CHAPTER-1

INTRODUCTION

CHAPTER 1

INTRODUCTION

1.1 OVERVIEW

In general, huge information is immersed in the wealth of recent insights across all industries, experience, and life to supply and guide discoveries and innovations. smart devices are the most actors in creations and utilization of big data that shifts tradition practices into the trendy lifestyle and even analysis direction is reversed from ‘theory to data' to ‘data to theory' paradigm. it is estimated that total population and total mobile phones are more or less 6.8 billion and six billion severally [9]. Moreover, mobile applications have modified the method people assume, live and interact in smart spaces and time.

What is big data? An information assortment, data visualization, human capital, impact, infrastructure & solutions and data analysis. Analytics and big data required all higher dimensions in business atmosphere. Huge scale web mining by using information Intensive Scalable Computing (DISC) system to extract data and models from web information necessitates traditional algorithms by mentioning power of similarity [11]. DISC is measured as powerful, inexpensive and fault tolerant method of huge information sets. Shown in [12], conceptual big data adoption model design is for organizations by technical, business case development, data privacy and structure to connected processes. As shown in [3], big data processing exploitation storm system is deployed rather than MapReduce that is suitable for execution.

Hadoop allows the process of huge volumes of structured and unstructured information exploitation cluster of Artefact hardware during an easy, scalable, economical and reliable method. Hadoop is primarily installed on Linux clusters although it may be installed on Windows platforms exploitation emulators like Cygwin. Hadoop gives the Hadoop conveyed document framework, which may store and duplicate data over a bunch misuse the MapReduce.

1.2. OBJECTIVE

The Objectives of this study are:

- To establish huge information technologies landscape
- To experiment unstructured data sets of big data exploitation Hadoop system
- To discover impacts and means of visualization related to big data
- To suggest study within the area

1.3. METHODOLOGY OF THE STUDY

1.3.1. LITERATURE REVIEW

The extensive literature review is directed from conference procedures, journals, books, and therefore the internet to realize an understanding of Big data landscape and its worth share at different level to society. It gives the position to challenges related to current information also as its wide application in various areas.

1.3.2. DATA SOURCES

Organization with datasets or eBooks available for free in some file formats equivalent to text, pdf. the link has wide range of over 50,000 free eBooks with totally different categories that are simply

and freely accessible however other companies give links as open information sources however, they will need payment or process in their custody then they charge service charge.
Additionally, claiming supply public open information set by needful registration.

1.3.3. DEVELOPMENT AND PROCESSING TOOLS

Apache Hadoop Framework, an open source framework usually made-to-order for educational purpose and it is Hadoop Distributed File System (HDFS) that alleviates current processors limitation that process capability is at its ceiling point [13].
Apache Hadoop Framework provides a platform, to set information with variety of phases from data stage to output or input as a result stage.
Hadoop is considered as open source platform that deploy process of large volume of information and its storage at low cost with high speed. It can make large-scale distributed processing system exploitation commodity computers that can lesser value computation.

1.4 DOCUMENT ORGANIZATION

The thesis is organized as follows:
Chapter 1:- This chapter introduces the thesis and therefore the motive of the analysis work. in addition, the chapter includes objectives, incentive, and clarification.

Chapter 2:- This chapter provides a short review of various breathing and emerging technologies that are relevant to the proposed work situations this chapter is titled as a literature review.

Chapter 3:- This chapter demonstrates the important world problems and analysis of the complete analysis area. in addition, describes, however, the solution is developed for improving the previous solutions.

Chapter 4:- This chapter involves the summary of the implementation of the proposed resolution and conjointly includes the navigational behavior of the enforced system.

Chapter 5:- This chapter describes the experimental outcomes with the implemented system. here for this chapter includes the results graphs and their description for simulating the performance of the enforced system.

Chapter 6:- This chapter draws conclusions or outline of entire work performed in addition includes the suggestions for extending the provided solution.

CHAPTER-2

LITERATURE REVIEW

CHAPTER 2

LITERATURE SURVEY

2.1 LITERATURE SURVEY

In the year 2016, Sachin Bendea et.al Hadoop archiving technique which is able to reduce the knowledge of overhead storage of namenode and help improve the performance by decreasing the map operation of the mapreduce program. Zilan Chen, Dan Wang, Lihua Fu, Wenbing Joe 9 create small indexes for every small file to mix small files together and improve the small file storage potency and reduce the pressure on the nanomas made by the meme day. Also, the cache policy that improves the small text process efficiency of HDFS. Kashmiri P. Jayakar, Y. B. Gaurav offered a resolution decision of 10 Extended Hadoop Distributed file system (EHDFS). During this approach, a collection of the time files is connected, as the client that's marked, during a file to reduce the file count. An indexing mechanism was created to access separate files from connected Joint files. Additionally, the prefetching of the index additionally provides I / O performance enhancements and reduces a load on Name node. Chatuporn Vorapongkitipun, Natawut Nupairoj11 proposed a way referred to as the New Hadoop Archive (NHAR) supported the Hadoop Archive (HAR).

In the year 2015, Harjit Singh Lamba et.al a systematic review may be a methodology won't to assess and interpret and synthesize all the offered analysis work conducted by many alternative authors, concerning an analysis question, discipline or interest event. [8] Institutional literary reviews offer powerful insight into the IT business's value model and the current evidence concerning the method and a quick summary of its relationship with the large information. The purpose of this analysis is to find the gap between model and method to evaluate the worth of IT business from huge day. whereas conducting this analysis, Kinchelm and the charter's guidelines are followed. The steps taken for this study are the analysis of initial analysis, quality and knowledge synthesis and results.

In the year 2015, J. Christy Jackson. et al. Hadoop is an open software system project by Apache code foundation. the rationale for Scheming pig was to form Hadoop additional Accessible and Useable for non-developers. It is script primarily based, execution atmosphere defensive Pig Latin a language accustomed knowledge flows[11]. The scripting language is employed in loading and process computer file with a series of pig operators that translates input file to realize the desired output. Pig execution atmosphere has a local mode and Hadoop mode. local mode runs all the scripts and doesn't need Hadoop map-reduce and HDFS.

In the year 2014, Savitha k and Vijaya MS et al. As the growth of information will increase over years, storage and analysis become incredible, this successively will increase the processing time and value potency. though various techniques and algorithms are utilized in distributed computing the problem remains still idle. To overcome this issue Hadoop Map-reduce is employed, to method large number of files in a parallel manner. The usage of world wide web produces the information in huge amount result being more interested users in their day to day online activities. Users interaction in a website is analysed by web server log files that is in semistructured format and a computer-generated information. This paper includes an analysis of web server log files exploitation Hadoop Map-reduce to pre-process log files and explore the network anomalies and session identification. Results discloses the process speed and time potency when compared with conventional one [1].

In the year 2014, Xindong Wu et al. Large-scale web mining by using Data-Intensive Scalable Computing (DISC) System to abstract data and models from internet data necessitates traditional algorithms by putting the power of similarity [11]. DISC system is considered in a concert of powerful, fault tolerant and cheap method of large information sets in case of restriction of computing primitive. The study handled 3 classical issues in web mining: suggesting new articles from the stream in real-time, content distribution from web 2.0 to users through graph matching and finding similar items from a bag of web content.

In the year 2013, Narkhedesayalee et. Al, The log file is generated at a record rate. Thousands of terabytes or petit log files are created on every day by a knowledge center. it's very difficult to save lots of and analyze these large volumes of log files. the problem of log file analysis isn't solely due to its volume, however, log files are complicated because of completely different structures. Common database solutions like log files don't seem to be appropriate for analysis as a result of

they are not with efficiency ready to manage giant amounts of logs. 2009, Andrew Pavlow and Eric Paulson [13] compared SQLWMs and HDOOP Mapiridas and instructed hand up appraisers with speed to work quicker and load faster than load information database management system. Also, the standard database management system can't handle large datasets. this can be wherever huge data Technology Rescue Comes [8] Hoodupmr applies to several Log files are a large variety of data, therefore for the analysis of parallel applications of appropriate platforms and manages [3] appropriate for saving the log files, the Hadoop is beneficial for his or her analysis. Enterprises are a replacement approach of applying for the information collection and analysis. Hadoop is an open source project created by Dog Cutting [17], operated by Apache software package Foundation. This app works with thousands of nodes and petabits information. In the year 2013, C. Kvitha et .al internet expertise mining and behavior analysis are mentioned. they're used for cluster analysis by exploitation Fiji C and exploitation world web log information. Baier et al (2009), internet User information is proposed on a Framework for Mining light Scale, that uses the map/retrospection reflections. In internet access to mining, we tend to use to extract data consistent with information by web content that is frequently utilized by the user and here we tend to use the strategy of exhausting the way to eliminate the path of user content and transaction. we tend to introduce the formula of path evacuation exploitation the sequent pattern cluster technique. The ways used are information prepressing and information improvement strategies. In internet access to mining, we won't to extract data consistent with information by websites that are often utilized by the user and here we tend to use the strategy of draining the way to eliminate the path of user content and transaction.

In 2012, Mohamed et al, 'The analysis in explain effects of huge information, expressed by 3Vs, analytics on organizations' worth creation. As per the study, rate of information is changing tremendously high that forced organizations to handle economically and seems to be novel technologies. Methodology of case study is employed to authenticate organization's worth creation in exploitation huge information analytics. The finding shows that huge information analytics may produce value in 2 ways: improving transaction efficiency and supporting innovation.

In the year 2012, Muneto Yamamoto et .al, in today's internet world logs are crucial segment of various computing systems, that supports capabilities from error management to audit, as variety of log sources increase (such as in cloud environments), an ascendable system is important to with efficiency process logs. Log file analysis is changing into a necessary task for analyzing the

customer's Behaviour so to enhance sales also as for datasets like atmosphere, science, social network, medical, banking system it's necessary to research the log information to get required knowledge from it. web mining is that the method of discovering the knowledge from the online information. Log documents are getting to be produced quick at the rate of 1-10 Mb/s per machine, one learning focus can create many terabytes of log data amid multi day. These datasets are huge. to break down such substantial datasets, we need parallel handling framework and dependable data stockpiling system. A virtual database framework is a proficient determination for coordination of the data anyway it winds up wasteful for huge datasets. The Hadoop structure gives solid data stockpiling by Hadoop Distributed File System and Map Reduce programming model that could be a parallel preparing framework for huge datasets. Hadoop dispersed record framework separates input data and sends portions of the principal data to a few machines in Hadoop bunch to convey squares of information. This system serves to technique log data in the parallel misuse of the considerable number of machines in the Hadoop bunch and figures result with productivity. The prevailing methodology gave by Hadoop to "Store first question later", loads the information to the Hadoop Distributed record framework thus executes inquiries written in Pig Latin. This approach diminishes the reaction time too in view of the heap on to the end framework. This paper proposes a log investigation framework misuse Hadoop Map-Reduce that will give exact prompts limit reaction time [2]. This training session investigates process logs with Apache Hadoop framework.

In 2010, Liu et al., high speed and real-time big data processing exploitation storm system is employed rather than Map-Reduce that fits for batch processing. A storm is a system with wrong acceptance that achieves process together with alternative tools such as Cassandra, Redis and Kafka over NoSQL. Additionally, the study is proposed system design that supports to method Twitter and Bitly streams of knowledge.

In 2009, Jian wan et al. Huge information elements, opportunities and challenges are involved to review current state and evolution of big information in terms of 7 dimensions, historical background, what's big data? information analysis, information collection, impact, information visualisation, infrastructure & solutions, and human capital. It distils, and surveys works of literature to understand the consequences of huge information in the business atmosphere. The study shows huge information rewards in business environments however additionally in way of

life activities of individuals. In general, it's conceptually indicating that huge information and Analytics need all the seven dimensions in today's business atmosphere.

In the year 2009, Murat Bayir et al., With the fast development of the web, it's necessary to method large volumes of documents during a short time. analysis of web mining focuses on changeable strategies applicable to mass documents [1]. Mass distribution information collection and computing is an alternate technique during a distributed technique [2]. In distributed computing, a haul is split into several tasks, every of that is resolved by a computer. However, several issues like task programming, fault tolerance, and inter-machine communication are very sophisticated with programmers parallel and with a little expertise with the distribution system. In this paper, we tend to describe the results of document cluster supported our expertise and MapReduce. MapReduce [3] could be a framework that needs programmers to run solely giant an outsized an oversized task parallel and products large cluster of machines to see the scale and reduce the perform.

CHAPTER-3 PROBLEM DEFINITION & PROPOSED SOLUTION

CHAPTER 3

PROBLEM DEFINITION & PROPOSED SOLUTION

3.1 PROBLEM DOMAIN

Slow Processing Speed: Hadoop Map-Reduce (MR) process large data-sets. Due to the utilization of Map and Reduce operation, increase in latency can be recorded, hence need time to process. and cause for decrease in processing speed and more time in speed of data processing is distribution and process of data.

Latency: Map Reduce(MR) framework is in format and slow to support structural data volume. In MR Map, consider various data set that can convert into various form of datasets. Where, various elements are broken into value and pair that reduce the map input and output that processes further. MR require time for performance of task hence, can result in increased latency.

No Interactive Mode to Use: MR has no interactive mode and in Hadoop its developers need hand code for various operations that convert it into difficult operation.

No Caching: Efficiency lacked by Hadoop when it's about caching. In Hadoop MR, intermediate data in memory cannot be cached, that can hamper performance of Hadoop.

3.2 PROPOSED SOLUTION

Currently, companies are dazed by pool of information that flows from frequency identification, alternative devices or sensors that are either unstructured or voluminous and can be processed in exploitation traditional manner. This data or real-time information can be used for new, service enhancements, development or way to reply to changes in the atmosphere [17]. Values embedded in a stream of structured and unstructured information that answers most of the queries that might not be raised by business however because of technological limitations. solely five-hitter of information offered in the organization is used however, 95^{th} value of information can be chosen to be proposed as big data technology. Through digitalization, it speaks related to its usage However, it is used in additional manner. BI has no abilities to method data that has many granular,

real-time and repetitive, demands organization to induce thorough data from the selected moment prior to any changes.

Files are brought and scattered in Hadoop HDFS to save in various computers within Hadoop cluster(s). File is chopped into smaller blocks with the size larger than or equal 64MB and distributed over different nodes so as to assure replications and fault tolerance. as an instance, whenever one or a lot of nodes (s) fail(s), the chunk of a file or information in the unsuccessful node(s) are replicated to different nodes. therefore there will never be information loss. As shown in Fig. 2.1, Hadoop HDFS contains Name Node as a master node, secondary Name Node as checkpoint and Data Node as slave node that stores actual information [15].

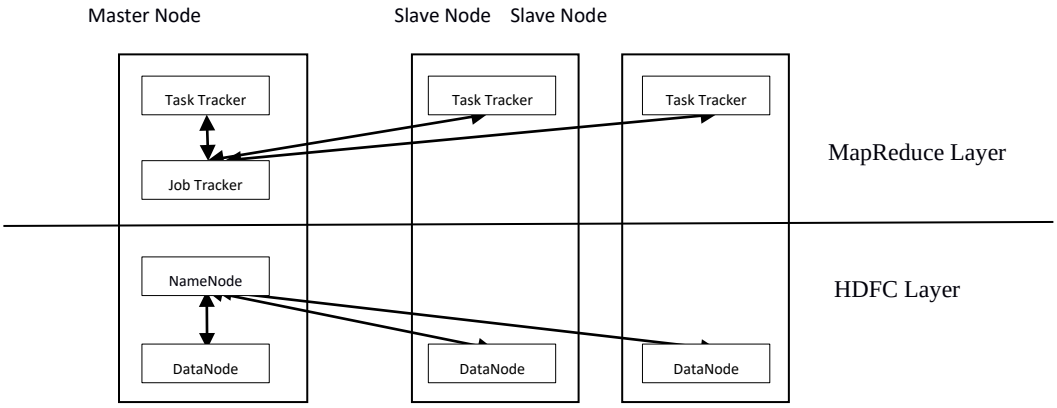

Fig. 3.1: High-Level Hadoop Architecture

Files are taken and distributed in Hadoop HDFS to save in Hadoop cluster(s) in various computers. A file will be sorted into reduced blocks with greater size or equal to 64MB. Afterwards, those files were distributed over nodes to fault tolerance and assure replications. For instance, whenever data or the chunk of file in the failed node(s) will be replaced in different nodes. Hence, it avoids data loss.

3.2.1. Big data and Its Challenges

Data streaming from various sources is collecting data in large volume with various forms at high velocity and speed. Big data is apart from "massive data" or "lots data" in a manner with incorporation of all Vs (velocity, variety and volume) to treat it as big data.

In real-time, information needs to be processed via tangible value of information in transit with extract insights. Big data as resource or tool helps to advance society; furthermore, it supports to deal with recurring world challenges such as energy, atmosphere, drought, poorness then on [23].

The upside of immense data is noteworthy inside and out ways of life regardless of the snags and furthermore the dangers, the potential worth of enormous data is unlimited NFS plans to boost the development and logical disclosure advance.

- Lead of request to new fields that cannot be potential
- Analytic tools development and information's algorithms
- Facilitate accessibility, accendibility and property of information infrastructure
- Rise in burden of social processes, interactions and human
- Promote economic process and improved health and quality of life

Discovery and innovation are ready for tools, new information, practices, and engineering, infrastructures created in science, education, national security, trade and medicine. However, exploitation and finding measures and standards for big data analysis is restricted due to infant stage; therefore, big data analysts particularly data scientists are creating methods furthermore as tools to line up information.

Testing hypothesis exploitation huge information resources may lead to false confirmation; therefore, forcing huge information to answer a specific question is an act of self-deceive which could produce the wrong conclusion. Huge information are processed without facilitating analytical software packages or statistical packages. On other side, people in general of personalities are higher in-process large data, organizing and visualizing it as acceptable. as an instance, "we tend to humans have a remembering capability within the petabyte vary which we tend to method several thousands of thoughts every day. additionally, we tend to get new data incessantly and fastly in various formats (auditory, visual, interception, gustatory and olfactory)."

3.2.2 Tools and Framework

3.2.2.1. Hadoop

Hadoop can be a framework consist of various parts for its returning and functioning that meant results. Major parts are NameNode, secondary NameNode, DataNode, JobTracker, and TaskTracker. Those elements have simple certain task generally to complete. NameNode, a brain or master of whole Hadoop system, listening all DataNodes, manages schedules of JobTracker, holds data regarding repose rack standing and then on. Secondary NameNode is taken as a backup node that takes an exposure of NameNode to restore traditional functioning once its failure. DataNode may be a slave node wherever knowledge is deposited, and data manipulation takes place before aggregation activities started. JobTracker is that the one that orchestrates all tasks to be distributed throughout across task allotted nodes. TaskTracker may be a slave by its terrible nature and its responsibility is completing the ordered task to be performed at the low level that is individual nodes or trade goods machines wherever knowledge is held on [25].

The hdfsdfs -ls /	Given HDFS list all directories and files for path of destination.
The hdfsdfs -ls -d / Hadoop	Listed directories as plain files. This command will show Hadoop folder's detailed list.
HDFS DFS –ls -H / data	Format size of file that is human-readable.
hdfsdfs -ls -r /hadoop	List all files in Hadoop directory in all subdirectories

Table3.1: List Files

Hadoop is additionally a system that consists of a collection of connected comes that are enforced to facilitate customization supported expertise and experience of organizations. the main projects are Hadoop Streaming that allows script writing for those that are acquainted on script languages, Hadoop Hive that provides SQL writing capabilities for those that are operating with SQL languages, Hadoop Pig that is only procedural language that supports information pipeline situations and Hadoop HBase that stands with real-time knowledge retrieval instead of batch processing. On high of those, Hadoop Distributed filing system and MapReduce is the most important comes that may be taken as the backbone of the system [27].

In general, Hadoop MapReduce design provides an atmosphere wherever data processing is completed in a giant set of artefact nodes. every node may be a single unit of the machine that executes allotted task fully responsibilities while not counting on different machines for its execution.

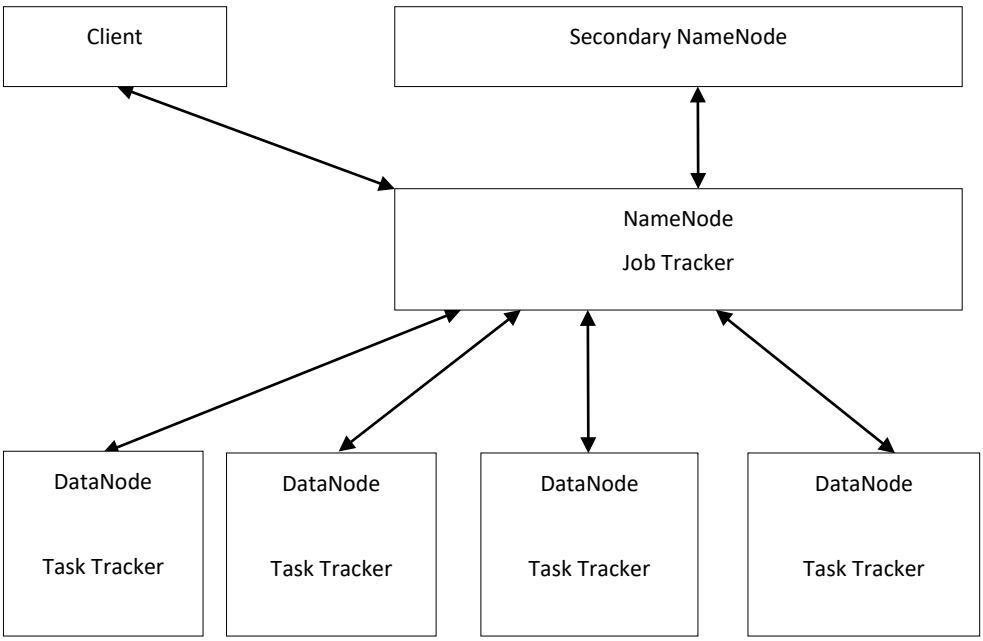

Fig. 3.2: Hadoop Components

mentioned higher than, Hadoop Map-Reduce framework is only software system resolution for current limitation of the area and process capability. rather than swing single machine with a huge area and high speed like a supercomputer that is more expensive; in addition, it demands high experience to set up and for in-progress operations likewise, there comes the lowest resolution that may be enforced with an affordable investment. come back on investment of latest huge information technologies is surprisingly high in terms of insight which will be extracted from process untapped, unstructured, knowledge set because of ancient technological limitation. From overall knowledge set, 90th information is unstructured data and that can shape traditional business practices to trendy one of completed or on progress activities. Limits of traditional RDBMS and tools of analysis are in main measurability challenge which suggests because the size of knowledge will increase their retrieving and on proportional manner manipulation won't be scaled up. In addition, schema orientated manipulation and knowledge storage has been changing into a blockage for variation in set process related to knowledge.

Traditional technologies' foundational building blocks include knowledge warehousing, transactional databases, ETL, business intelligence etc. is directly connected with structured knowledge. So, its application for semi-structured and unstructured knowledge would be terribly

effortful. although there are a variety of tries to alleviate measurability limitations of those technologies, their ceiling point to embrace changes is not elastic enough [29]. Moreover, ACID (Atomicity, Consistency, Integrity, and Durability) [30] property of relative databases isn't reposeful to elasticity for the growth of knowledge. Transactional nature of relative databases that is all dealings process shall be committed directly or fail altogether makes it a strict rule to be abided [3].

On contrary, CAP (Consistency, availableness, and Partition) theorem is an area to realize two of the three CAP tolerances considering it as a guideline and to secure all three tolerance variables in distributed computing atmosphere. Particularly, a great knowledge is considered as a platform to method giant knowledge set at high speed of various data in a distributed setting, certain there is a tradeoff among CAP tolerance variables to appropriately complete all at once. As theorem indicates, there is always a compromise between convenience and consistency in distributed computing state of affairs. Hadoop has modified the knowledge analytics landscape in method that simplifies processing despite knowledge structure in fault-tolerant manner and high performance.

3.2.2.2 Hadoop Distributed File System (HDFS)

File storage structure has been modified to take care of distributed file storage in conjunction with guaranteeing fault tolerance. In 2004 [31], Google started to modification algorithmic program to spice up its search capability by classification whole files within the web. As result, it's released a report on Google file system that was initiated new classification system, Hadoop Distributed file system, to be developed by open source community. it's a mechanism to handle giant files in a distributed manner over multiple of nodes within the type of chunks that every chunk are replicated as per set replication issue at the time of configuration. Whenever there's failure of 1 or additional nodes, knowledge is affected from unsuccessful nodes to active nodes wherever accommodation area is obtainable.

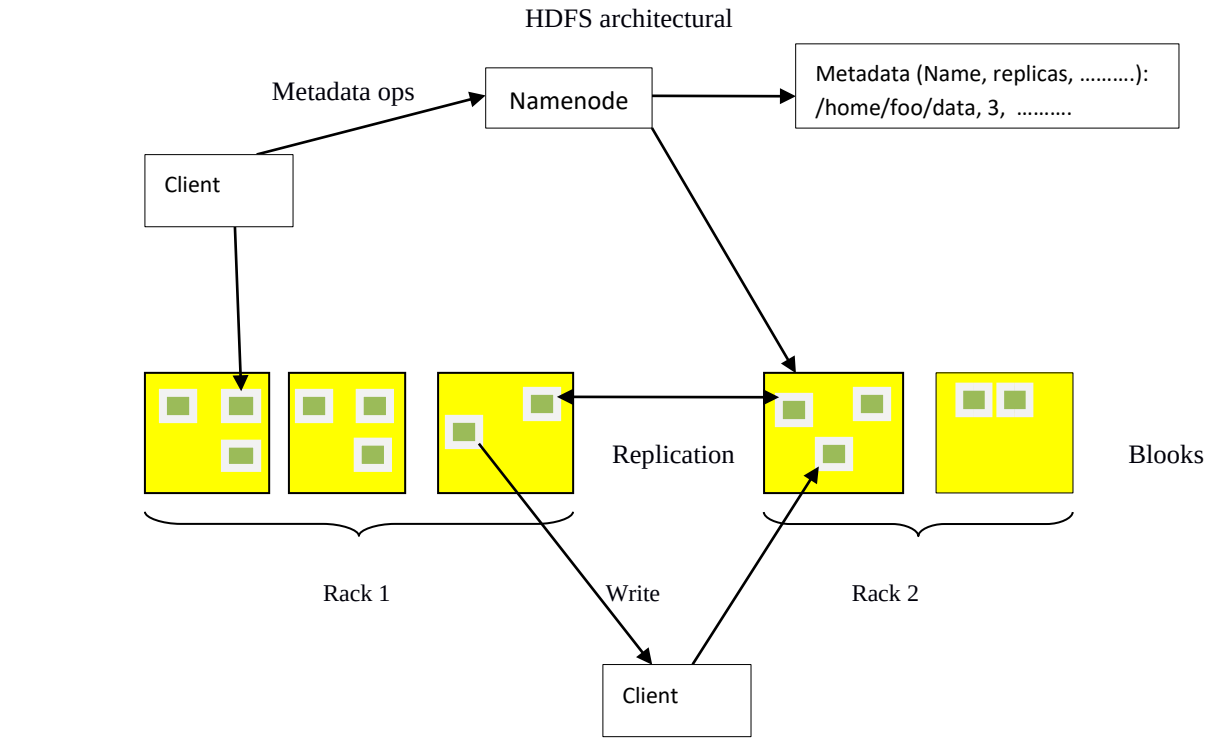

Fig. 3.3: HDFS architectural view

In addition, it creates an atmosphere wherever horizontal scaling is definitely achieved to scale out to many thousands of artefact machines [25].

In addition, as shown in Fig. 3.3, HDFS is changing into the middle of field modification for current procedure followed by improving the performance of throughput and latency. The effect of performance enhancement at a level of software system (Hadoop Framework) instead of hardware, is attracting big firms like facebook, google, yahoo etc. therefore on adopt the principle and follow moreover. It enhances read/write operations of local file chunks by moving computation to wherever knowledge is held on. It handles terribly giant files that be gigabytes or additional by reading or writing sequentially to/from nodes so there's no ought to bring knowledge to memory to control therefore the role of primary memory is turning into insignificant [12].

Hadoop's-text / hadoop1 / webdata.csv	HDFS command that takes a source file and output file into the terminal's text format.
hdfsdfs-cat / hadoop / test	This command will display the HDFS file test content in your output.
hdfsdfs -appendToFile / home / ubuntu / test / hadoop / text	append Contains text in a local file with text in an HDFS file

Table 3.2: Read/Write Files

3.2.2.3. MapReduce

MapReduce could be a programming model for process giant scale datasets during a single pass in clusters of thousands of nodes by reassuring fault tolerance and it supports 2 varieties of functions for the various purpose of duties [33]. Map Task is a function deployed to assign information to supported nodes replication issue set. Conversely, reduce task operate for information results aggregation in consistent request initiated by the client.

Proposed work: The mapper produces a middle of the road key-esteem match amid a work. The reducer entireties up all means each doc id.

1: Class Mapper

2: technique Map (doc id a, doc d)

3: For all term t ∈ doc d do

4: Emit (term t, check 1)

1: Class Reducer

2: Strategy Reduce (term t, checks [c1, c2, . . .])

3: total ← 0

4: for all Count C ∈ Count (c1, c2, . . .) do

5: aggregate ← total + c

6: Emit (term t, tally total)

Even though Map Task and reduce Task are 2 functions that are clearly visible to any or all parties, there are different functions in between Map Task and reduce Task to play a job for subsidiary activities like cacophonous, sorting, shuffling etc. Map Task depends on split function before distributing chunks of a file to nodes as per replication issue. Within the same fashion, Task is a type functions and heavy reliance to syndicate the result. Split function completes the task of chopping file into a preset size, so Map Task can send preset size chunks to delegated nodes once congregation of data done at no cost area accessibility. Mappers produce key/value combine for all coming back chunks whereas storing them. Shuffle operate, additionally, is responsible for taking input from Mappers and categorizing keys supported their groups. sort operate plays a task of sorting keys according to their values before Reducers absorb. Finally, Reducers mix similar keys and combine their values at every node that is local disk wherever the information resides.

3.2.2.4. Data Processing (Technology Stack)

In huge information technology, stack eventualities are type of data processing shifted from information retrieval from magnetic disk and interconnect primary memory processing to computation sending wherever data saved. This can be a great innovation for petascale information to avoid memory access and network traffic bottlenecks to achieve results in affordable time. Major processing paradigm shifts have been brought through the implementation of the Map-Reduce framework on high of Hadoop Distributed file system. although this has reduced burden of information transfer and manipulation to the amount of uniformity in dealing huge data, it's still

challenge in terms of generality to the specialists within the field by forcing them to understand programming language implementation and its complexness.

Java [35] is programming language that has been accustomed implement as an open source code and customizable by interested parties to adopt where Hadoop is deployed as a method of huge information. Many corporations such as knowledgeable communities are adopting Hadoop system as their favourite atmosphere by adding projects as an example, Microsoft is one in every of huge suppliers of huge information product furthermore as services however it's adopted Hadoop for large information storage and process, therefore, its projects are altogether addicted to Java libraries as a foundation. Different programming and scripting languages are now a part of Hadoop system as plugin onto MapReduce framework that perform on Hadoop is created one of the easy performing projects of those languages. Various language include Ruby, Python, Script-like languages SQL-like languages etc. run on highest of MapReduce framework.

Apache Hive is a project that can be considered as data warehouse for Hive Query Language (HQL) that provides an ability SQL-like language with exploitation for users. Generally, facts can be extracted by execution of map-reduce users. In turn, that help to inject their task in MapReduce while not delving however it functions. The tasks are either sending information for storage or retrieval specific result when process information from a collection of nodes, trade goods hardware. Hive queries are regenerate into Hadoop Jobs to run whether or not Map Task or reduce Task that doesn't mean that rational information structure is imposed on

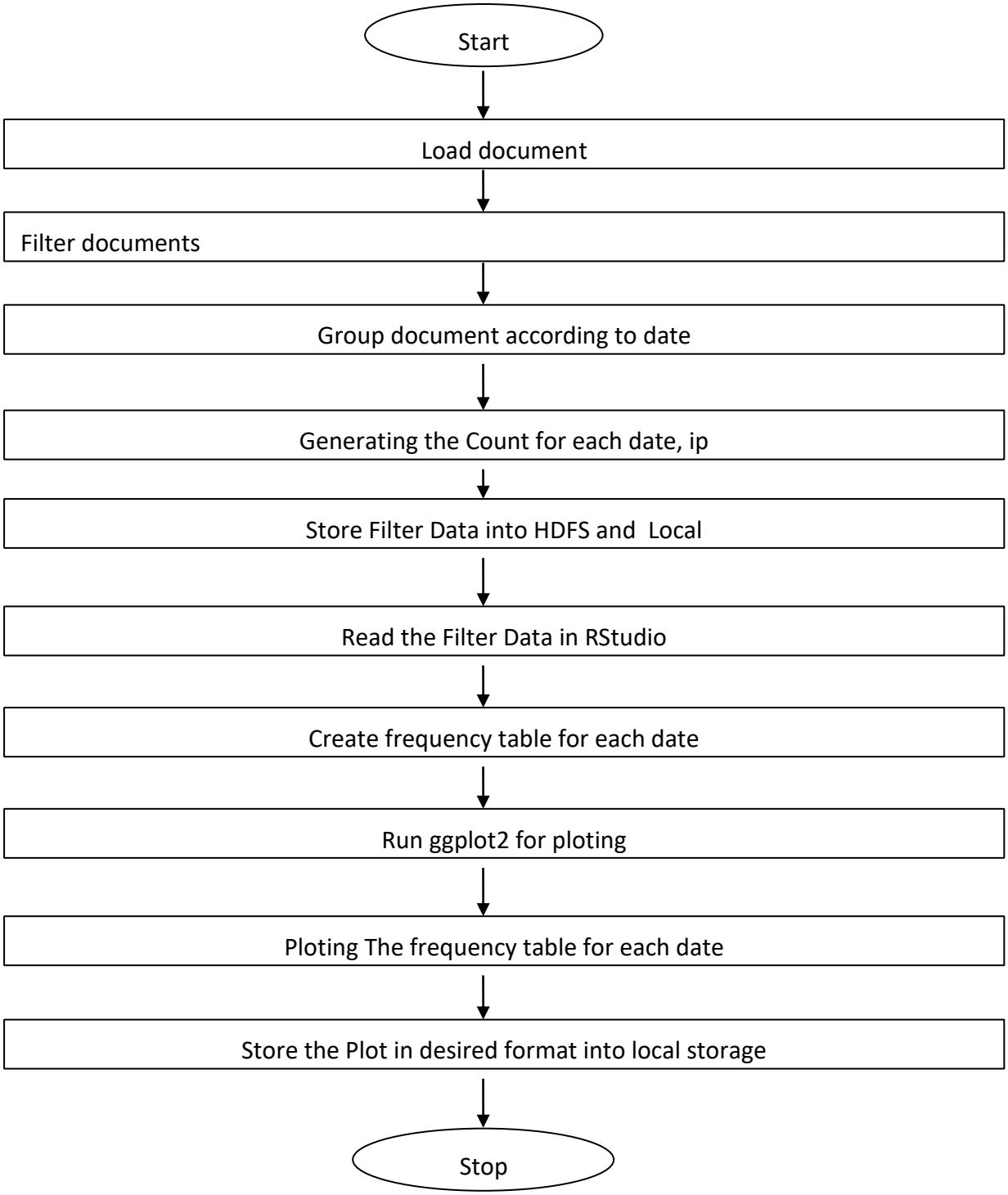

Fig3.4: Flowchart of Proposed Methodology

Map-Reduce framework rather than HQL queries can be interpreted as a task so that there won't be any force for users to map down or lessen programs related to task that realize information with

analysis of objectives. Although, HQL is a type of SQL-like language, that features different to SQL queries, to illustrate struts, maps (key/value pairs) and array.

In reasonable time it will avoid memory access and network traffic .Major data shifts have been pulled through the implementation of MapReduce structure of HDFS. Although this has reduced burden of transfering and manipulating of data to the extent of steadiness in dealing huge amount of data, it still pose challenge in terms of generality to the specialists within the field by forcing them to understand programming language implementation and its complexities. Java is the programming language that has been accustomed to implement as an open source code and it's customizable by interested parties so as to adopt where Hadoop is employed as a way to process huge information. Lots of corporations As well as knowledgeable communities are adopting Hadoop ecosystem so adapting it to their own unique atmosphere by adding projects as represented in Fig. 2.6. as an instance, Microsoft is one among huge suppliers of huge information product As well as services however it adopted Hadoop for large information storage and process, therefore, its projects are related to Java libraries as foundation. Different programming and scripting languages are becoming a part of Hadoop system as plugin onto MapReduce framework on Hadoop. Various languages such as Ruby, SQL-like languages, Script-like languages, Ruby, Python etc. all of them run on the highest of MapReduce framework.

Apache Hive may be a tool that acts like information warehouse for Hive query language (HQL) that provides for users a capability to method information For SQL-like language. In general, it abstracts details of MapReduce implementation such users will inject their task into MapReduce while not delving On how it functions. The tasks are either sending information for storage or retrieval specific result when process information from a collection of nodes, commodity hardware. Actually, Hive queries are regenerated into Hadoop Jobs to run whether or not Map Task or reduce Task that doesn't mean that rational database structure is imposed on MapReduce framework rather HQL queries are taken as a task so users won't be forced to write Map or reduce Task programs to realize information analysis objectives. although HQL is SQL-like language, it's extra options that are utterly dissimilar to SQL queries, as an example structs, maps (key/value pairs) and array.

Apache Pig [33] may be a scripting language that eases to write down jobs and send as MapReduce jobs therefore on be executed against Hadoop. it's a platform that is overtly extensile for information loading, manipulating and remodelling by victimization scripting language is termed Pig Latin. It supports advanced and complicated information manipulation although it's easy scripting language.

hdfsdfs -Chmod -r 755 /dir	Changes permissions of the files recursively.
hdfsdfs -chownhduser:hadoop /dir	Changes owner of the file. hduserin the command is owner and one is group hadoop.
hdfsdfs -chown -R hduser:hadoop /dir	Recursively changes owner of the files.
hdfsdfs -chmod 755 /dir/file1.txt	Changes permissions of the file.

Table3.4: Ownership and Validation

SQOOP [7] is one in all highest work that's accustomed link relational database and Hadoop comes together. Therefore, it facilitates information movement from relational databases, structured information, to Hadoop, schema-less or unstructured information, and the other way around. it's plug and play extensile framework that helps developers to program through the SQOOP application programming interface (API) therefore on adding new connectors.

Apache HCatalog [30] contains a role to abstract information read from HDFS files hold on in Hadoop into tabular type. It provides integrated abstraction kind for all alternative comes that relay on the tabular structure of knowledge view. for example, Pig and Hive use this abstraction to minimal the amount of difficulties in reading information from HDFS. Apart from the real fact that HDFS can be any formatting and hold on anyplace within the cluster, HCatalog can suggest a method for mapping related to formats of file and locations in the tabular read of the information. additionally, it's open and extensible for proprietary file formats.

HBase [36] could be a project that supports the practicality of NoSQL (Not solely SQL) database on the highest of HDFS. it's a storage giant|of huge|of enormous} column that would be a limitless variety of columns together with billions of rows that facilitate quick access to large datasets or large tables that is sparsely held on. it's a practicality of knowledge Modification Language (DML) [37] that supports inserts, updates, and deletes; but, Hadoop by its nature it's a write once and

browses several or infinite times. In spite of its rational information nature, it doesn't offer full options of relative databases similar to typed columns, security, increased information programmability and query language capabilities.

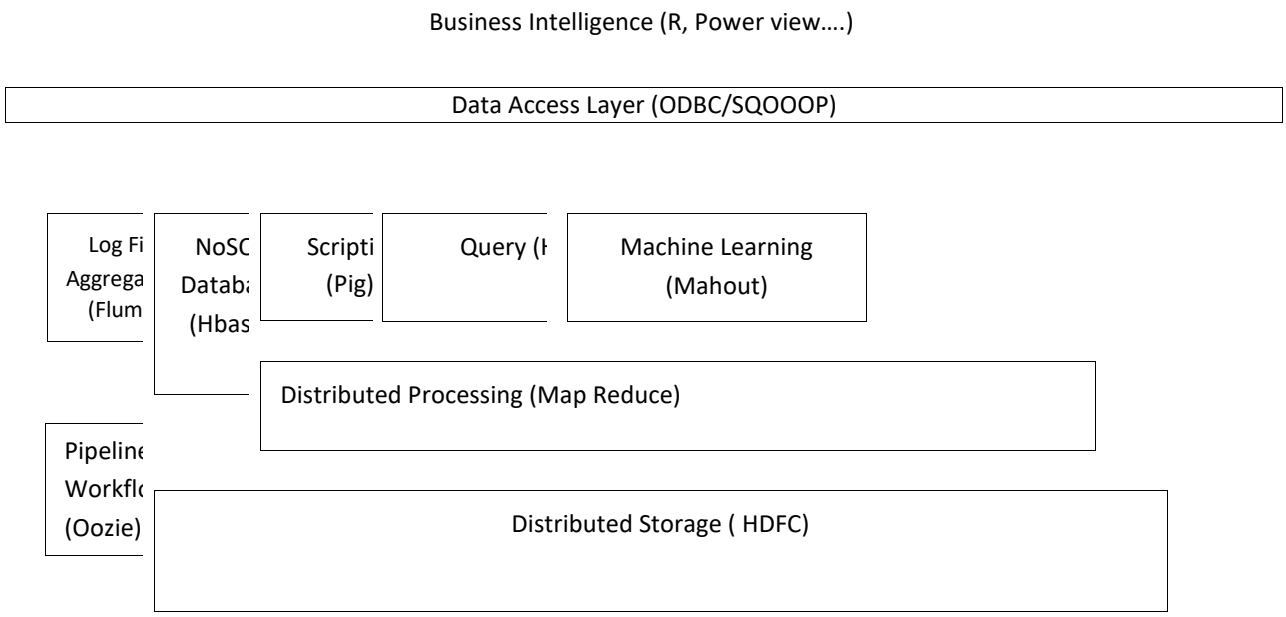

Fig. 3.5: Hadoop ecosystem [43]

Flume [38] can be a framework that knobs streaming event information besides instruction of execution system that is a Hadoop scheme nature. It ingests flowing information stream into stage that are enlisted as combination, shifts, and collect huge quantity of knowledge that commits to HDFS. Major elements of flume are enlisted as source, client, channel, destination and sink to flow events through approximately all elements to destination from client. Apache Mahout [39] could be machine learning and its goal is related to develop ascendable machine learning libraries that are imposed on highest of framework related to Hadoop exploitation MapReduce. Presently, 4 cases, with which it was supported were recommendation mining, core case for recommendation engine, the cluster is employed to cluster documents supported connected topics, classification could be an algorithmic program that consumers already classified documents therefore on classify

new documents and frequent itemset mining is a means that of understanding bucketed items along.

Oozie, Zookeeper and Ambari acts as supporting tools for Hadoop system to paired effectively and efficiently in method of information analyzing. Ambari [40], that can be a system center to Hadoop system for operational insight, management and provisioning of clusters. Oozie [41] could be a programming application for Hadoop that manages a chain of events, process or processes that should be initiated and completed at the specific measure. Zookeeper [42], on the other side is deployed to provide support for storing and managing data of configuration.

CHAPTER-4 IMPLEMENTATION AND RESULT ANALYSIS

CHAPTER 4

IMPLEMENTATION AND RESULT ANALYSIS

4.1 IMPLEMENTATION

4.1.1 PIG

Pig is a high-level scripting language used in Apache Hadoop. Data flow excels the pig describing problem. In Pig, all required information manipulations can be done with the help of Apache Hadoop. Additionally, User Defined Function (UDF) facility, the Pig code can launch code in various languages such as Jathon, Java and Jarbis. On contrast, pig scripts can be run in different languages. Finally, pig as a component can be used to build complex and larger applications that can deal in intricate issues of business.

Pig application is transaction model related to ETL that can remove data from a process source, convert it to rule set and formerly upload it in datastore. Pig can use UDF to access information from files, streams or other sources. Permit out the transition in intricate algorithm data for UDF features alteration. At last, Pig can save results in Hadoop data file system. It is translated into a work series that runs on a MR series Apache Hadoop Cluster. At apache Hadoop, Pig interpreters can optimize at good speed as a fragment of translation. Apache Pig is known to be a platform for analysis of large data set that consists of high-level language for

programs related to data analysis. Evaluate these programs to know analytics programs. The major trait of pig program is the helpful structure in paralysis, that enables to manage big data sets. Presently, Pig's infrastructure level comprises of compiler that plots the program to precise for which large-parallel applications already exist (e.g., Hadoop subproject. Pig language level currently consists of Pig Latin, a textual language with following key traits:

Extensibility: Users can create their own functions for special-purpose processing.

Optimization Opportunities: The process encoding process allows the system to optimize automatically their functionality, that allow the users to focus on words instead of skills.

Programming: Programming is easy for parallel execution, trivializing "embarrassingly parallel" to achieve tasks related to data analysis. Consist of multiple interrelated data such as compound tasks, encoded as information stream shocks, easy to maintain, write and understand.

Two main components are present in Pig system: Pig Latin, that integrates SQL's low-level processing programmable map compression and high-level announcement style and. Pig program is like specifying an execution plan for query. It shows steps' sequence, where every individual uses high-level data manipulation structure to conduct conversion related to single data using joining, group, filter etc.Pig system consider an input in Pig Latin program, then compiles it into DAG of Map Reduce work and coordinate execution on a given Hadoop cluster. The following specification shows a simple example of a Pig program. It describes a task operates over a table URL that stores and load data with two traits: ip,date. Url = LOAD 'myinput' USING PigStorage(','); usr = FOREACH Url GENERATE $1 AS ip,$2 AS date; grp = GROUP usr BY (date,ip); result = FOREACH grp GENERATE group, COUNT(usr);

Store result into 'myoutput';

4.1.2 Data Visualization (Presentation)

R, a software language for statistics analysis either an easy or complex one. It include routines for data search and summaries, data modelling and graphical representation.

While working on R, objects are created that stored in current workspace (also called as image). To save workplace, there is a need to save the session. Hence, workspace can be saved at any time by clicking at top of the control panel on disc icon. R writes the memory throughout the session and store it.

R composes that the memory all through the whole session is put away in memory. You can look back to past summons wrote by utilizing the `up' bolt key (and `down' to look back once more). You can likewise `copy' and `paste' utilizing standard windows editorial manager systems (for instance, utilizing the `copy' and `paste' discourse catches). On the off chance that anytime you need to spare the transcript of your session, tap on `File' and after that `Save History', which will empower you to spare a duplicate of the orders you have utilized for later utilize. As an option, you can physically reorder charges in a scratch pad editorial manager or something comparative.

4.2 Experiment and Results

The proposed session recognizable proof calculation utilizes the Map-diminish approach for productive handling of log documents. The log records encase HTTP ask for made on the site, for a time of multi year nine months, gathered on the discrete day and age. The procedure is done in Ubuntu 14.04 OS with Apache Hadoop-2.7.1 in pseudo circulated mode. In this mode, every one of the components of Map-lessen and HDFS keep running in its own JVM of a solitary machine. The Hadoop a java-based system is equipped for handling pet bytes of information by part into autonomous squares of a similar size. The anticipated work is done in single JVM from Sun jdk1.6 along with Eclipse IDE, along these lines the occupations of MapReduce undertaking is performed utilizing java. Every soul begun in experimental mode and information is set in the HDFS lives of localhosts.

The logs pre-processing is approved out in map task and the results are proved in Hadoop Web interfaces. The name node logs and the files can also be downloaded from the localhost for further

analysis. The pre-processed data is again scrutinized to find the session length with one map task and zero reduce task. Executing the proposed job in Hadoop MapReduce takes 2.48 minutes for pre-processing and 3.02 minutes for session identification. The Hadoop approach of preparing log documents is performed on Java 1.6 in single JVM. The log document is stacked as content, which is put away as a table. The record is then pre-processed to clean the conflicting information to which session calculation is performed. Data around 40MB of log records with session recognition is made in pre-handling 0.47 minutes. Executing the work for the entire dataset takes around 6.15 minutes. In pseudo-dispersed mode all the five hubs begin in a different JVM which can be seen in the logs index. The content document put away in HDFS is recovered and broke down in R to deliver a measurable outcome. The yield record of session distinguishing proof calculation is stacked into R comfort as a table [18] utilizing R orders.

4.3 Overview 'localhost:9000' (active)

Started: Wed Mar 21 12:11:28 IST 2018

Version: 2.4.0, r1583262

Compiled: 2014-03-31T08:29Z by Jenkins from branch-2.4.0

Cluster ID: CID-de816074-0d29-451b-b551-963a1b7e92be

Block Pool ID: BP-1546548648-127.0.1.1-1435566748426

Safe mode is off.

383 files and directories, 223 blocks = 606 total filesystem object(s).

Configured Capacity: 18.58 GB

DFS Used: 53.32 MB

Non DFS Used: 6.67 GB

DFS Remaining: 11.85 GB

DFS Used%: 0.28%

DFS Remaining%: 63.8%

Block Pool Used: 53.32 MB

Block Pool Used%: 0.28%

Number of Under-Replicated Blocks 223

Number of Blocks Pending Deletion 0 **NameNode**

Journal Status

Current transaction ID: 4199

Journal Manager State

FileJournalManager(root=/usr/tmp/dfs/name)

EditLogFileOutputStream(/usr/tmp/dfs/name/current/edits_inprogress_0000000000000004199)

NameNode Storage

Storage Directory Type State

/usr/local/hadoop/tmp/dfs/name IMAGE_AND_EDITS Active

The previous analysis was done in Java and Hadoop MapReduce (HMR) comparison between time take in both approaches is as shown in the table given below.

Total 516 MB of NASA Server Logs	Seconds
Java	375
HMR	350

TABLE 4.2. PREVIOUS APPROACH

Time take through Current approach for the same analyses done in R-base is as shown in the table given below.

Total 516 MB of NASA Server Logs	Seconds
PIG	270
R-Base	127

TABLE 4.3. CURRENT APPROACH

Total 516 MB of NASA Server Logs	Seconds
Java	375
HMR	350

PIG	270
R-Base	127

TABLE 4.4. COMPARISION BETWEEN PREVIOUS AND CURRENT APPROACH

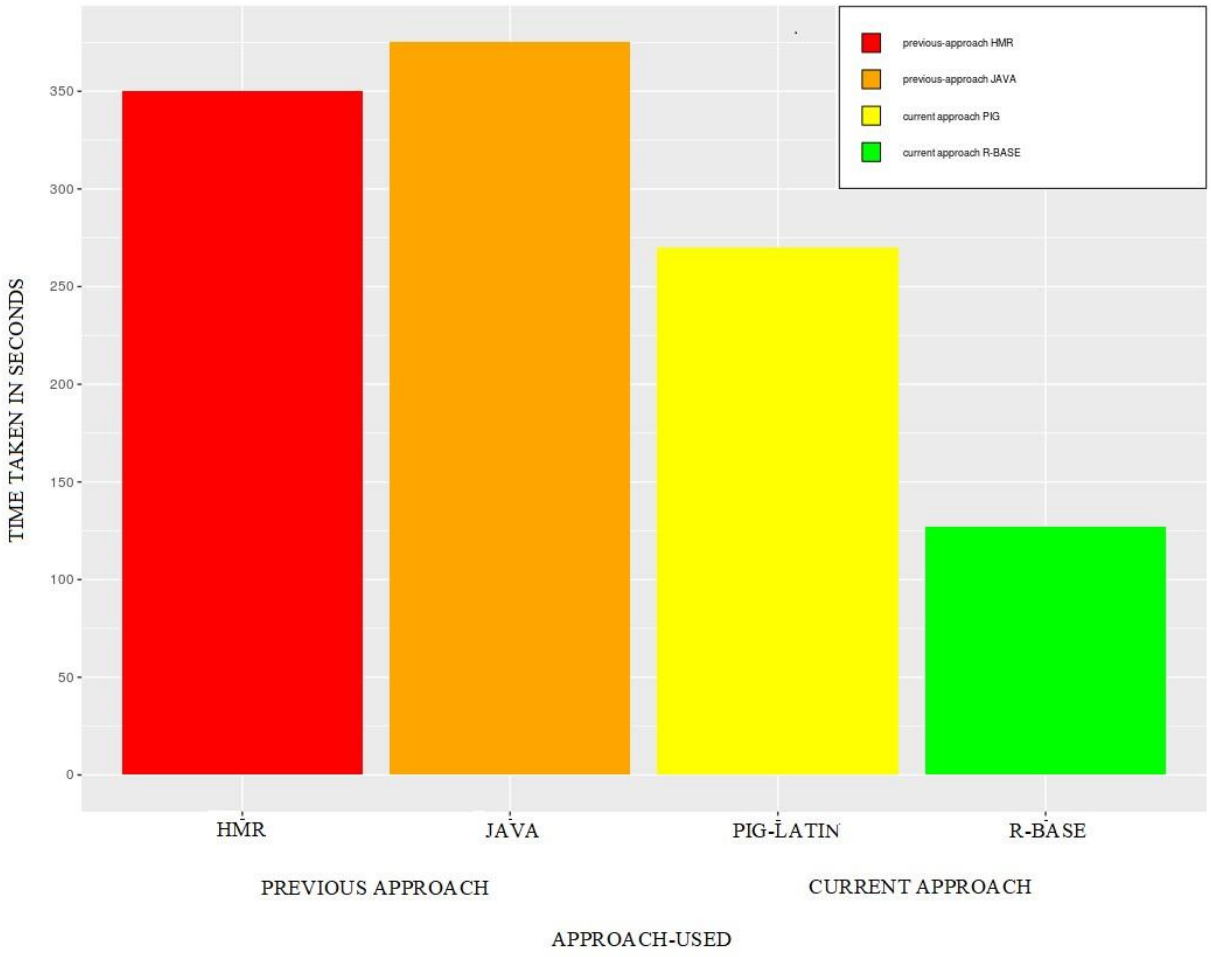

Figure 4.1 Comparison Graph between the Previous and Current approach

The one of a kind include of IP address influenced guide to diminish undertaking is examined by putting away the records as comma isolated qualities and arranged utilizing the request charge and tally > 10000 is recovered and replicated to another document for encourage investigation.

The segment fields are isolated by space, thus design coordinating is set to recover the date, time and aggregate visit made by the clients on the specific date. A bar outline is plotted for the aboveindicated fields to have a more profound examination as appeared in Figure 4.1.

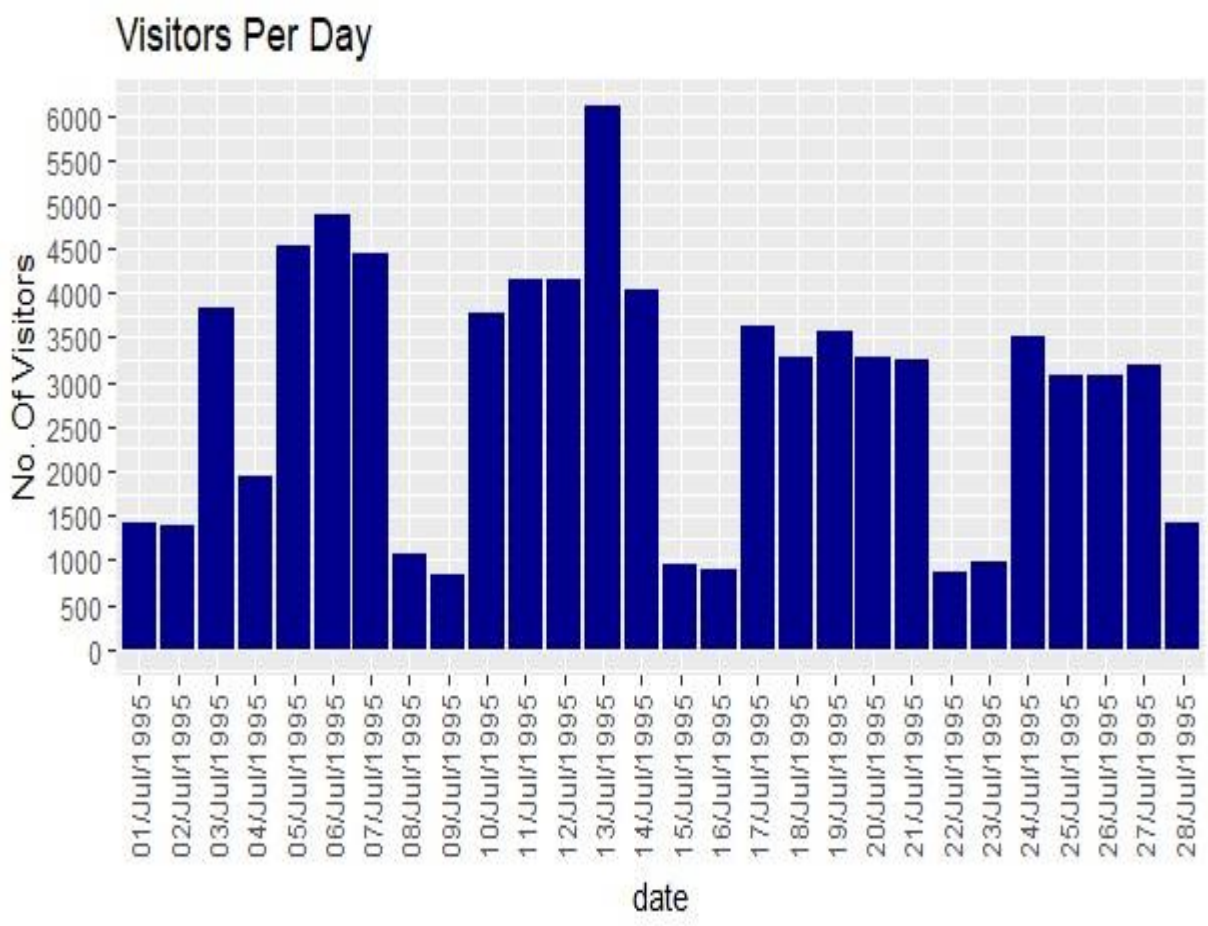

Figure 4.2. Barchart for visit by the users in a particular date

From the above outcomes, it is demonstrated that handling of content documents in single JVM on java takes additional time than preparing a similar record in Hadoop MapReduce. Despite the fact that HMR procedure a similar activity in java the information is dealt with line by line which is the predefined usefulness of guide lessen. This capacity is proficient to do the preparing in less measure of time when contrasted with java.

From Table 4.4 it is watched that an opportunity to execute 516 MB of the dataset in both the earth demonstrates a distinction of 1minute 20 seconds. While scaling the dataset to petabytes the time

discrepancy would change in hours. Expanding a similar work in bunch still lessens the effectiveness in time.

CHAPTER-5

CONCLUSION AND FUTURE WORK

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1. Conclusion

In this work, we tend to applied Hadoop Map-reduce programming model to analyze internet server log files wherever information gets a hold on multiple nodes during a cluster in order that access time needed will be reduced and Map Reduce works for big datasets giving efficient results. To get summarized results for a specific web application, we want to try log analysis that may facilitate the business strategies to enhance also to make reports in statistical manner. To access geographical reports in long analysis we can use exploitation visualization tool that shows web content hit, traffic sources, user's activity, during which a part of web site users has an interest, etc. Log investigation will be finished by differed procedures anyway what makes a difference is the time interim. Hadoop Map Reduce structure gives parallel disseminated handling and solid data stockpiling for enormous volumes of log documents. Here, Hadoop normal for moving calculation to the data rather moving information to calculation improves reaction time. From these reports, business communities will evaluate that elements of the website got to be improved in a manner such as what are the potential clients, from that topographical district site is getting most hits, and others, that can help in future promoting plans.

Big data processing and visualization is a challenge that needs the new way of tackling which is otherwise cannot be solved with the current practice of data management because data deluge and data creation frequency in varieties of formats are inevitable scenarios. The approach that is employed in this study to undertake these challenges are reviewing problem areas in detail, followed by designing solution, the implementation of the designed solution after that testing implemented a solution using big data sets. As a result shows, the Hadoop ecosystem provides a platform to process unstructured data sets of Big Data in cheap, fault tolerant and high speed. The achievement of the study expounds next generation of IT in areas of data storage, processing, and visualization. Especially, reliability and computational power do not need to scale up in terms of

hardware and processor capacities. Therefore, Big Data processing and visualization challenges are able to handle using software solutions rather than in placing specialized machines with increased hardware and processing capabilities.

5.2. Future work

This study has strong points to be raised for practical study in the area. These are data dimension and technological dimensions which indicates glimpse of light that sheds for upcoming challenges how to confront and extract insights from huge unstructured data sets. As we have seen in the study, it is possible to manage big data regardless of size and nature of data. However, full-scale experimentation on all data types including multimedia has not been carried out in this study, due to time and resources constraints, which can be researched further. Apart these, the points that require further investigation and study are fully distributed environments or clustered machines to exploit full potential by processing Terabytes and Petabytes of data sets of big data in general and its specific application for decision making by implementing revealing insights.

REFERENCES

[1] Savitha k and Vijaya MS, "An Efficient Analysis of Web Server Log Files for Session Identification using Hadoop Mapreduce", Elsevier, 2014

[2] ThanakornPamutha, SiripornChimphlee and ChomKimpan, "Data Preprocessing on Web Server Log Files for Mining Users Access Patterns" .International Journal of Research and Reviews in Wireless Communications of Vol. 2, No. 2, ISSN: 2046-6447 ,June 2012.

[3] Konstantin Shvachko, HairongKuang, Sanjay Radia, Robert Chansler, "The Hadoop Distributed File System" Yahoo, IEEE, 2010.

[4] Chris Sweeney, Liu Liu, Sean Arietta and Jason Lawrence, "HIPI: A Hadoop Image Processing Interface for Image-based MapReduce Tasks", University of Virginia,2010.

[5] Mohamed H. Almeer, "Cloud Hadoop Map Reduce For Remote Sensing Image Analysis" Journal of Emerging Trends in Computing and Information Sciences, Vol. 3, No. 4, ISSN 2079-8407, April 2012.

[6] Muneto Yamamoto and kunihiko Kaneko, "Parallel Image Database Processing With Mapreduce And Performance Evaluation In Pseudo Distributed Mode" International Journal of Electronics Commerce Studies, Vol.3,No.2,pp.211228,doi: 10.7903/ijecs.1092, 2012.

[7] Murat Ali Bayir, Ismail HakkiToroslu, "Smart Miner: A New Framework for Mining Large Scale Web Usage Data" WWW 2009, Madrid, Spain.ACM 978-160558-487-4/09/04, April 20–24, 2009.

[8] P. SrinivasaRao, K. Thammi Reddy and MHM. Krishna Prasad, "A Novel and Efficient Method for Protecting Internet Usage from Unauthorized Access Using Map Reduce". I.J. Information Technology and Computer Science, 03, 49-55, 2013.

[9] SayaleeNarkhede and TriptiBaraskar, "HMR Log Analyzer: Analyze Web Application Logs over HadoopMapReduce", International Journal of UbiComp (IJU) vol.4, No.3, July 2013.

[10] Jian Wan, Wenming Yu and XianghuaXu, "Design and Implement of Distributed

Document Clustering Based on MapReduce", Proceedings of the second Symposium International Computer Science and Computational Technology (ISCSCT '09), pp.278-280, 26-28 Dec.2009.

[11] J. Christy Jacksona, V. Vijayakumarb, Md. Abdul Quadirc, C. Bharathid "Survey on Programming Models and Environments for Cluster, Cloud, and Grid Computing that defends Big Data " 2nd International Symposium on Big Data and Cloud Computing ISBCC'15,pp.517 – 523, ELSEVIER.,2015.

[12] Ramesh Rajamanickam and C. Kavitha, "Fast Real Time Analysis of Web Server Massive Log Files Using an Improved Web Mining Architecture". Journal of Computer Science 9 (6): 771-779, ISSN: 1549-3636, 2013.

[13] Jeffy Dean, Sanjay Ghemawat. "MapReduce: Simplified Data Processing on Large Clusters", OSDI04: Sixth Symposium on Operating System Design and Implemention, Ssn Francisco, CA, December, 2004.

[14] Xindong Wu ,Xingquan Zhu, "Data Mining with Big Data" IEEE Transaction on knowledge and dataEngineering, vol .26,no.1, January 2014

[15] Harjit Singh Lamba, Sanjay Kumar Dubey , "Analysis of Requirements for Big Data Adoption to Maximize IT Business Value" ,978-1-4673-7231-2/15/2015, IEEE

[16] Sachin Bendea, RajashreeShedgeb ,"Dealing with Small Files Problem in Hadoop Distributed File System", Procedia Computer Science ,79 (2016) 1001 – 1012, ELSEVIER.