International Journal of Humanities and Social Science Invention (IJHSSI)
ISSN (Online): 2319 – 7722, ISSN (Print): 2319 – 7714
www.ijhssi.org ||Volume 11 Issue 8 August. 2022 || PP. 78-82

Role of Hadoop in Information Communication Technology (ICT)

Dr. Dayawanta Raut & Dr. Santoshi Saulkar
[1 2] SL Mankar college of Education Amgaon District Gondia. Pin code 441902

ABSTRACT
This paper enunciates the role of Hadoop technology for Information technology. Hadoop is a free, Java-based program design framework that wires the doling out of large data sets in a circulated computing environment. It is part of the Apache project sponsored by the Apache Software Foundation.If we use Hadoop technology in Information Technology then we can find speedy, accurate and precise information. The crux of this paper is to explore the knowledge of Hadoop and Map Reduce for facilitation of Information Technology. Hadoop technology has been conversed hand in hand with huge data for some time now, but IT professionals still don't know the full extent of what the technology can do or how to use it. So the objective of this paper is to highlight the role of Hadoop in Information Technology.

Keywords*: Name node, HDFS, Map Reduce, Petabyte, Data node.*

I. INTRODUCTION

What is Hadoop Technology?

Hadoop is an open, Java-based software design framework that backings the handling of large data sets in a dispersed computing environment. It is part of the Apache project sponsored by the Apache Software Foundation. Hadoop technology has been conversed hand in hand with giant data for some time now, but IT specialists quiet don't know the full scope of what the technology can do or how to use it.The open source Hadoop framework is based on Google's MapReducesoftware and can process large data sets at a gritty level. It deals analytics at a truncated price tag and high speed that some analysts say can't be achieved any other way. Essential to the usefulness of Hadoop is the Hadoop Distributed File System (HDFS), which sanctions parallel processing by straddling data over dissimilar nodes in a single cluster and provides fault tolerance. The Hadoop framework is used by chief players comprising Google, Yahoo and IBM, largely for presentations involving search engines and advertising. The preferred operating are Windows and Linux but Hadoop can also work with BSD and OS X.

Hadoop has two main layers:

A. **Computation layer:** The computation tier uses a framework called Map Reduce.

B. **Distributed storage layer:** A distributed file system called HDFS provides storage. Hadoop and Its Ecosystem Hadoop is an open source framework for processing large amount of data in batches. Hadoop is created for pipelining massive amount of data for data processing to achieve excellent end result. The idea of Hadoop is to provide a cost-efficient High Performance Computing using the cloud infrastructure. Hadoop is an Apache top level project, and licensed under Apache 2.0 To learn the basics of Hadoop, check out this Hadoop Essential Training and learn the fundamental principles behind Hadoop, and how you can use the power of Hadoop to make sense of your Big Data. As a system that allows Big Data Management to be managed in the commodity hardware that work simultaneously in parallel computing environemt, Hadoop must be able to be fault tolerant, meaning that it has to be able to continue operating properly in the event of the failure of some of its components. Built in a modular approach,

Core component of Hadoop consists of:

- Hadoop Common – libraries and utilities that provides common functionality of Hadoop
- Hadoop Distributed File System (HDFS machines in the cluster. HDFS is made to be fault tolerant and capable of running
- Hadoop Map Reduce – a component model for large scale data processing in a parallel manner.

Why is Hadoop important?

- Ability to collection and progression of huge volumes of any kind of data, swiftly.
- Calculating and computing power.
- Error tolerance

- Springiness.
- Short cost.
- Scalability.

sources :http://www.sas.com/en_us/insights/big

Hadoop Distributed File System:

HDFS, the capacity layer of Hadoop, is a disseminated, versatile, Java putting away expansive volumes of unstructured information. Map Reduce is a product system that serves as the process layer of Hadoop. Map Reduce emplo "Guide" capacity separates an inquiry into various parts and procedures information at the hub level. The "Decrease" capacity totals the aftereffects of the "Guide" capacity to decide the"answer" t Hive is a Hadoop-based information warehousing clients to compose questions in a SQL Reduce. This permits SQL software engi center and makes it less demanding to incorporate with business insight and representation devices, for example, Micro strategy, Tableau, Revolutions Analytics, etc.Pig Latin is a Hadoop Yahoo. It is moderately simple to learn and is adroit at profound, long information pipelines (a restriction of SQL.) HBase is a non-social database that takes into consideration low It adds value-based capacities to Hadoop, permitting clients to lead overhauls, embeds and erases. EBay and Face book use HBase vigorously. Flume is a structure for populating Hadoop with information. Oozie is a work process handling framework that gives clients a chance to in numerous dialects –, for example, Map Reduce, Pig and Hive other. Oozie permits clients to determine, for instance, that a specific question is just to be started after indicated past employments on which it depends for information are completed. Flume is a system for populating Hadoop with information. Ambari is a web based set of instruments for sending; overseeing and observing Apache Hadoop bunches. Its advancement is b which incorporates Ambari in its Horton works Data Platform. Avro is an information serialization framework that takes into account encoding the composition of Hadoop records. It is capable at parsing information a

Hadoop Distributed File System (HDFS) – a distributed file-system that stores data on multiple machines in the cluster. HDFS is made to be fault tolerant and capable of running a component model for large scale data processing in a parallel manner.Ability to collection and progression of huge volumes of any kind of data, swiftly. ower.

http://www.sas.com/en_us/insights/big-data/hadoop.html

HDFS, the capacity layer of Hadoop, is a disseminated, versatile, Java-based record framework capable at putting away expansive volumes of unstructured information. Map Reduce is a product system that serves as the process layer of Hadoop. Map Reduce employments are partitioned into two (clearly named) parts. The "Guide" capacity separates an inquiry into various parts and procedures information at the hub level. The "Decrease" capacity totals the aftereffects of the "Guide" capacity to decide the “answer" t based information warehousing-like system initially created by Face book. It permits clients to compose questions in a SQL-like dialect called HiveQL, which are then changed over to Map Reduce. This permits SQL software engineers with no Map Reduce experience to utilize the distribution center and makes it less demanding to incorporate with business insight and representation devices, for example, Micro strategy, Tableau, Revolutions Analytics, etc. Pig Latin is a Hadoop Yahoo. It is moderately simple to learn and is adroit at profound, long information pipelines (a restriction of social database that takes into consideration low-inactivity, brisk lookups in Hadoop. ed capacities to Hadoop, permitting clients to lead overhauls, embeds and erases. EBay and Facebook use HBase vigorously. Flume is a structure for populating Hadoop with information. Oozie is a work process handling framework that gives clients a chance to define a progression of occupations written, for example, Map Reduce, Pig and Hive - then shrewdly interface them to each other. Oozie permits clients to determine, for instance, that a specific question is just to be started after indicated past employments on which it depends for information are completed. Flume is a system for populating Hadoop with information. Ambari is a web based set of instruments for sending; overseeing and observing Apache Hadoop bunches. Its advancement is being driven by designers from Hortonworoks, which incorporate Ambari in its Hortonworks Data Platform. Avro is an information serialization framework that takes into account encoding the composition of Hadoop records. It is capable at parsing information a system that stores data on multiple on commodity hardware a component model for large scale data processing in a parallel manner.

Ability to collection and progression of huge volumes of any kind of data, swiftly. . data/hadoop.html based record framework capable at putting away expansive volumes of unstructured information. Map Reduce is a product system that serves as aments are partitioned into two (clearly named) parts.

The "Guide" capacity separates an inquiry into various parts and procedures information at the hub level. The "Decrease" capacity totals the aftereffects of the "Guide" capacity to decide the"answer" to the question. like system initially created by Face book. It permits like dialect called HiveQL, which are then changed over to Map neers with no Map Reduce experience to utilize the distribution center and makes it less demanding to incorporate with business insight and representation devices, for example, Microstrategy, Tableau, Revolutions Analytics, etc. Pig Latin is a Hadoop-based dialect created by Yahoo. It is moderately simple to learn and is adroit at profound, long information pipelines (a restriction of inactivity, brisk lookups in Hadoop. ed capacities to Hadoop, permitting clients to lead overhauls, embeds and erases. EBay and Facebook use HBase vigorously. Flume is a structure for populating Hadoop with information. Oozie is a define a progression of occupations written then shrewdly interface them to each other. Oozie permits clients to determine, for instance, that a specific question is just to be started after indicated past employments on which it depends for information are completed.Flume is a system for populating Hadoop with information. Ambari is a webbased set of instruments for sending, overseeing and eing driven by designers from Hortonworoks, which incorporate Ambari in its Hortonworks Data Platform. Avro is an information serialization framework that takes into account encoding the composition of Hadoop records. It is capable at parsing information and

Performing evacuated strategy calls.

Mahout is an information mining library. It takes the most well-known information digging calculations for performing bunching, relapse testing and measurable demonstrating and executes them utilizing the Map Reduce model. Sqoop is a network device for moving information from non-Hadoop information stores –, for example, social databases and information distribution centers – into Hadoop. HCatalog is an incorporated metadata administration and sharing administration for Apache Hadoop. BigTop is a push to make a more formal procedure or structure for bundling and interoperability testing of Hadoop's sub-extends and related segments with the objective enhancing the Hadoop stage as a whole.

Assumptions and Goals

Equipment disappointment is the standard instead of the exemption. A HDFS example might comprise of hundreds or a huge number of server machines, each putting away part of the document framework's information. The way that there are countless and that every segment has a non-minor likelihood of disappointment implies that some segment of HDFS is dependably non-utilitarian. Accordingly, discovery of flaws and speedy, programmed recuperation from them is a center compositional objective of HDFS. Applications that keep running on HDFS need spilling access to their information sets. They are not broadly useful applications that regularly keep running on universally useful record frameworks. HDFS is composed more for cluster preparing as opposed to intuitive use by clients. The accentuation is on high throughput of information get to as opposed to low dormancy of information access. POSIX forces numerous hard prerequisites that are not required for applications that are focused for HDFS. POSIX semantics in a couple key territories has been exchanged to expand information throughput rates.

What are the difficulties of utilizing Hadoop?

Map Reduce writing computer programs is not a decent match for all issues. It's useful for straightforward data solicitations and issues that can be isolated into autonomous units; however it's not productive for iterative and intuitive scientific undertakings. Map Reduce is record serious. Since the hubs don't intercommunicate with the exception of through sorts and rearranges, iterative calculations require different guide mix/sort-diminish stages to finish. This makes various documents between Map Reduce stages and is wasteful for cutting edge investigative figuring. There's a broadly recognized ability hole. It can be hard to discover section level software engineers who have adequate Java aptitudes to be profitable with Map Reduce. That is one reason circulation suppliers are dashing to put social (SQL) innovation on top of Hadoop. It is much less demanding to discover software engineers with SQL aptitudes than Map Reduce abilities. Furthermore, Hadoop organization appears to be part workmanship and part science, requiring lowlevel information of working frameworks, equipment and Hadoop portion settings.

Information security. Another test revolves around the divided information security issues, however new devices and advancements are surfacing. The Kerberos authentication convention is an incredible stride toward making Hadoop situations secure. Undeniable information administration and administration. Hadoop does not have simple to-utilize, full-include devices for information administration, information purifying, administration and metadata. Particularly missing are devices for information quality and institutionalization.

Hadoop Applications

Making Hadoop Applications More Widely Accessible Apache Hadoop, the open source Map Reduce system, has drastically brought down the cost obstructions to preparing and investigating huge information. Specialized hindrances remain, be that as it may, subsequent to Hadoop applications and advances are very unpredictable and still outside to most designers and information examiners. Talend, the open source combination organization, makes the huge making so as to figure force of Hadoop genuinely available it simple to work with Hadoop applications and to consolidate Hadoop into big business information streams.

A Graphical Abstraction Layer on Top of Hadoop Applications

With regards to our history as a pioneer and pioneer in open source information coordination, Talend is the primary supplier to offer an immaculate open source answer for empowers enormous information mix. Talend Open Studio for Big Data, by layering a simple to utilize graphical advancement environment on top of capable Hadoop applications, makes huge information administration available to more organizations and a bigger number of designers than any time in recent memory.

With its Eclipse-based graphical workspace, Talend Open Studio for Big Data empowers the designer and information researcher to influence Hadoop stacking and preparing innovations like HDFS, HBase, Hive, and Pig without writing Hadoop application code. By essentially selecting graphical parts from a palette, orchestrating and arranging them, you can make Hadoop occupations that, for instance:

- Load information into HDFS (Hadoop Distributed File System)
- Use Hadoop Pig to change information in HDFS
- Load information into a Hadoop Hive based information distribution center
- Perform ELT (separate, load, change) accumulations in Hive
- Leverage Sqoop to coordinate social databases and Hadoop

Hadoop Applications, Seamlessly Integrated

For Hadoop applications to be really available to your association, they should be easily incorporated into your general information streams. Talend Open Studio for Big Data is the perfect instrument for incorporating Hadoop applications into your more extensive information design. Talend gives more inherent connector segments than whatever other information coordination arrangement accessible, with more than 800 connectors that make it simple to peruse from or keep in touch with any significant record organization, database, or bundled undertaking application.

For instance, in Talend Open Studio for Big Data, you can utilize drag 'n drop configurable parts to make information incorporation streams that move information from delimited log records into Hadoop Hive, perform operations in Hive, and concentrate information from Hive into a MySQL database (or Oracle, Sybase, SQL Server.

II. Literature Review

S. Vikram Phaneendra & E. Madhusudhan Reddy et.al. Illustrated that in olden days the data was less and easily handled by RDBMS but recently it is difficult to handle huge data through RDBMS tools, which is preferred as “big data”. In this they told that big data differs from other data in 5 dimensions such as volume, velocity, variety, value and complexity. They illustrated the hadoop architecture consisting of name node, data node, edge node, HDFS to handle big data systems. Hadoop architecture handle large data sets, scalable algorithm does log management application of big data can be found out in financial, retail industry, health-care, mobility, insurance. The authors also focused on the challenges that need to be faced by enterprises when handling big data: - data privacy, search analysis, etc

Kiran kumara Reddi & Dnvsl Indira et.al. Enhanced us with the knowledge that Big Data is combination of structured , semi-structured ,unstructured homogenous and heterogeneous data .The author suggested to use nice model to handle transfer of huge amount of data over the network .Under this model, these transfers are relegated to low demand periods where there is ample ,idle bandwidth available . This bandwidth can then be repurposed for big data transmission without impacting other users in system. The Nice model uses a store –and forward approach by utilizing staging servers. The model is able to accommodate differences in time zones and variations in bandwidth. They suggested that new algorithms are required to transfer big data and to solve issues like security, compression, routing algorithms.

Objectives

* To enlighten the role of Hadoop Technology on Information technology
* To explore the Map Reduce technology.
* To find the different roles of Hadoop in different field.
* To highlight the advantages and disadvantages of Hadoop.

Research Methodology

I have used the secondary data for this research paper. I have used different journals, periodicals, different sites that are mentioned on the headings of references.

Recommendations

My recommendation for this paper is to disseminate the Hadoop technology for different field to make processing work easy.

Limitations

*The vital limitation of this technology is lack of awareness.
*Second important limitation of Hadoop Technology is its usability.
*Third impediment of Hadoop technology is tough and complicated computer programming.

III. Conclusions

We have entered a time of Big Data. The paper portrays the idea of Big Data alongside 3 Vs, Volume, Velocity and assortment of Big Data. The paper additionally concentrates on Big Data preparing issues. These specialized difficulties must be tended to for proficient and quick preparing of Big Data. The difficulties incorporate the undeniable issues of scale, as well as heterogeneity, absence of structure, mistake taking care of, security, convenience, provenance, and representation, at all phases of the investigation pipeline from information procurement to result understanding. These specialized difficulties are normal over a vast assortment of utilization areas, and along these lines not costeffective to address in the connection of one space alone. The paper depicts Hadoop which is an open source programming utilized for preparing of Big Data.

References

[1]. http://www.ijsrp.org/research-paper-1014/ijsrp-p34125.pdf.
[2]. http://searchcloudcomputing.techtarget.com/definition/MapReduce
[3]. http://searchcloudcomputing.techtarget.com/definition/Hadoop
[4]. https://www.talend.com/resource/hadoop-applications.html
[5]. https://www.quora.com/What-is-the-relationship-between-MapReduce-and-Hadoop
[6]. http://www.sas.com/enus/insights/big-data/hadoop.html
[7]. http://www.sas.com/enus/insights/big
[8]. data/hadoop.htmlhttp://searchstorage.techtarget.com/essentialguide/Complete-guide-to-Hadoop-technology-andstorage
[9]. http://www.tutorialspoint.com/hadoop/hadoop_mapreduce.htm
[10]. http://hortonworks.com/hadoop
[11]. https://blog.udemy.com/hadoop-ecosystem

Shri LAXMANRAO MANKAR COLLEGE OF Education Amgaon Acknowledge to Shri Keshavrao Mankar Secretary BSS and Dr DK Sanghi Principal Astt. Professor