. Voodoo	Deliver results with real-world impact [11,12,26].
37. Veil	Examine latent variables from behind the curtain [11,12,26].
38. Vulpine	Data leads to a new technology [11,12,26].
39. Verdict	People affected by model's decision [11,12,26].
40. Vet	Vetting the assumptions with evidence [11,12,26].
41. Vane	Unclear direction of decision-making [11,12,26].
42. Vanilla	Simple methods if tackled with care, can provide value [11,12,26].
43. Victual	Big Data fuel of data science [11,12,26].
44. Vantage	Privileged view of complex systems [11,12,26].
45. Varmint	As data gets bigger, so do software bugs [11,12,26].
46. Vivify	Ability of data science to cope with every real-life aspect [11,12,26].
47. Vastness	Bigness of Big Data [11,12,26].
48. Voice	Ability to speak with knowledge [11,12,26].
49. Vaticination	Ability to forecast [11,12,26].
50. Veer	Change direction according to customer need [11,12,26].
51. Voyage	Increasing knowledge [11,12,26].
52. Varifocal	It is about trees and forest [11,12,26].
53. Version control	You are using it right? [11,12,26].
54. Vexed	Potential of data science to handle complicated problems [11,12,26].
55. Vibrant	Provision of insight by data science [11,12,26].
56. Vogue	Artificial intelligence will become? [11,12,26].

2.8 Trials and Problem Associated with Big Data Processing

Big data is a big challenge [29]. Data storage will flood your access. The data- processing tools you need are new and intimidating. Business-value insights to analyze data to find real work.

2.8.1 Storage

The data for companies that deal with the most mysteries is how it is stored [29]. The old traditional method of "spinning dish" was to use a hard drive. These are slow, but not much to save a lot of money. Speed is a priority, so companies with SSDs can move in for a bit more money. Ten to a hundred times faster, but more expensive. In fact, the turbo charge who wants to establish his company, the computing "in memory of" is present. This figure is 10,000 times faster than the old methods maybe stored in RAM, which allows you.

2.8.2 Process

Talk to a data advisor and Hadoop before starting to shit about how long you can [29]. In fact, Big Data and Hadoop are inseparable from the many folks pretending to be themselves. For newcomers, Hadoop is a way to store and manipulate data in a separate departments or groups. It has huge reserves for handling data, making it ideal for growth, easy-to-manage and intuitive means. So what's new? The whisper on the road Hadoop began to see that. Here is Sanjay Joshi, Indian giant Mahindra Tech Global Big Data owner is: "Hadoop has been around for nearly a decade, and it comes to technology, that's a long time ago and it looks like there are some limitations.

2.8.3 Analysis

OLDE people in large numbers, data scientists who get paid can only analyze [29]. He baffled and astonished the spectators with their coding skills. Today, it is possible to use Additional employees by using simple graphical interface data to search for valuable patters. Drag and drop most popular chart which use Tableau, Qlikview, and SAS Visual Analytics and are Opinurate. Founded by Lucky Voice karaoke series, Martha Lane-Fox uses Tableau for planning. Staff songs, the most popular food and beverage for customers who prefer to look at reservation data, and can play an important role in rebooking. It avoids the Tableau coding format completely. Users simply take over and create a new diagram that lists the data transfer. Walmart is suitable for promoting cross selling products and additional selling of its Neo4j chart database to identify uses. This tech-less employee is fairly easy to use. Big data software makers recognize this trend and now these systems collaborate easily to ensure the network. For example, log analytics, the maker of the concept tool, scientists recently rely on data to end with the stated goal, and HP has partnered with Vertical Analytics Platform. Communications between the two are automatic. Scientists hardly have data. The second is the desire to tools should be considered.

2.8.4 Information growth

In big data it's the most important issue that is size [30]. We heard the word big data, the first thing we see in our minds is volume and size. Managing the large and rapidly growing amount of data is a challenging task. The data is growing faster and the CPU speed is constant and static.

2.8.5 Speed

Size matters speed. If the data set is larger and contains large information, it will take longer to respond [30].

2.8.6 Privacy and security

Data privacy is the big issue that arises in big data. In the United States, there is much fear about inappropriate use of personal data [30].

3. HADOOP AS A SOLUTION TO OVERCOME BIG DATA TRIALS AND PROCESSING PROBLEMS

Big data is a big challenge. Data storage will flood your Hadoop on large data sets in a distributed computing arrival [29]. The data-processing tools you need are new and intimidating. And environment used to support the processing of a programming commercially valuable insights to analyze the data to find there framework. Hadoop is the most popular technology beginning of twenty –first century that is able to mine and sort data. Apache Hadoop is based on google file system named Hadoop Distributed File System(HDFS) for storing data and Map Reduce Programming paradigm for processing stored data as explained in Fig. 4 [30,31]. Apache Hadoop Ecosystem is explained in Fig. 5 [27]. Hadoop is the most implemented

solution for handling big data using Map Reduce software framework that divides(breaks down, splits) the task into small parts and assigns it to separate computer and collect the result in data set. Hadoop has enough flexibility to work with multiple data sources, or even assemble multiple systems to be able to do large scale processing.

3.1 Hadoop Advantages

3.1.1 Range of data sources

The data collected from different sources will be in different shapes variety of structured, semi-structure, multi-structured, and unstructured form. The sources can be social media like Facebook and twitter or even email conversations [30]. Much time will be required to convert all collected data into one format. Hadoop saves this time because it can extract valuable data from any form of data. It also has a variety of functions such as data warehousing, fraud detection etc.

3.1.2 Cost effective

The companies had to spend lots of amount of their benefits into storing large amounts of data. In some cases they had to delete large groups of raw data in order to make space for new data [30]. There was a chance of waste important information in such cases. With Hadoop, this issue is completely resolved. It is a cost practical solution for data storage purposes.

3.1.3 Speed

Every organization wants to use a platform to get the work done as fast as possible rate [30]. Hadoop enables the organization to do just that with its data storage needs. It uses a storage system where the data is stored on a spread file system.

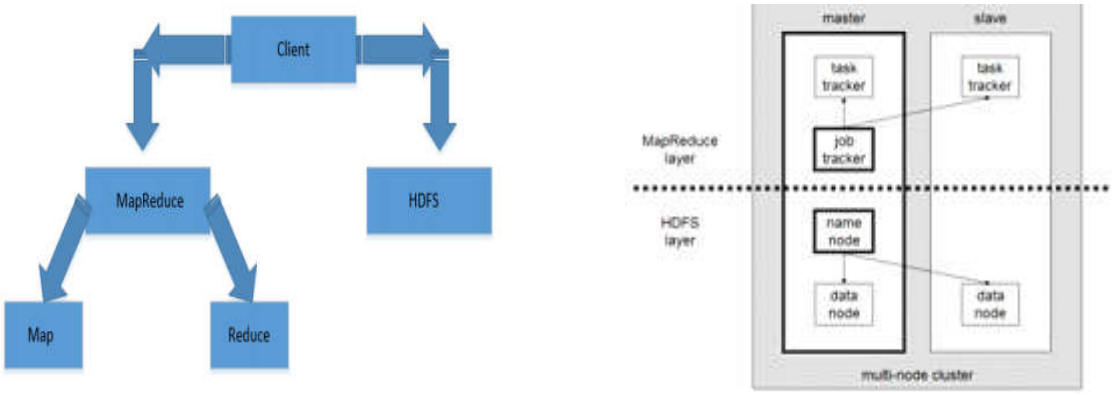

Fig. 4. General components and structure of hadoop

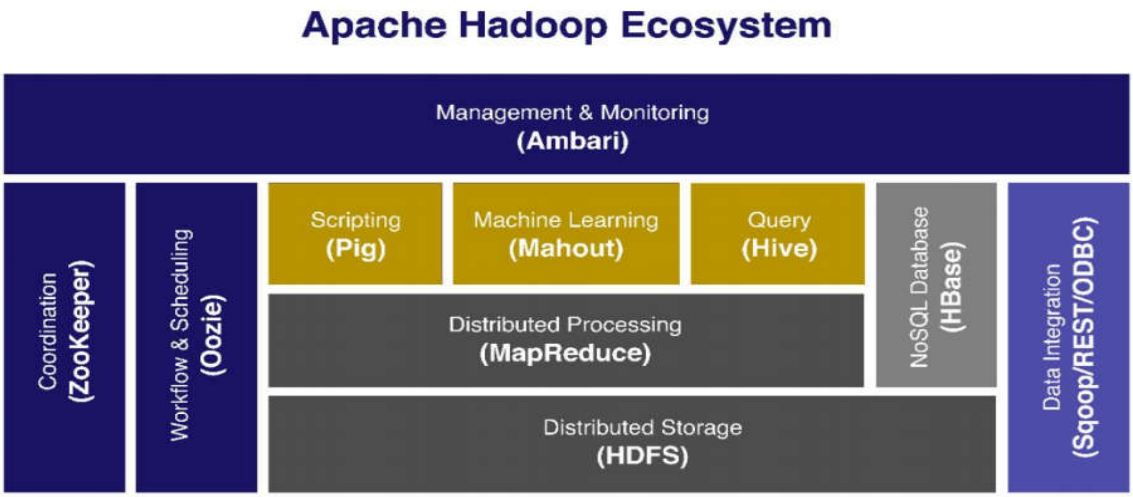

Fig. 5. Apache hadoop ecosystem

3.1.4 Multiple copies

Hadoop automatically duplicates the data that is stored in it and creates multiple copies [30]. This is done to ensure that data is not lost in the event of a malfunction. Hadoop understands that the data stored by the organization is important and should not be lost unless the company ignores it.

3.2 Problems with Hadoop

Although Hadoop has many advantages, it still faces some problems. The main issue for Hadoop is lower data processing efficiency than traditional DBMSs [32,33].

3.2.1 Lack of query optimization (Limited SQL support)

HDFS does not support SQL and improved query techniques. Map and Reduce functions are used to complete the process [32]. This requires users to create new functionality in Java if they need special functionality to meet their specific requirements. MapReduce has difficulties trying to access data from SQL database, and since the SQL database is still used in the required major database systems, this is a problem for Hadoop. Fig. 6 illustrates MapReduce functionality within Hadoop. Before the Mapper function works, a single file entry is uploaded in HDFS, and the file is split into multiple blocks during upload. Each block is assigned to each mapper task and yields an inter-mediate result. Transfer all intermediate results to the Reducer task over HTTPS. MapReduce works after loading all data into central storage from various locations and manipulating the data without modeling the data [32]. This process will be idle because MapReduce engine is not running until data acqusition. With traditional databases, different systems fit different data models and reduce data transfer. However, Hadoop will ignore all implementation plans and improve data modeling and raw data processing plan. So MapReduce has lower performance than DBMS.

3.2.2 Small data concerns

There are a few big data platforms in the market that are not fit for small data functions [30]. Hadoop is one such platform where only large companies that generate big data can benefit from its functionality. It cannot perform efficiently in small data environments.

3.2.3 Risky functioning

Java is one of the most widely used programming languages [30]. It has also been connected to various communities because cyber criminals can easily exploit the frameworks that are built on Java. Hadoop is one such structure that is built completely on Java. Therefore, the platform is weak and can cause unexpected damages.

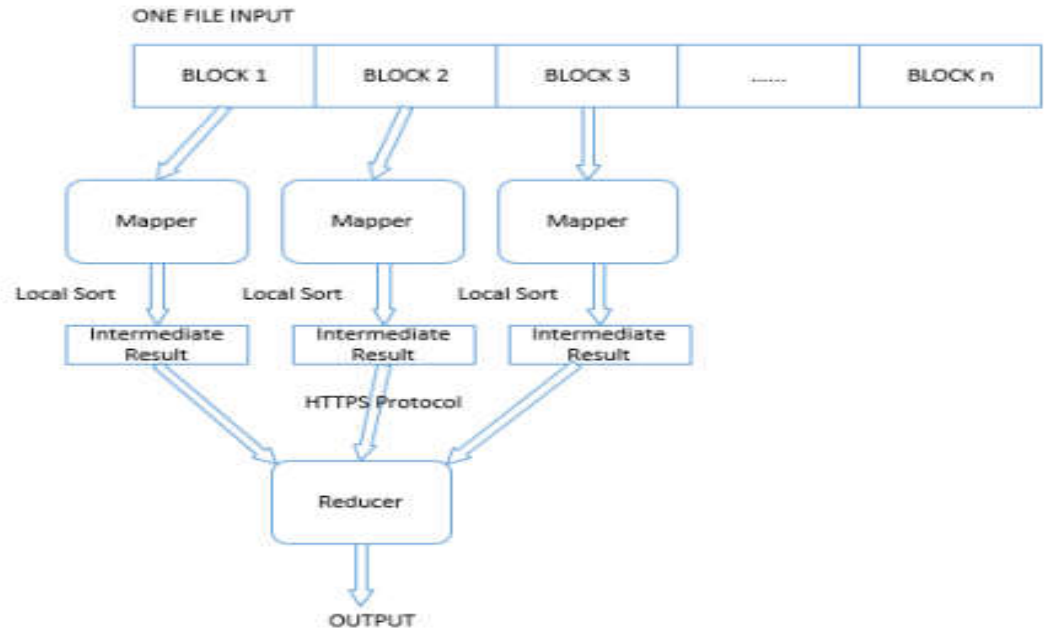

Fig. 6. Graphical representation of map reduce collective functions

3.2.4 Parallelism is absent in hadoop

Parallel processing refers to multiple processors that access data at the same time; however, MapReduce uses single workflow processing instead of parallel data processing [32]. Meanwhile, MapReduce designs a single read input and output, which means regardless of the number of data processing, to achieve higher processing requests messages are sent, MapReduce only works on one at a time. Transfer to the next stage will not be granted until the current stage ends; therefore, MapReduce is a blocking process.

3.2.5 Simplified scheduling schemes in hadoop

In order to achieve ease of use and fault tolerance, Hadoop has a simple data- processing scheduler [32]. Hadoop uses default FIFO planning for Job Tracker vertex .In this method, JobTracker manipulating data one by one, entering first in and first out; hence this simple planning is not able to deal with the real problem. Hence, this single strategy and simple operation scheduling show less efficiency in real-time processing. Another problem with Hadoop is the lack of data optimization before the central station reaches local results. In DBMS, data is always protected before committing, but in real time processing, since data is never stored, the system must provide a suitable method to avoid data loss, processing delay or data overhead.

3.2.6 Multiple copies of data

HDFS inefficiency creates of multiple copies of data (at least 3 copies) [5].

3.2.7 Inefficient execution

Lack of query optimizer results in inefficient cost-based plan to implement resulting greater cluster compared to similar database [5].

3.2.8 Challenging framework

It is not possible to take advantage of Complex transformational logic with the MapReduce framework [5].

3.2.9 Lack of skills

Knowledge of algorithms and skills are required to develop distributed MapReduce for proper implementation [5].

One of the biggest challenges is having a computing infrastructure that can analyze large and varied (structured and unstructured) data from multiple sources and enable real-time analysis of unexpected content without a clear schema or structure [5].

3.3 Proposed Improvements to Overcome Hadoop Problems

Since many issues have surfaced since Hadoop was developed, users in big organizations complain that Hadoop is sometimes slow [34]. In real data manipulation, data is always moving. Big storage and particular technologies are necessary. Unique techniques could provide potential improvements for efficiency of Hadoop frame-work. In the following section we introduce some recommendations.

3.3.1 Keeping big data in movement

In the real-time data manipulating, to accomplish higher efficiency, the system must make ensure that all request messages are processed without a critical processing path [34]. To perform this operation, the system must use special functions to handle data processing. They introduce PCF and DCF into MapReduce function of Hadoop. This hybrid method can efficiently raise data transfer speed. The wireless MAC sub-layer includes two basic access methods:

- distributed coordination function (DCF) and
- The point coordination function (PCF).

DCF uses carrier-sensing multiple access with collision avoidance (CSMA/CA) approach [34]. The PCF relies on polling to determine which can transmit next. According to these methods, we can use same idea with data manipulating. According to these methods, we can use a similar idea with data processing. While transferring data from different locations, we can add DCF and PCF between the MapReduce function and the sites. For example, for DCF, it uses the CSMA/CA policy. During sending, when one middleman requests, it will send a Request to Send (RTS) to MapReduce, it will send Clear to Send (CTS) back for confirmation, then they can connect successfully. When other sites see CTS, they will stop trying to send the request. For PCF, point coordinator (PC) is an important role in the whole

process. PCF requires a computer to control all the transmission. During a special time period, the computer will ask all sites to set their requirements and create a request list. In this polling handling it can create an additional overhead for systems. However, it will avoid unexpected collisions and retreat. We will see the architecture for PCF in Fig. 7. A computer plays a controller and controls all different points, and then make a job list to processing.

Therefore, when the big number of data manipulating is involved into polling, the data manipulating is polled in sequence from the polling list maintained by the computer, which avoids collision and produces more efficient process.

3.3.2 Massive parallel operations

With great computing power, MapReduce uses one single I/O operations [34]. Like MPP (Massive Parallel Processing), we must distribute the MapReduce function in different locations connected to a switch. In the switch, it should retain an optimizer that can be used for data mining. By using this hybrid method, you will improve your handling efficiently. Moreover, in real time system waiting for process is not a good idea so it needs time restrictions. While moving many queues (data processing), the system can split the queues into different sectors and process them using different parameters, as shown in Fig. 8. This method can save time when submitting many data processing requests simultaneously.

3.4 Hadoop Tools for Handling Big Data

There are many tools that help in achieving these goals and help data scientists to process data for analyzing them. Many new languages, frameworks and data storage technologies have emerged that supports handling of big data.

3.4.1 Hadoop distributed file system (HDFS)

Hadoop includes a Hadoop Distributed File System or HDFS a fault tolerant storage system [29]. HDFS which stands for Hadoop Distributed File System that based on the Google File System (GFS) and provides a distributed file system designed to work on clusters on commodity hardware devices [35]. HDFS is the Java based distributed file system used by Hadoop [36]. It consists of Blocks, Name node (master) and the Data node (slave). A block is the minimum amount of data that can read or write. The default size of HDFS blocks are 128Mb.Files stored in the HDFS are split in to multiple blocks called Chunks that are independent of each other; for example, If the file size is 50 Mb then the HDFS blocks takes only 50Mb of memory space within in the default 128 Mb [36,37,38,39,40,41,42].The name node is responsible for storing the Metadata which means it contains all the information about which shelf the data nodes data are stored. It contains the directory and location of data. Data nodes contain actual user data. On one hadoop cluster, there is only one name node and multiple numbers of data nodes present. Fig. 9 illustrates HDFS structure [36]. HDFS is data storage infrastructure without losing large chunks of failure, massive amounts of information stored in the scale are progressively viable. Clusters can be made using inexpensive computers. One fails, the remaining devices in the cluster running on Hadoop when going from work without data lose or interruption continue to work. HDFS, a cluster known as" splitting files and servers into redundant parts across the cluster to store blocks from each storage group management. In normal case, three different HDFS servers are stored by copying each piece three complete copies of each file. The reason for large size of files or blocks is to reduce the number of disk seeks [43]. The server keeps a small block of the whole data set [30]. Hadoop is an open source framework that allows the storage and processing of big data in a distributed environment across groups of computers using simple programming models. It is designed to scale from single servers to thousands of devices, each offering local computation and storage [35].

HDFS Advantages [36]:

- It has very high bandwidth to support map reduce jobs.
- It is very less expensive.
- We can write the data once and read many times.

HDFS Disadvantages [36]:

- Cluster management is hard.
- The process of join multiple data base is slow and difficult.

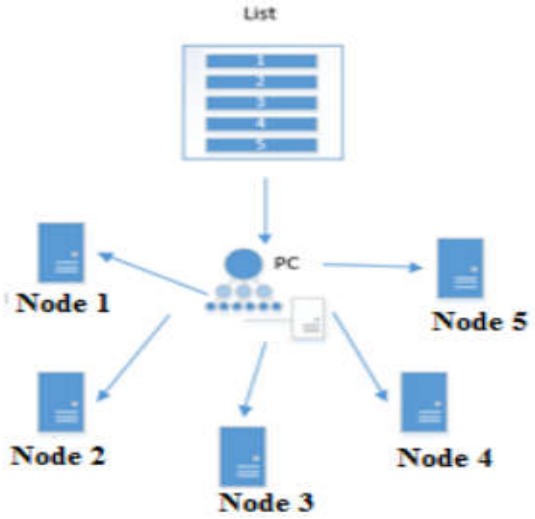

Fig. 7. Proposed architecture of PCF for big data

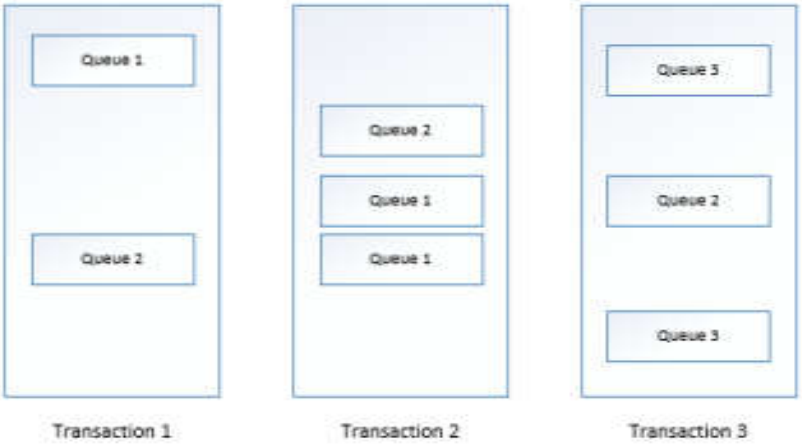

Fig. 8. Segregation of data processing among transactions and queues

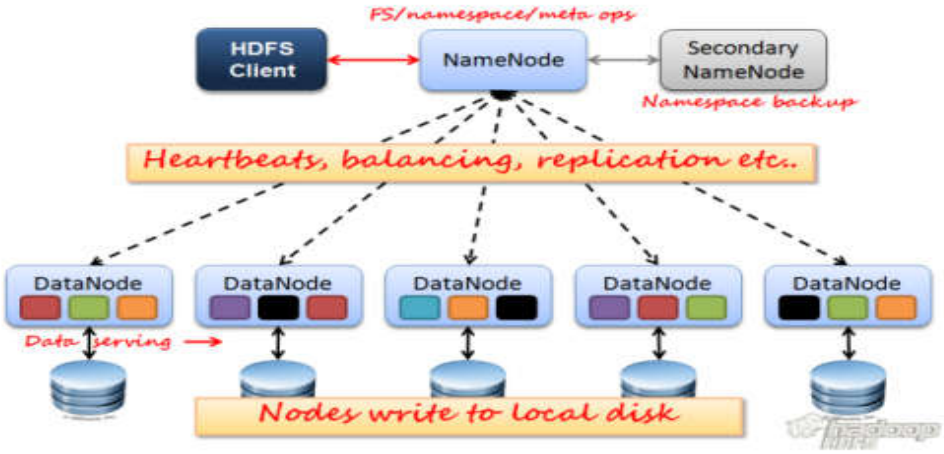

Fig. 9. HDFS structure

3.4.2 Hadoop MapReduce (programming paradigm)

MapReduce is a software framework for processing pole of hadoop. The processing process is applied on huge divided amounts of data that run parallel in a reliable, fault-tolerant manner. The two different phases of MapReduce are :

Map Phase: this phase, the workload is divided into smaller sub-workloads [5]. The tasks assigned to the map or mapper's job which assigned the divided workload (smaller sub-workloads "input") stored in HDFS to Mapper function line by line , which processes each unit block of data to produce a sorted list of (key , value) pairs. The mapper produces output list and passed it to the next phase. This process of mapper is known as shuffling as explained in Fig. 10 [30].

Reduce phase: In this phase, the reducer job is to process the input data that comes from the mapper by analyzing and merging it to produce the final output which is written to the HDFS in the cluster .Some other programming models such as Spark [44,45] and DataMPI [46] are competing with MapReduce. Table 3 summarizes the big data capabilities and the available primary technologies [5].Since MapReduce is an open source with high performance which is used by many big companies for processing batch jobs [47,48].

MapReduce Advantages [36]:

- It supports wide range of language Java.
- It is a platform individualistic.

MapReduce Disadvantage [36]:

- It is applicable only for batch oriented process.
- Does not apply to interactive analysis.
- Does not work with intensive algorithm, learning and graphing.
- To overcome the limitation of Hadoop 1.0 move to hadoop 2.0

3.4.3 Hadoop YARN

YARN (Yet Another Resource Negotiator) [36]. In hadoop YARN, the multiple name node server manages the entire namespace in the hadoop cluster. YARN is the heart of the hadoop 2.0 OS. It consists of four components such as:

1) Client,
2) Resource Manager,
3) Node Manager and
4) Map Reduce Application Master.

Client submits their job to the resource manager for processing [36,37,38,49].

Framework for scheduling jobs and managing cluster resource [27]. YARN was developed to fix MapReduce limitations of. YARN has divided the main functions of MapReduce1 into two components related to resource management, job scheduling and monitoring via two separate daemons: Resource Manager and Application Master. This new architecture has solved the limitations of Map-Reduce 1. Fig. 11 illustrates moving from Hadoop 1 to Hadoop 2 [36].

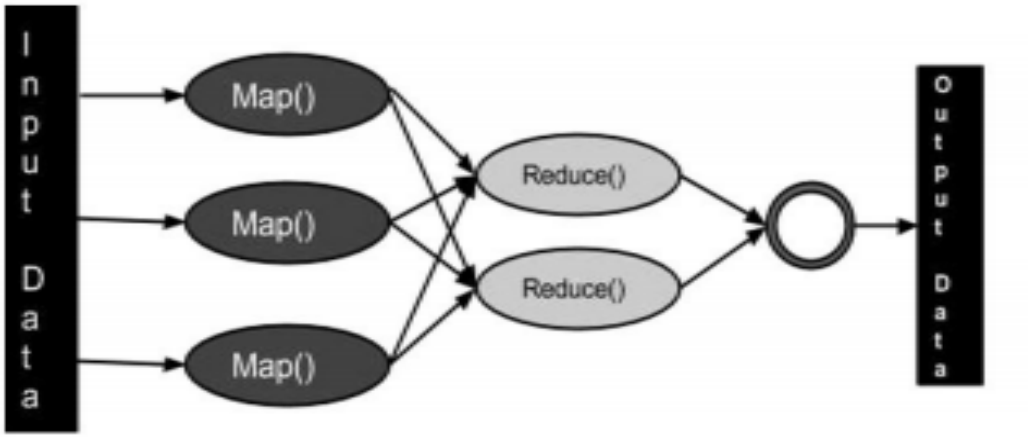

Fig. 10. MapReduce architecture

Table 3. Big data capabilities and their primary technologies

BigData Capability	Primary Technology	Features
Storageand management capability	Hadoop Distributed File System (HDFS)	Open source distributed file system, Runs on high performance commodity hardware, Highly scalable storage and automatic data replication
Database capability	Oracle NoSQL	Dynamic and flexible schema design, Highly scalable multi-node, multiple data center, fault tolerant, ACID operations, High- performance key-value pair database
	Apache HBase	Automatic failover support between Region servers , Automatic and configurable sharding of tables
	Apache Cassandra	Fault tolerance capability for every node, Column indexes with the performance of log-structured updates and built-in caching
	Apache Hive	Query execution via MapReduce, Uses SQL-like language HiveQL, Easy ETL process either from HDFS or Apache HBase
Processing capability	MapReduce	Distribution of data workloads across thousands of nodes ,Breaks problem into smaller sub-problems
	Apache Hadoop	Highly customizable infrastructure, Highly scalable parallel batch processing, Fault tolerant
Data integration capability	Oracle big data connectors, Oracle data integrator	Exports MapReduce results to RDBMS, Hadoop, and other targets, Includes a Graphical User Interface
Statistical analysis capability	R and Oracle R Enterprise	Programming language for statistical analysis

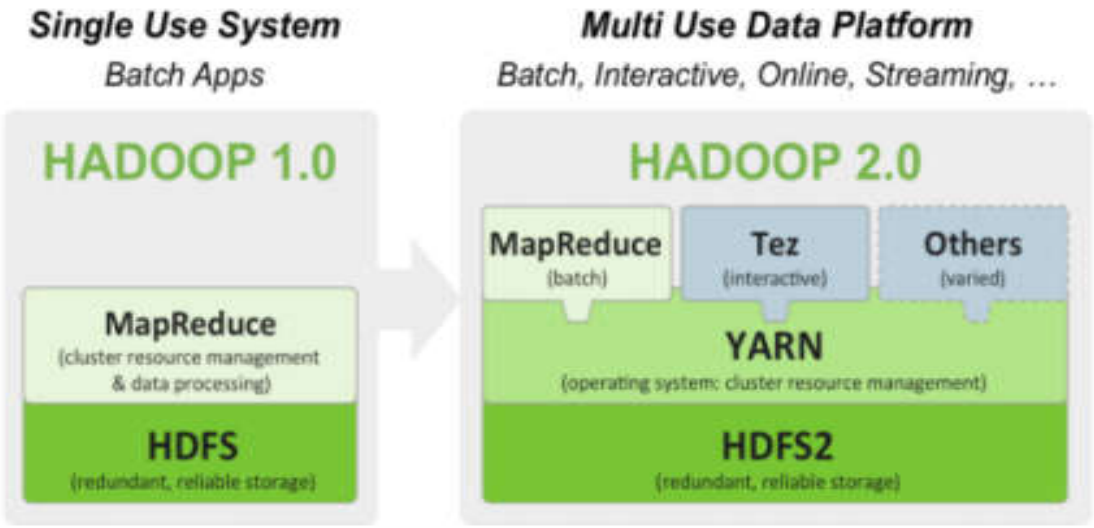

Fig. 11. Hadoop 1.0 To Hadoop 2.0

Is considered as the resource manager that manages all the resources through the cluster. The node manager is responsible for maintaining the Meta data information, monitors and keep track of user application. Map reduce application master is responsible for executing the application. It makes more interactive to multiple applications, manages all the resources management, job scheduling, provide security controls and high availability of data. Here it can execute the both non- map reduce and map reduce application.

YARN Advantages [36]:

- It provides efficient utilization of resources.
- It can run the application which doesn't follow.
- It supplies high data accessibility.

3.4.4 Hive hadoop

Hive is an open source platform that introduces facilities for querying and managing large datasets in distributed storage [5]. Also, Hive could be defined as data warehousing software that addresses how data is structured and queried in distributed Hadoop [30]. Hive is also a popular development environment used to write queries for data in the Hadoop environment. Hive is a annotation language used to develop applications for the Hadoop environment; however, it does not stands for real-time queries. Hive is a technology developed at Facebook that turns Hadoop into a full SQL query dialect data warehouse. Being a dialect of SQL, Hive is a declarative language. You define the data flow, but in Hive we describe the result we want and and Hive figure out how to build the data flow to achieve that result. Unlike Pig, a chart is required in the cell, but you are not limited to just one schema (chart). Like PigLatin and the SQL, Hive itself is a complete relational language but not a complete Turing language. It can also be extended by UDFs just like Piglatin to be a complete Turing. Hive is a technology for turning Hadoop into a data warehouse, complete with SQL dialect for querying. Hive works in terms of tables. There are two types of tables you can create: managed tables whose data is managed by Hive and external tables whose data is managed outside of Hive.

Hive Advantages [36]:

- The users need not to write their jobs in to Map Reduce programs.
- It provides tools for easy data extraction, conversion and loading of data from the big data ware house.
- It offers infrastructure for storing the data.
- Any SQL developer can easily write hive query.
- It can be integrated with Hbase for easy query and data retrieval.
- Map reduces programs that need more lines of code than HiveQL.

Hive Disadvantages [36]:

- It doesn't support for processing unstructured data.
- Complex jobs cannot be executed with Hive.
- The output of one job can to be used to query the input for other jobs.

3.4.5 Pig (programming tool)

Pig is a platform that allows analysts to analyzing large data sets [5]. Pig could be defined also as procedural language for developing parallel processing applications for large data sets in the Hadoop environment [30]. Pig is an alternative to MapReduce, and automatically generates MapReduce functions. Pig includes Pig Latin, which is a textual language. Pig translates Pig's Latin texts into MapReduce. Pig consists of both a language and an implementation environment. Pig, called PigLatin, is a data flow language - this is the kind of language that you program with by linking things together. Pig can operate with complex data structures, even those that can contain levels of nesting. Unlike SQL, Pig does not require data to have a chart or schema, so it is well suited for manipulating unstructured data. But, Pig can still tap into the value of the schema if you want to save on it. PigLatin is relatively complete as SQL, which means it is at least as powerful as Relational algebra. Turing completion requires conditional constructs, an infinite memory model, and loop combinations. PigLatin is not complete Turing in itself, but it can be when extended with a User defined Function.

Pig Advantages [36]:

- It decreases the Duplication of data.
- It reduces the number of lines of code and save the development time.
- The user defined functions can be easily programmed for read and write operations.
- It aids nested data models.
- The programmer who knows SQL language can easily able to learn and write pig scripts. Disadvantages:

Pig Disadvantages [36]:

- It doesn't provide JDBC and ODBC connectivity.
- There is no committed Metadata database.
- It doesn't offer web interface.

3.4.6 HBase

Base is a scalable and distributed database that supports structured data storage for large tables [50]. It was designed to store structured data in tables that could have many of rows and many of columns [30]. HBase is not a relational database and wasn't designed to stands for both transactional and other real-time applications. Apache HBase is distributed column based database like layer built on Hadoop designed to supporting billions of messages per day, HBase is highly scalable and offers fast random write operations as well as random and streaming readouts. It also provides atomic guarantees at the grade level, but there isn't support for transactions across the parent grades. From a data model perspective, column orientation provides maximum flexibility in storing data and broad rows allow the creation of billions of indexed values in one table. HBase is ideal for workloads that require a lot of writing and need to hold large amounts indicators and maintain the flexibility to scale quickly.

HBase Advantages:

- It is highly Fault Tolerant.
- It provides low latency access to small amounts of data from within a large data set.
- It is highly flexible data model.
- Strongly consistent

HBase Disadvantages:

- It cannot be applicable to complicated data access patterns (Such as joins).
- It is not applicable for transactional applications or relational analytics.

3.4.7 R

R is a free software package for statistics and data visualization [51]. It is available for UNIX, Windows and MacOS systems and is the result of the work of many programmers from all over the world. R contains facilities for data handling, provides high performance procedures for matrix computations, large set of tools for data analysis, graphical functions for data visualization and a straightforward programming language. R comes with around 25 standard packages and many others available for download through the CRAN website family of Internet sites (http://CRAN.R-project.org). R is used as a computational platform for regular statistics production in many official statistics agencies and could be an open-source statistical computing language that provides a wide variety of statistical and graphical techniques to derive insights from the data [5]. Besides official statistics, it is used in many other sectors like finance, retail, manufacturing, academic research etc., making it a popular tool among statisticians and researchers. Is an open-source statistical computing language that introduces a wide variety of statistical and graphical methods for extracting insights from data [5]. It has an efficient method for processing and storing data and supports vector operations with a set of operators for faster processing. It contains all the features of a standard programming language and supports conditional arguments, loops, and user-defined functions. R is supported by a large number of packages through the Comprehensive R Archive Network (CRAN). It has robust documentation for every package. It has strong support for data settings, data mining and machine learning algorithms along with a good support for reading and writing in a distributed environment, making it suitable for handling big data. However, memory management, speed, and efficiency are probably the biggest challenge faced by R, R Studio is an IDE developed for programming in the R language. It is distributed for standalone Desktop machines as well as it supports client-server architecture, which can be accessed from any browser.

3.4.8 Python

Is a popular programming language, it is open source and is supported by Windows, Linux and Mac platforms [5]. It hosts thousands of modules packages contributed by a third-party or community. NumPy, Scikit, and Pandas support some popular packages for machine learning and data mining for data preprocessing, computing and modeling. It adds support for large multidimensional arrays and matrices with Python. Scikit supports classification, regression, clustering, dimensional reduction, feature selection, preprocessing and model selection algorithms. Pandas help with data managing and preparation for data analysis and modeling. It has strong graph support with NetworkX and nltk library for text analytics and Natural language processing. Python is extremely easy to use and great for quick and dirty analysis of a problem. It also integrates well with spark with the pyspark library.

3.4.9 Scala

Is an object-oriented language and has an acronym for "Scalable Language" [5]. Object and every method in Scala is a method-call, just like any object-oriented language. It needs java virtual machine domain. Spark, the in-memory cluster computing framework is written in Scala. Scala has becoming popular programming tool for dealing with big data problems.

3.4.10 Apache spark

Is an in-memory cluster computing technology designed for quick processing, which is implemented in Scala [5]. It is a memory-centric computational engine framework that is fast and scalable [27]. Since it's memory-driven, it assumes MapReduce as the main framework for working with Big Data. Spark includes build-in libraries that support ETL, stream processing, machine learning, and graph computation, SQL and Data Frames. Spark is designed to extend the MapReduce model. Spark is faster because of in-memory arithmetic computation, and faster than MapReduce for complex application on disk. Besides, Spark can be used for a wide range of workloads (batch application, iterative algorithms, interactive queries, streaming data), and it is easy to use because it supports wide range of APIs for Scala, Python, Java, and R [27]. Most recently, it also started supporting R. It comes with 80 high-level interactive query operators. In-memory computation is supported by the Resilient Distributed Data (RDD) framework, which distributes the data frame into smaller pieces on different machine devices for faster computation [5]. It also supports both Map and Reduce for data manipulating. It supports SQL, data flow, graph processing algorithms and machine learning algorithms. Although Spark can be accessed with Python, Java, and R, it has a strong support for Scala and is much more stable at this time.

3.4.11 Amazon elastic compute cloud (EC2)

Is a web service that provides compute capacity over the cloud. It gives complete control over computing resources and allows developers to run their manipulation in the desired computing environment [5]. It is one of the most successful cloud computing platforms. It works on the basis of the pay-as-you-go prototype. Few other structers that assist big data are MongoDB, BlinkDB, Tachyon, Cassandra, CouchDB, Clojure, Tableau, Splunk and others.

3.4.12 Ambari

A web-based tool for provisioning, managing, and monitoring Hadoop clusters [27]. It also provides a dashboard to display cluster health and the ability to visually display MapReduce, Pig and Hive apps. Apache Ambari provides easy to use RESTful APIs which allows application developers to easily integrate Ambari with their own applications.

3.4.13 Mahout

A Scalable data mining and machine learning library essentially focuses on classification; clustering and batch based collaborative filtering (recommendation) [27]. The Aim of Mahout is to find insights out from big data stored in Hadoop. Mahout works well with spread environment and can scale productively in the cloud. It includes several MapReduce enabled clustering implementation like K-means and Canopy [27].

Mahout Advantages [36]:

- It supports supplementary and distributed naive bayes classification.
- It extracts an enormous amount of data.
- The company's such as Adobe, Twitter, Foursquare, Face book and LinkedIn internally uses mahout for data mining.
- Yahoo is specially used for pattern mining.

Mahout Disadvantages [36]:

- It doesn't support scala version in the development.
- Does not contain a decision tree algorithm.

3.4.14 Avro

A data serialization efficient binary format that facilitates interoperability with applications of different programming languages due to binary encoding that can be used for long term storage in Hadoop [27]. Avro uses a human readable JSON text file format to define data types and protocols because it supports the transfer of data objects in the form: attribute-value. Avro supports versioning of MapReduce that handles field addition and deletion of forward and backward alignment [27].

3.4.15 ZooKeeper

A centralized coordinator for reliable distributed applications coordination in Hadoop ecosystem. ZooKeeper is concerned with maintaining configuration information, naming, spread synchronization, handling partial failures that are inevitable in spread environment, and general group service [27].

ZooKeeper Advantages [36]:

- It provides reliability and availability of data.
- It provides high concurrency and serialization.
- Atomicity removes data inconsistencies between clusters.
- It is fast and simple.

ZooKeeper Disadvantages [36]:

- The large number of stacks needs to be maintained.

3.4.16 Oozie

Is a Java-based workflow server-based system that coordinates, manages, and process Hadoop jobs. Oozie represents workflows as control and action flows as nodes via Directed Acyclic Graphs (DAG) [27].

Oozie Advantages [36]:

- It allows the workflow of execution can be restarted from the failure.
- Availability of an API for the web service (meaning we can control jobs from anywhere).

Oozie Disadvantages [36]:

- It is not a resource scheduler.
- Not suitable for off-grid planning.

3.4.17 Sqoop

Is a software tool designed to transfer bulk data between Hadoob and relational database. Sqoop is used to import data from external database into HDFS or HBASE or HIVE [50]. It allows for data import from and data export to external relational database and parallel data transferring. Uses a simple SQL query as well as saved functionality that can run multiple times to import data-related updates regarding the data between Hadoop and relational database. Fig. 12 represents Sqoop transform.

Sqoop Advantages [36]:

- It offers the migration of heterogeneous data.
- It offers easy integration with Hive,HBase and oozie.
- We can import the whole data base or the single table in to HDFS.

3.4.18 Cassandra

A scalable multi-head database with no single points of failure [50]. Apache Cassandra is a high availability, highly scalable and high performance open source distribute database management system having capability of handling hugs amount of data across multiple servers. It provides for tolerance and is decentralized.

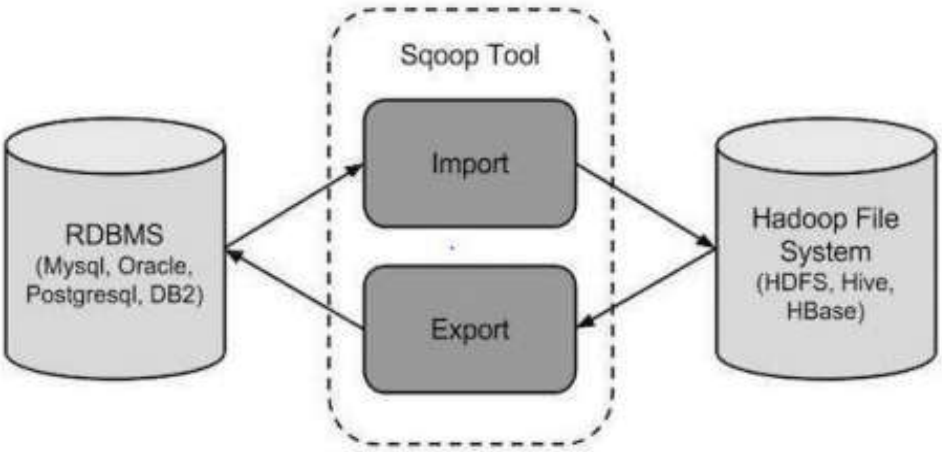

Fig. 12. Sqoop transformation

Cassandra Advantages [36]:

- The biggest companies such as Face book, Twitter, Rack space and Cisco uses the Cassandra to handle their data.
- It contains Dynamo-Style Replication model to supply no one point of failure.
- It provides high throughput and quick response time if the number of nodes in the cluster increases.
- Transaction ACID is supported.
- It is a column directed data base.

Cassandra Disadvantages [36]:

- It does not support sub query and join operation.
- Limited support for data collection.
- Limited storage space for one column value.

3.4.19 Tez

A generalized data flow programming framework, which provides a robust and flexible engine for arbitrary DAG implementation of tasks to process data for both batch and interacting use cases [50]. Tez is certified by Hive, pig and other Hadoop environment frameworks and also by other commercial software (e.g.HTL tools), to replace Hadoop MapReduce as the underlying execution engine.

3.4.20 Flume

Is a reliable distributed service for efficiently collecting aggregating and moving large amount of LOG data [50]. It helps the users to make the most of valuable log data. It allows streaming of data from multiple sources, and real-time bulk web log collection.

Table 4 compares the features of MapReduce, Spark, and DataMPI. MapReduce has been evolved from MapReduce1to MapReduce2 or the so-called YARN which concerns with scheduling and cluster resource management and covers the limitations of the first release [27,52].

Table 4. Comparing MapReduce, spark, and data MPI

Features	Hadoop	Spark	HBase	Hive	Pig	MapReduce
Data processing	Hadoop is a batch process-ing system.	Spark is micro-batch processing and system.	HBase is used to store data into a column-oriented database.	Hive is used for data sumarization , query, and analysis.	Pig is a tool used for analyzing of huge data sets.	Map Reduce is a tool which is suitable for parallel proc-essing of huge data.
Streaming engine.	Hadoop takes large data sets as input, proc-esses it and produces the output.	Spark streams data in micro-batches.	The batch load is opt-imized to run on the Spark exec-ution eng-ine	Hive streaming provides for Soft-ware based enterprise con-tent deli-very that is done behind fire-wall for effi-ciency and security.	Pig provides a parallel arch-itecture orient-ed streaming engine that can update Hadoop data over small portions.	Map Reduce is a type of native batch processing en-gine.
Data flow	Hadoop is a chain of stages.	Spark repre-sents a data flow in a form of a direct acy-clic graph (DAG).	HBase is run on top of HDFS and it stor-es data in the key / value form.	Data flow in Hive beha-vees at the qucry execution level right from the UI. Meta store sends	The Pig is used to analyze larger sets of data that represents them as data flows.	Map Reduce is a distributed programming model that was designed for processing of large volumes of datasets in parallel such

Features	Hadoop	Spark	HBase	Hive	Pig	MapReduce
				metada-ta info back to the compiler.		that it is in-dependent of sub work.
Scalability	Hadoop provides scalable flexible data storage and analysis.	Spark rovides linear scalabilityin the istribut-ed environ-ment.	HBase pro-vides xtreme scalability, reliability, and flexib-ility.	Hive is much familiar, fast, scalable and extensible.	Pig provides high level scability.	Map Reduce provides scal-ability means that single ser-ver to thousa-nds of machines.
Latency	Hadoop gives higher latency than both Spark and Fink.	Spark gives low latency than Hadoop.	HBase is fast and used for low aten-cy data access. It stores data in an in memory table called aMem-Store.	Hive has high latency as compared to HBase.	Pig is streaming writes, just like Map Reduce. Low latency queries are not suported in Pig, thus it is not suitable for OLTP.	Map reduce gives low latency.
Cost	A midrange Intel server is recommended an enterprise-class Hadoop cluster.	Spark is very costly	It depends on your us age pattern,S3 listing and file transfer might cost money.	Hive is also open source, and built on top of Hadoop for data querying.	Pig is lower in cost to write and maintain compared to Map Reduce.	Map reduce Cost is high but Hadoop cluster, a mid-range Intel server is reco-mmended.
Development	Hadoop is developed Apache Soft-ware Foundat-ion.	Spark is dev-eloped in the Uni-versity of california after some time it's code base	HBaseis an open source project that was incubated by Apache Software Foundation.	Hive was init-ially developped by Face-book, but also some other companies develop and use it.	Pig is origina-lly developed by Yahoo & Facebook.	MapReduce is developed by Google for a new style of data process-sing.
Scheduler	Hadoop provides fair scheduler and Capacity scheduler they are two type sch-eduler in Ha-doop. The scheduler in Hadoop becomes the plug-gable component.	Spark acts as its own flow scheduler due to in-memory computation.	HBase scheduler uses a polling to change state at regular int-ervals. Also, if required based on configuratio-n it can trigger jobs.	Hive schedules table every hour by use of Oozie schedule.	Oozieis the tool for work-flow scheduler in Hadoop for Apache Pig – Secondly, write a brief Pig script for each data file to extract the req-uired data fields.	Map Reduce provides the Fair Scheduler, which provides a way to share large clusters.

3.5 Hadoop Applications

3.5.1 Amazon

To build search pointers for Amazon products; process millions of sessions per day for analytics, using both the Java and streaming APIs; clusters vary from 1 to 100 nodes [30].

3.5.2 Yahoo!

More than 100,000 CPUs in about 20,000 Hadoop computers; the largest cluster: 2000 nodes (2*4cpu boxes with 4TB disk each); used to support research for Ad Systems and Web Search [30].

3.5.3 Facebook

To store copies of the internal registry and dimension data sources and use them as the source for reporting/analytics and machine learning; 320 machine cluster with 2,560 cores and about 1.3 PB raw storage [30].

3.5.4 Google

Apache Hadoop [53,54] is an open source implementation of the Google's MapReduce [55] parallel processing framework.

3.5.5 Twiter

Twitter runs multiple large Hadoop clusters that are among the biggest in the world. Hadoop is at the core of our data platform and provides vast storage for analytics of user actions on Twitter.

3.5.6 LinkedIn

LinkedIn relies heavily on Hadoop especially for offline data infrastructure needs.

New Yourk Times software engineer talks about how Hadoop is driving business innovation at the newspaper and Web site

3.5.7 AOL

Is an American web portal and online service provider based in New York City. It is a brand marketed by Verizon Media. It is used for running an application that analyzes the behavioral pattern of their users so as to offer targeted services [53].

4. HADOOP IMPROVEMENTS USING SCHEDULING

Hadoop-MapReduce has become a powerful algorithm for processing big data on distributed commodity device clusters like Clouds [53]. Through our deep study of researches that study Hadoop performance enhancement, we found that there are many factors can affect Hadoop performance as explained in next sub-section. Between different researches that intended to enhance Hadoop performance, we chosen to make our improvements through Scheduling Policy because it can be considered as promising policy. In all Hadoop applications, the default FIFO scheduler is available where jobs are scheduled in FIFO order with support for other priority-based schedulers as well.

4.1 Scheduling in Hadoop

There are various factors affecting the performance of scheduling policies like data volume (storage), data source format (various data), speed (data rate), security and privacy, cost, connectivity and data sharing [56]. Ability to make Hadoop scheduler resource aware is one the emerging research problem that attracts the attention of most of the researchers because the current implementation relies on statically configured slots [53]. Each Scheduler takes into account resources like CPU, Memory, Job deadlines and IO etc. The default Scheduling algorithm is based on FIFO as jobs were executed in the order they were sent. Later the ability to set the priority of the Job was added [53]. Facebook and Yahoo have both contributed a lot of work developing scheduling software i.e. Fair Scheduler [57] and Capacity Scheduler [58] respectively which subsequently released to Hadoop Community. Several researchers are working on opportunities to enhance Hadoop's scheduling policies. Recent efforts such as Delay scheduler [59], Dynamic Proportional Scheduler [60] offer differentiated service for Hadoop jobs allowing users to adjust the priority levels assigned to their jobs. However, this is no guarantee that the job will be completed by a specific deadline. Deadline Constraint Scheduler [61] addresses the issue of deadlines but focuses more on increasing system utilization. The Schedulers described above attempt to allocate capacity fairly between users and jobs, and make no attempt to consider resource availability on a more precise basis. Resource Aware Scheduler [62] considers the resource availability to schedule jobs. All the previous

schedulers address one or more problem(s) in scheduling in homogeneous Hadoop clusters. Other new work considers scheduling in Hadoop in Heterogeneous Clusters.

4.2 Proposed Scheduling Algorithm in Hadoop Heterogeneous Clusters

Heterogeneity in Hadoop systems is a substantial challenge in scheduling algorithms. This topic has not gained much awareness from researchers. We proposed to choose a promising algorithm that enhances performance of scheduling policies used in Heterogeneous Clusters. Using this algorithm, we are able to reduce the complexity of estimating the execution time for heterogeneous systems by considering the base unit. This algorithm has been presented to run on heterogeneous Hadoop clusters and runs job in parallel [56]. This algorithm first distributes data based on the performance of the nodes and then schedules the jobs according to their cost to perform and reduces the cost of executing the jobs. The chosen proposed algorithm achieves better performance in terms of execution time, cost and location compared to FIFO and Fair schedulers. They carried out part of the scheduling operation on the users' system. Executing some scheduling components on users' system can decrease the workload of name nodes.

4.3 Experimental Results for Chosen Proposed Algorithm

Fig. 13 presents the average job execution time for the algorithms based on the type of jobs in a real Hadoop system [56]. According to the obtained results, the chosen proposed algorithm has on average executed the jobs 2.29 times faster than the FIFO algorithm and 2.12 times faster than the Fair algorithm. Fig. 14 presents the average cost, and Fig. 15 presents the locality of the algorithms. Considering cost while assigning the jobs to the nodes and distributing the data of the jobs based on the performance of the nodes in the proposed algorithm has led to the creation of a higher performance difference. The chosen algorithm offers better performance for all three types of jobs compared to the FIFO algorithm and the Fair algorithm.

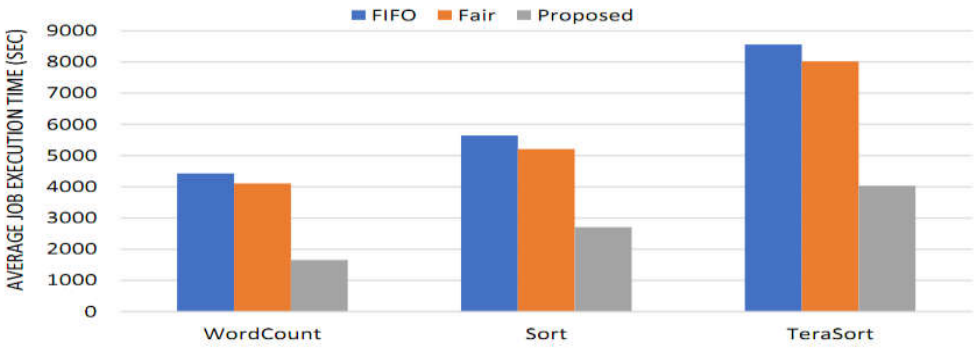

Fig. 13. Average job execution time for micro benchmarks

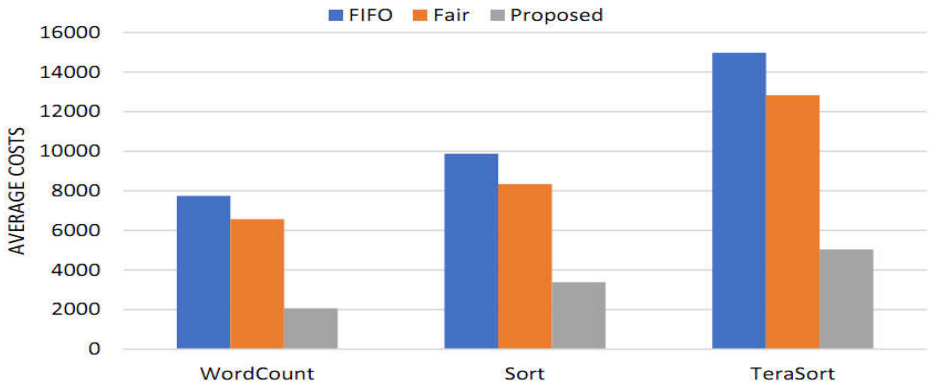

Fig. 14. Average costs for micro benchmarks

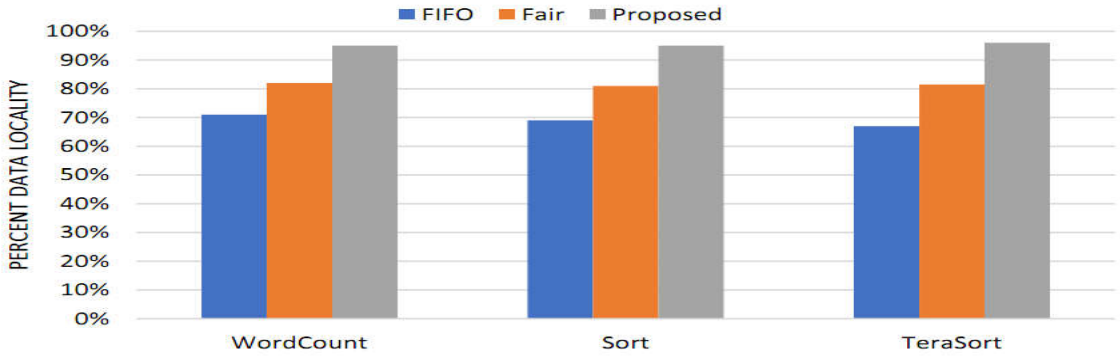

Fig. 15. Percent data locality for micro benchmarks

5. CONCLUSION

Managing the data is the big issue. Now an enormous amount of data is produced in the creation process so, the concept of big data occupies a large place in the scene. For data management, big data technology used i.e.Hadoop. Hadoop can handle the huge amount of data with it's very effective cost, can also handle these ocean amount of data with very fast processing speed, and it can create a double copy of data in case of system failure or to prevent the loss of data. Although all these advantages of Hadoop, it suffers from some problems that we introduced in section 4. There are various factors affecting the performance of scheduling policies like data volume (storage), data source format (various data), speed (data rate), security and privacy, cost, connectivity and data sharing. Between different researches that intended to enhance Hadoop performance, we chosen to make our improvements through Scheduling Policy because it can be considered as promising policy. Ability to make Hadoop scheduler resource aware is one the emerging research problem that attracts the attention of most of the researchers because the current implementation relies on statically configured slots. All the previous schedulers address one or more problem(s) in scheduling in homogeneous Hadoop clusters. Other new work considers scheduling in Hadoop in Heterogeneous Clusters. We proposed to choose a promising algorithm that enhances performance of scheduling policies used in Heterogeneous Clusters. Using this algorithm, we are able to reduce the complexity of estimating the execution time for heterogeneous systems by considering the base unit. This algorithm has been presented to run on heterogeneous Hadoop clusters and runs job in parallel. According to the obtained results, the chosen proposed algorithm has on average executed the jobs 2.29 times faster than the FIFO algorithm and 2.12 times faster than the Fair algorithm (used by Facebook). This chosen Algorithm achieved promised important goal of enhancing Hadoop performance which we aim for with full force.

6. RESEARCH LIMITATIONS AND FUTURE WORK

This paper (BD3) is considered completion in our series of big data papers. We started our series with paper titled," How Many Old and New Big Data V's Characteristics, Processing Technology, and Applications (BD1)" [11]. Second paper in this series is titled as, "Big Data Fifty Six V's Characteristics and Proposed Strategies to Overcome Security and Privacy Challenges (BD2)" [12]. Our fourth paper that is in Progress is titled, "Blockchain-based Trusted Distributed System for Big Data and Proposed Solution approaches to Overcome Blockchain Problems (BD4)".

DISCLAIMER

The products used for this research are commonly and predominantly use products in our area of research and country. There is absolutely no conflict of interest between the authors and producers of the products because we do not intend to use these products as an avenue for any litigation but for the advancement of knowledge. Also, the research was not funded by the producing company rather it was funded by personal efforts of the authors.

COMPETING INTERESTS

Author has declared that no competing interests exist.

REFERENCES

1. Sitalakshmi Venkatraman, Ramanathan Venkatraman. Big data security challenges and strategies. AIMS Mathematics. 2019; 4(3):860–879
   DOI: 10.3934/math. 2019.3.860
   Accepted: 01 July 2019, Published: 19 July 2019.
2. McNeely CL, Hahm J. The big (data) bang: Policy, prospects, and challenges. Review of Policy Research. 2014;31:304–310.
3. Gandomi A, Haider M. Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management. 2015;35:137–144.
4. Lawal Muhammad Aminu. Implementing big data management on grid computing environment. International Journal of Advanced Trends in Computer Science and Engineering; 2014.
5. Abhay Kumar Bhadani, Dhanya Jothimani. "Big data: Challenges, opportunities, and realities", Effective big data management and opportunities for implementation, Pennsylvania, USA, IGI Global. 2016;1-24.
6. Chandhini C, Megana LP. Grid computing-a next level challenge with big data. International Journal of Scientific & Engineering Research. 2013;4(3).
7. Bhadani A, Chaudhary S. Performance evaluation of web servers using central load balancing policy over virtual machines on cloud, Proceedings of the Third Annual ACM, Conference, Bangalore, ACM; 2010.
8. Bhadani A. Cloud computing and virtualization. Saarbrucken: VDM Verlag Dr. Muller Aktiengesellschaft & Co. KG. 116 s; 2011.
   ISBN: 9783639347777.
9. Assunção MD, Calheiros RN, Bianchi S, Netto MAS, Buyya R. Big data computing and clouds: Trends and future directions. Journal of Parallel and Distributed Computing. 2015;79–80:3-15.
10. Alexandru Adrian TOLE. Big data challenges. Database Systems Journal. 2013;4(3).
11. Abou_el_ela Abdou Hussien. How many old and new big data V's characteristics, processing technology, and applications (BD1). International Journal of Application or Innovation in Engineering & Manage-Ment. 20209(9).
    Available:www.ijaiem.org, editor@ijaiem.org
12. Abou_el_ela Abdou Hussien. Big data fifty six V's characteristics and proposed strategies to overcome security and privacy challenges (BD2). Journal of Information Security. 2020;11(4).
13. Mircea Raducu Trifu, Mihaela laura Ivan Big data: Present and future. Database Systems Journal. 2014;1(1).
14. George Firican. The 10 V'S big data. Work Paper; 2017.
15. Kirk Borne. Top 10 big data challenges-a serious look at 10 big data V's. blog post; 2014.
16. William Vorhies, View Blog. How many V'S in big data. Work Paper; 2014.
17. Dhamodharavadhani S, Gowri Rajasekaran. Unlock different V's of big data for analytics. International Journal of Computer Sciences and Engineering, Open Access Research Paper 2018;6(4).
18. Darrin; 2016.
    Available:https://educationalresearchtechniques.wordpress.com /2016/05/02/characteristics-of-big-data/
19. Gartner; 2013
    Available:http://www.forbes.com/sites/gartnergroup/2013/03/27/ gartners -big-data-definiti-consists-of-three-parts-not-to-be-confused-with-three-vs/#5040d1cc 3bf6
20. 2012. Available:https://hbr.org/2012/10/making-advanced-analytics-work-for-you/ar /1
21. Good Strat Tweetl; 2015.
    Available:http://www.informationweek.com /big-data/big-data-analytics/big-data-avoid-wanna-v-confusion/d/d-id/1111077?page_number=1
22. Laney D.; 2012
    Available:http://blogs.gartner.com /doug-laney/deja-vvvue- others-claiming-gartners-volume-velocity-variety-construct-for-big-data/
23. Gewirtz D.; 2016
    Available:http://www.zdnet.com /article/"volume-velocity-and-variety-understanding-the-three-vs-of-big-data"/
24. Chuck Cartledge. How many Vs are there in Big Data? Working Paper; 2016.
25. Borne D; 2014.
    Available:https://www.mapr.com /blog/top-10-big-data-challenges-serious-look-10-big-data-vs
26. Uthayasankar Sivarajah, Muhammad Mustafa Kamal, Zahir Irani, Vishanth Weerakkody. Critical analysis of big data

challenges and analytical methods. Journal of Business Research. 2017;70;263-286.
27. Abedallah Zaid Abualkishik. Hadoop and big data challenges. Journal of Theoretical and Applied Information Technology. 2019;97(12).
28. Mudassir Khan. Big data analytics evaluation. International Journal of Engineering Research in Computer Science and Engineering, (IJERCSE). 2018;5(2).
29. Zahid Javed, Tariq Shahzad, Badarqa Shakoor, Muhammad Tehseen Qureshi, ozia Mushtaq. Review: Big data and Hadoop; 2015.
Available:https://www.researchgate.net/publication/ 303988873
30. Rupali Jagadale, Pratibha Adkar. A review paper on big data & Hadoop. International Journal on Recent and Innovation Trends in Computing and Communication. 2018; 6(5).
ISSN: 2321-8169,.
31. Alexandru Adrian Tole. Big data challenges. Database Systems Journal. 2013;4(3).
32. Ye Zhou, Amir Esmailpour. Improvements in big data Hadoop several hybrid efficiency methods. ASEE 2014 Zone I Conference, University of Bridgeport, Bridgeport, CT, USA. 2014.
33. Kyong-Ha Lee, Yoon-Joon Lee, Hyunsik Choi,Yon Dohn Chung, Bongki Moon. Parallel data processing with MapReduce: A survey. ACM SIGMOD 2011;40(4).
34. Michael Stonebraker, Uğur Çetintemel, Stan Zdonik. The 8 requirements of real-time stream processing. ACM SIGMOD 2005;34(4).
35. Hadoop big data analysis framework. tutorialspoint; 2016.
Available:www.tutorialspoint.com
36. Suguna S, Devi K. Improvement of HADOOP ecosystem and their pros and cons in big data. International Journal of Engineering and Computer Science. 2016;5:16680-16685.
ISSN: 2319-7242.
37. Ishwarappa, Anuradha. A brief introduction on big data 5Vs characteristics and Hadoop technology. Ishwarappa and J. Anuradha / Procedia Computer Science. 2015;48:319–324.
38. Manoj Kumar Singh, Parveen Kumar Hadooo: A big data framework for storage, scalability, complexity, distributed files and processing of massive data sets. International Journal of Engineering Research and General Science. 2014;2(5).
39. Amrit pal, Pinki Agrawal,Kunal Jain and Sanjay Agrawal. A performance analysis of map reduce task with large number of files data sets in big data using Hadoop" 978-1-4799-3070-8/14 $31.00, IEEE; 2014.
40. Shakil Tamboli, Smita shukla patel. A survey on Innovative approach for improvement in efficient of caching Technique for Big Data applications" -1-4799-6272-3/15/$31.00, IEEE; 2015.
41. Shankar Ganesh Manikandan, Siddarth Ravi. Big data analysis using apache Hadoop. 978-1-4799-6541-0/14/$31.00, IEEE; 2014.
42. Ankita Saldhi, Abniav Goel, Dipesh Yadav, Ankur Saldhi, Dhruv Saksena and S.Indu,"Big Data Analysis Using Hadoop Cluster" 978-1-4799-3975-6/14/$31.00, IEEE, 2014.
43. Avia Katal, Mohamed Wazid, Goudar RH. Big data: Issues, challenges, tools and Good practices. Conference: Contemporary Computing (IC3), 2013 Sixth International Conference on; 2013.
44. Neda Maleki, Amir Masoud Rahmani, "MapReduce: An Infrastructure Review and Research Insights. The Journal of Supercomputing; 2019.
DOI: 10.1007/s11227-019-02907-5
45. Available:http://spark.apache.org/
46. Available:http://datampi.org/
47. Soualhia M, Khomh F, Tahar S, Task scheduling in big data platforms: A systematic literature review. J Syst Softw. 2017;134:170–189.
48. Zhang B, Wang X, Zheng Z. The optimization for recurring queries in big data analysis system with MapReduce. Future Gener Comput Syst. 2018;87:549–556.
49. Available:http://www.ibm.com/developerworks/library/bd-yarn-intro/
50. Ishwarappa, Anuradha J. A brief introduction on 5Vs characteristics and Hadoop technology, science direct, procedia computer science. 2015;48:319-324.
51. Bogdan Oancea, Raluca Mariana Dragoescu. Integrating R and Hadoop for big data analysis. Evista Română de Statistică nr. 2014;2.
52. Toshifa, Aniruddh Sanga, Shweta Mongia. Big data Hadoop tools and technologies: A

review. International Conference on Advancements in Computing & Management, ICACM; 2019.
53. Thirumala Rao B, Reddy LSS. Survey on improved scheduling in Hadoop MapReduce in cloud environments. International Journal of Computer Applications (0975 –8887). 2011;34(9).
54. Apache Hadoop; 2019. Available:http://hadoop.apache.org
55. Dean J, Ghemawat S. Mapreduce, Simplified data processing on large clusters. OSDI '04. 2004;137–150.
56. Abdol Karim Javanmardi, Hadi Yaghoubyan S, Karamollah Bagherifard, Samad Nejatian, Hamid Parvin, "A unit‑based, cost‑efcient scheduler for heterogeneous Hadoop systems. The Journal of Supercomputing Springer Science+ Business Media, LLC, part of Springer Nature; 2020. Available:https://doi.org/10.1007/s11227-020-03256-4
57. Hadoop's Fair Scheduler;2018. Available:http://hadoop.apache.org/common/docs/r0.20.2/fair_ schedu ler.html
58. Hadoop's Capacity Scheduler; 2020. Available:http://hadoop.apache.org/core/docs/current/capacity _ sched uler.html
59. Matei Zaharia, Dhruba Borthakur, Joydeep Sen Sarma, Khaled Elmeleegy, Scott Shenker, Ion Stoica. Delay scheduling: A simple technique for achieving locality and fairness in cluster scheduling. In EuroSys '10", Proceedings of the 5th European conference on Computer systems, pages, New York, NY, USA, ACM, 2010;265–278.
60. Thomas Sandholm, Kevin Lai. Dynamic proportional share scheduling in hadoop. In JSSPP '10: 15th Workshop on Job Scheduling Strategies for Parallel Processing; 2010.
61. Kc K, Anyanwu K. Scheduling Hadoop jobs to meet deadlines. in Proc. CloudCom. 2010;388-392.
62. Mark Yong, Nitin Garegrat, Shiwali Mohan: Towards a resource aware scheduler in hadoop. in Proc. ICWS. 2009:102-109.

Peer-review history:
The peer review history for this paper can be accessed here:
http://www.sdiarticle4.com/review-history/62638