ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE
International Journal of Advanced Research in Computer and Communication Engineering
ICITCSA 2017
Pioneer College of Arts and Science, Coimbatore
Vol. 6, Special Issue 1, January 2017

A Survey of Characteristics and Emerging Technologies in Big Data

R. Anusuya[1], B. Prabadevi[2], R. Divya[3]

Assistant Professor, Department of Computer Applications, Pioneer College of Arts and Science, Coimbatore, India

Abstract: Big Data is a turn of phrase used to mean a massive volume of both ordered and unordered data that is so huge it is complicated to process using traditional database and software techniques. This gives grow to the time of Bigdata. The term Big Data comes with the new dispute to input, process and output the data. The paper centre of attention on restraints of traditional approach to manage the data and works that are useful in treatment big data. One of the approaches used in processing big data is HADOOP framework the paper presents the major components of the framework and working process within the framework.

Keywords: Big Data, Hadoop Framework. Components of Hadoop, Working of Hadoop.

I. INTRODUCTION

The word big data is used for data that go beyond the dealing out control of conventional database systems. The general individuality of the big data is that, the size of data is too huge, the invention of data is too high-speed and most of the times the data is not straight in the form appropriate for the database systems. As the term data diverges from the big data, so also the processing required managing the big data differs from the conventional computing techniques. The digitization process is extremely fast [1] and suitable to that the invention of data is almost in digital form and the data created is increasing in size exceeding Exabyte.

In unity to data generation the computer systems are much sooner than the old systems, yet examine [2] of large scale data is a decisive factor. "The truth is that the tools are still rising, and the promise of the [Hadoop] platform is not at the stage it needs to be for business to rely on it," says Loconzolo. But the authority of big data and analytics are rising so promptly that businesses need to wade in or threat being left behind.

II. HANDOOP FRAME WORK

Hadoop is an Apache open source framework written in java that assent to distributed dispensation of huge datasets diagonally clusters of computers using easy programming models. In conventional approach, a layout will have a computer to mount up and course big data as shown in figure 1. Here data will be mount up in an RDBMS like Oracle Database, MS SQL Server or DB2 and obscure software can be written to pass on with the database, process the essential data and at hand it to the users for analysis point. Hadoop, a framework and set of tools for commerce out very huge data sets, was previously designed to work on clusters of physical machines. That has imprecise. "Now an going up number of technologies are obtainable for giving out data in the cloud," says Brian Hopkins, an analyst at Forrester Research.

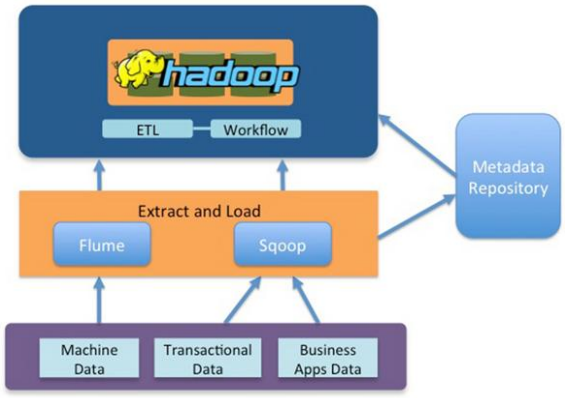

Fig. 1 Store and Process Big Data

Hadoop framework is able to enlarge applications proficient of running on clusters of computers and they could create complete statistical analysis for vast quantity of data.

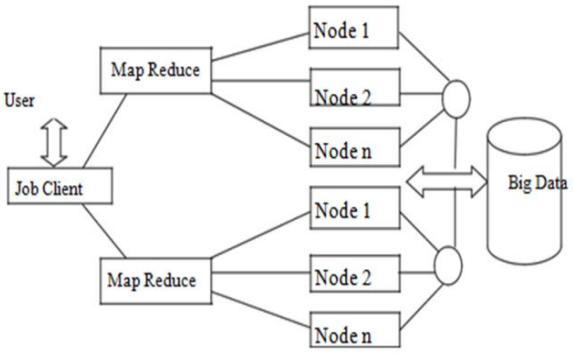

Fig. 2 Hadoop Framework

III. COMPONENTS OF HADOOP

A Hadoop frame-worked application works in an location that provides dispersed storage and working out across clusters of computers as shown in figure 2. Hadoop is deliberate to scale up from only server to thousands of

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICITCSA 2017

Pioneer College of Arts and Science, Coimbatore

Vol. 6, Special Issue 1, January 2017

machines, each input local computation and storage. Hadoop framework includes following four modules:

- Hadoop hold the Java libraries and utilities requisite by other Hadoop modules. The libraries afford filesystem and OS stage construct and enclose the essential Java files and scripts vital to start Hadoop.
- Hadoop YARN is a framework for job forecast and cluster resource management.
- Hadoop Distributed File System (HDFS) is a distributed file system that endow with high- throughput contact to application data. HDFS uses a master/slave architecture where master grip the file system metadata and one or more slaves store up the real data.

IV. WORKING OF HADOOP

Hadoop working can be estranged into phases, in the first phase, a user or application can propose a job to Hadoop job client with necessity of location of input and output files in the speckled file system, jar files contain map and diminish functions and configuring job by locale a array of parameters allied to the job.

In the second phase Hadoop job client tender the job and pattern to the MapReduce master called as JobTracker, which split out the jars/executables to the slaves, it diplomacy tasks monitor them.
In last phase the slaves on a variety of nodes perform task as per MapReduce completion and output of the diminish function is stored into the output files on the file system.

4.1 Advantages of Hadoop

- It is compatible on all the platforms.
- Distribute data and Computation. The Computation local to data prevents the network overload.
- It distributes the data and tasks across the nodes automatically, which allows the users to write and execute the distributed systems rapidly.
- The library has the API to notice and grip failures at the application layer which alleviate the framework to rely on the hardware for fault lenience.
- The framework prolongs to operate smoothly with the adding up and exclusion of servers dynamically.
- Fault tolerance by perceive faults and concern quick, automatic recovery.
- Data will be written to the HDFS formerly and then read numerous times.

A. Disadvantages of Hadoop

- Security Concern
- Still solitary master which involve care and may boundary scaling.
- Susceptible by Nature
- Not robust for Small Data
- Prospective stability Issue
- Common Limitations

V. THE HADOOP ECOSYSTEM COMPRISES OF 4 CORE COMPONENTS

5.1 Hadoop Common

Apache Foundation has pre-defined set of utilities and libraries that can be worn by further modules within the Hadoop ecosystem. For example, if HBase and Hive want to right to use HDFS they need to build of Java archives (JAR files) that are lay up in Hadoop Common.

5.2 Hadoop Distributed File System

The evade big data storage layer for Apache Hadoop is HDFS. HDFS is the "Secret Sauce" of Apache Hadoop components as consumer can plunk huge datasets into HDFS and the data will convene there satisfactorily until the user wants to authority it for examination. HDFS component produce numerous replication of the data block to be distributed diagonally unlike clusters for trusty and rapid data access. HDFS comprises of 3 vital components NameNode, DataNode and Secondary NameNode. HDFS manage on a Master-Slave architecture model where the NameNode perform as the master node for observance a follow of the storage cluster and the DataNode acts as a slave node summing up to the a variety of systems within a Hadoop cluster.

5.3 MapReduce - Distributed Data Processing Framework of Apache Hadoop

MapReduce is a Java-based system formed by Google where the definite data from the HDFS accumulate gets method powerfully. MapReduce shatter down a big data processing job into lesser tasks. MapReduce is reliable for the evaluate huge datasets in comparable before sinking it to locate the results.

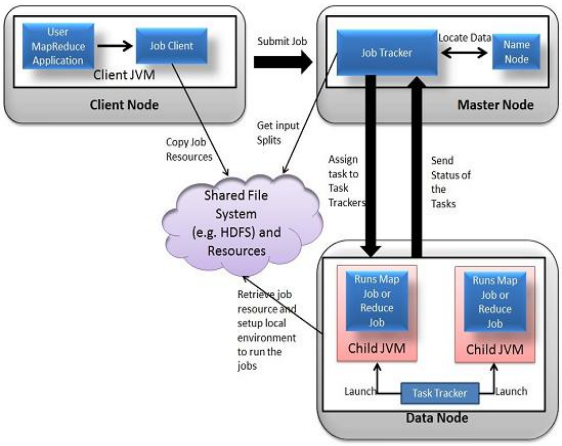

Fig. 3 MapReduce framework

In the Hadoop ecosystem, Hadoop MapReduce is a framework pedestal on YARN architecture. YARN based Hadoop architecture, ropes parallel dispensation of huge data sets and MapReduce compose obtainable the framework for simply writing applications on thousands of nodes, in view of mistake and failure management.The basic principle of process in the rear MapReduce is that the "Map" job propel a query for dispensation to a variety of nodes in a Hadoop cluster and the "Reduce" job gather all the results to output into a solitary value. In The equal

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ICITCSA 2017

Pioneer College of Arts and Science, Coimbatore

Vol. 6, Special Issue 1, January 2017

Hadoop ecosystem diminish task unite Mapped data tuples into slighter set of tuples. Meanwhile, both input and output of responsibilities are hoard in a file system. MapReduce takes care of setting up jobs, supervise jobs and re-executes the unsuccessful task.The delegation tasks of the MapReduce module are tackled by two daemons-Job Tracker and Task Tracker as shown in the image beneath. –

5.4 YARN

YARN forms an necessary part of Hadoop 2.0.YARN is vast enabler for dynamic supply operation on Hadoop framework as users can sprint various Hadoop applications lack having to bother about rising workloads

V. CONCLUSION

In this paper, we presented the gist of big data and exposed the Hadoop framework with its magnitude and concert. The advantages and disadvantages of Hadoop Framework and encompass the four components of Hadoop Ecosystem like Hadoop File System, MapReduce - Distributed Data Processing Framework of Hadoop and YARN.

REFERENCES

[1] Lyman P, Varian H. How much information 2003? Tech. Rep, 2004. [Online]. Available: http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/printable_report.pdf.
[2] Xu R, Wunsch D. Clustering. Hoboken: Wiley-IEEE Press; 2009.Google Scholar
[3] Ding C, He X. K-means clustering via principal component analysis. In: Proceedings of the Twenty-first International Conference on Machine Learning, 2004, pp 1–9.
[4] Kollios G, Gunopulos D, Koudas N, Berchtold S. Efficient biased sampling for approximate clustering and outlier detection in large data sets. IEEE Trans Knowl Data Eng. 2003;15(5):1170–87. View Article Google Scholar.
[5] Fisher D, DeLine R, Czerwinski M, Drucker S. Interactions with big data analytics. Interactions. 2012;19(3):50–9. View Article Google Scholar.
[6] Apache Hadoop, February 2, 2015. [Online]. Available: http://hadoop.apache.org.
[7] Sagiroglu S, Sinanc D, Big data: a review. In: Proceedings of the International Conference on Collaboration Technologies and Systems, 2013. pp 42–47.
[8] Chandarana P, Vijayalakshmi M. Big data analytics frameworks. In: Proceedings of the International Conference on Circuits, Systems, Communication and Information Technology Applications, 2014. pp 430–434.
[9] Big data and analytics—an IDC four pillar research area, IDC, Tech. Rep. 2013. [Online]. Available: http://www.idc.com/prodserv/FourPillars/bigData/index.jsp.
[10] Research A. Big data spending to reach $114 billion in 2018; look for machine learning to drive analytics, ABI Research, Tech. Rep. 2013. [Online]. Available: https://www.abiresearch.com/press/big-data-spending-to-reach-114-billion-in-2018-loo.
[11] Mayer-Schonberger V, Cukier K. Big data: a revolution that will transform how we live, work, and think. Boston: Houghton Mifflin Harcourt; 2013.Google Scholar.
[12] Saletore V, Krishnan K, Viswanathan V, Tolentino M. HcBench: Methodology, development, and full-system characterization of a customer usage representative big data/hadoop benchmark. In: Advancing Big Data Benchmarks, 2014. pp 73–93.

BIOGRAPHIES

Mrs. R. Anusuya completed MCA., M.Phil., in Computer Science and currently working as an Assistant Professor, Dept. of Computer Applications in Pioneer College of Arts and Science. Thirteen years of experience in teaching and published papers in International Journals and also presented papers in various National and International conferences. Area of research are Data mining and warehousing.

Mrs. B. Praba devi completed M.Sc., M.Phil., in Computer Science and currently working as an Assistant Professor, Dept. of Computer Applications in Pioneer College of Arts and Science. Ten years of experience in teaching and presented ten papers in various National and International conferences. Area of research is Data mining and warehousing.

Ms. R. Divya pursuing M.Phil in Computer Science and currently working as an Assistant Professor, Dept. of Computer Applications in Pioneer College of Arts and Science. Three years of experience in teaching and presented papers in various **National and International conferences.** Editorial Members in College Magazine and our Proceedings. Area of research is Data mining and warehousing, Computer Networks, and Cloud Computing.