JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

PERFORMANCE ANALYSIS USING HADOOP CONFIGURATION

Mr. Akhil Kadian, Dr. Ankit
Research Scholar, Assistant Professor
Department of Computer Science and Applications
Baba Mast Nath University, Asthal Bohar, Rohtak

Abstract: -

Big data can be in the form structured or non-structured information. This information can be any image or audio or video or any type of text. As bid data itself represent huge amount of data and is in the unit of terabytes or petabytes. Such huge amount of data be generated via any social network site e.g., Instagram, mobile network, Facebook etc. One of the best tools which is used in Hadoop to analyze the data is Apache Hadoop Yarn. It is an open-source operating system and is easy to install and the best part of this operating system is that after installation it hardly slows down the system. To resolve various parametric problems, an innovative framework has been introduced in this study in Apache Hadoop to solve the parametric issues during data analysis in Hadoop. To increase the performance of analysis is Hadoop it is mandatory that there must be acceptable parameter configuration. The aim of the study is to line up the performance of Hadoop via parametric configuration. Proposed methodology will result in the improvement of various issues during configuration of data analysis in Hadoop.

Keywords: - Big Data, Apache Hadoop YARN, HDFS architecture, Execution in Map Reduce, Proposed algorithm, Jobs of Hadoop YARN.

I. BIG DATA:

Big data includes heterogeneous type of huge amount of data. Such data requires different innovative techniques to analyze the data and to detect the performance after execution. As the Big data consists of complex data and of its high volume hence to process the data and to analyze this type of data an integrated framework in Hadoop is used which makes the execution much more reliable.

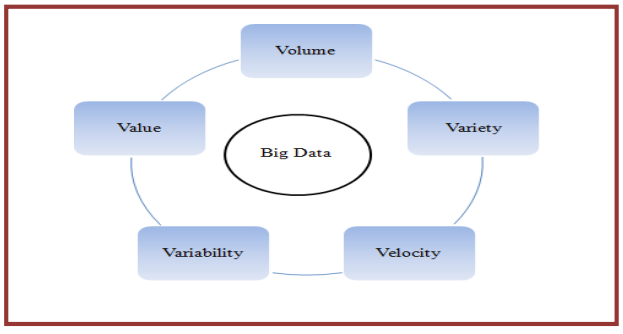

Fig.1: Components of Big Data

Fig.1 therefore shows the various components of the Big data which helps to analyze the data in a precise manner. Such components helps to analyze the growth of the data as per regular follow-up of huge amount of data.

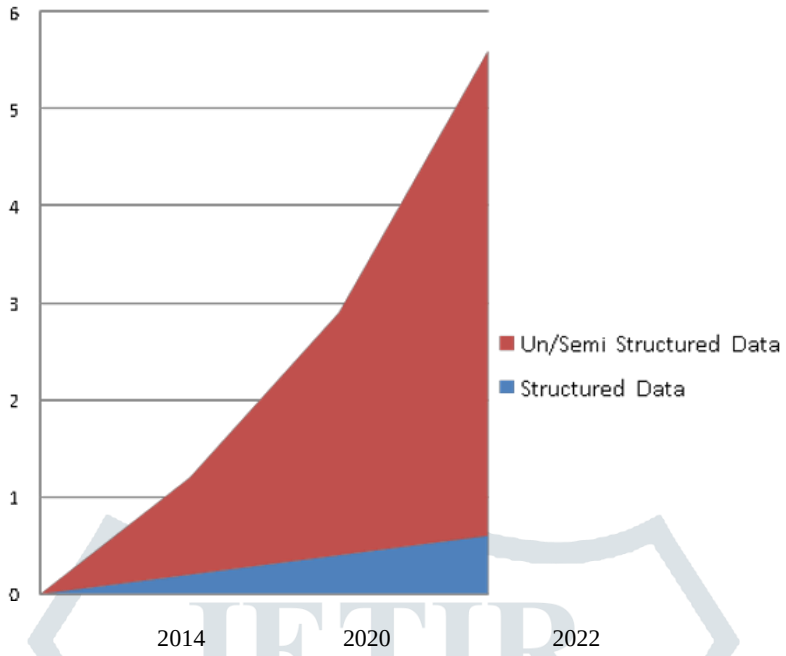

Fig.2: Growth of Big Data

Fig.2 shows the analysis of increase in big data from 2014 to 2022 as per regular analysis done in Hadoop using proposed technology.

The growth of big data is also depended on the Domain name system i.e., DNS. It verifies the request of client /user which was made prior to communication and if that request will be having any IP Address, then that IP address will be fetched via DNS Server and the connection will be established.

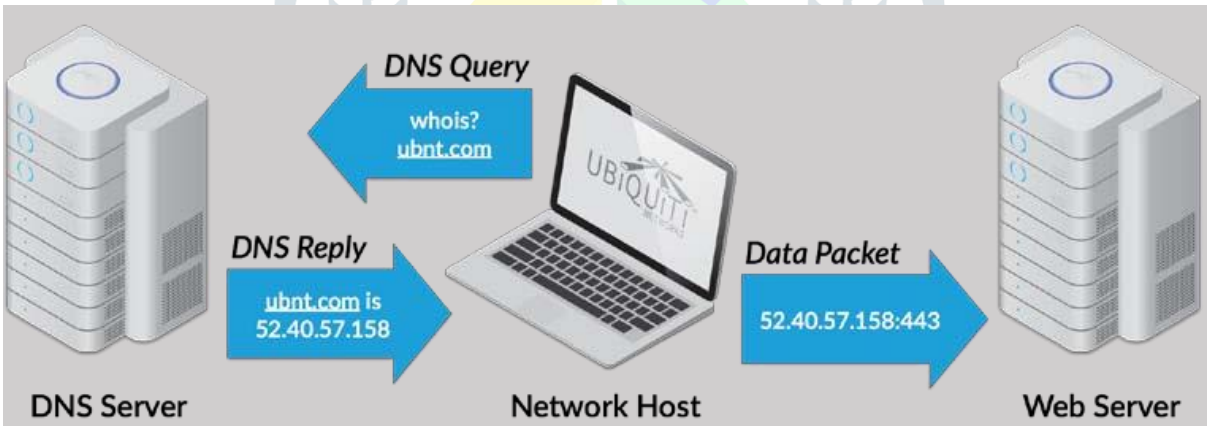

Fig.3: DNS System

As it has been clearly shown in Fig.3 which is clearly representing the communication establishment via Domain Name System only after getting IP address as mentioned in the Fig.3 .

Initially the user made a request to enter in any particular website and that request will be send towards the DNS Server which detects the IP address of that website where user want to have an access, after query received from Network host to DNS Server, DNS Server will detect the IP address for same and send back to the Network host where user will enter to that website as shown in We server side in Fig.3.

II. APACHE HADOOP YARN: -

YARN stands for Yet another Resource Negotiator. It is introduced remove the source of process finder. Hadoop Yarn is also called Map Reduce. It consists of various component like Application Master, Container, Resources Manager and no. of clients as well.

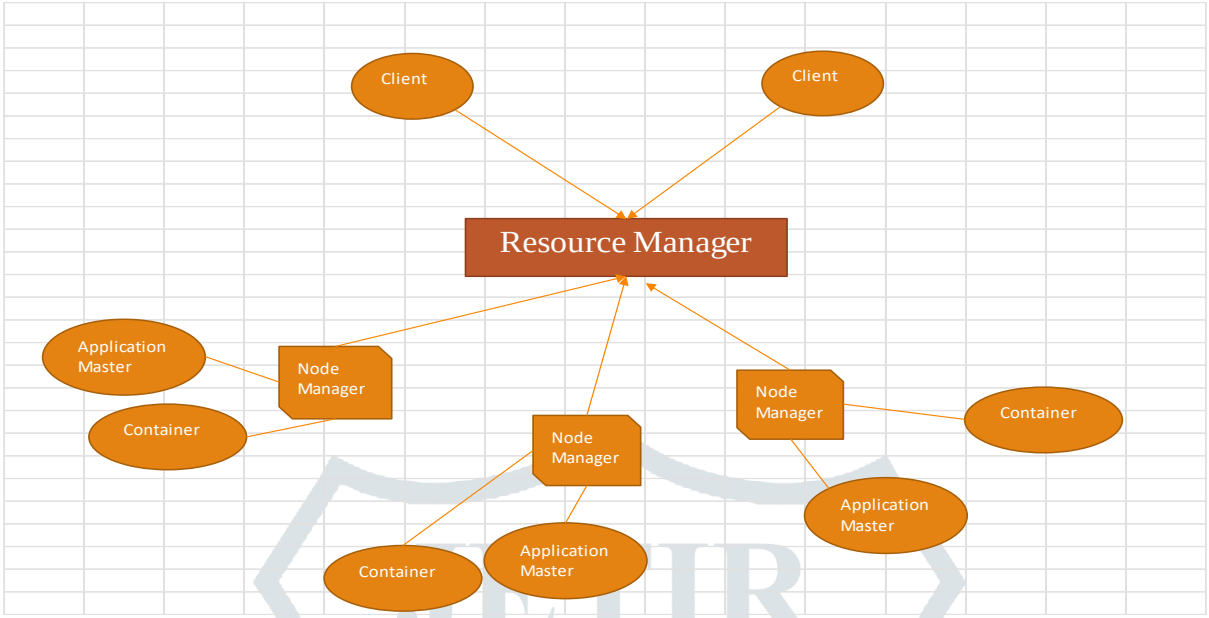

Fig.4: Architecture of YARN

Fig.4 clearly detects the communication made and the important role of resource manager which is directly accessed y the clients. The resource manager is linked with the various node manager which are having 2 components as a whole :

a) Application Manager
b) Container

As **YARN** clearly defines the concept of another resource manager, number of resource manager may be used only when there is a huge amount of data with various categories which will be further divided in to number of resource manager as per the categories.

Client and the request to resource manager and the save request be forwarded from Resource Manger to nodes which consists of components as container Application Master. There is one scheduler which is required to make all the request in a sequential manner that is one by one execution of requests.

III. HDFS Architecture in Hadoop:

It is an open-source software operating system of a google file system. HDFS is used to save huge amount of data, modify the data and execute the same data in a sequence.

In HDFS, huge data is fragmented in to number of blocks and such data in blocks will be executed locally or globally as per the nature of data. As shown in fig. 5, it clearly describes the Architecture of HDFS in Hadoop to analyze the data.

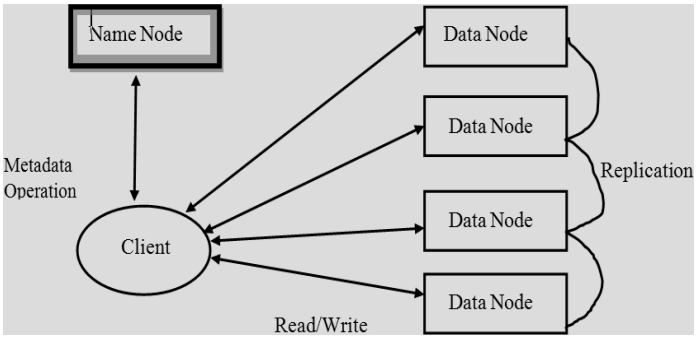

Fig.5: HDFS Architecture

Fig.5, represents the architecture of Hadoop Distributed File System (HDFS) in which data will be distributed or divided into number of blocks as per their respective categories. Here both Read and Write operations will be performed by the client such that data may be generated , modified or may be deleted by the client side. Client is working as a host in the HDFS Architecture which make the execution much faster and reliable. For the Hadoop configuration to analyze the data in a parametric form following components are required:

- CPU
- NETWORKING
- INPUT / OUTPUT
- MEMORY
- DEFAULT CONFIGURATION OF THE DATA

IV. MAP REDUCE DATA:

In Hadoop analysis of the data sorting play a vital role in the parametric analysis of huge amount of data. Mapping is also used during the execution Map Reduce 2.0 version for which all the requests are mapped in a sequential form and in the sorting all the requests or inputs are sorted as per the time and space required by individual data blocks. It basically used for large amount of data which gets fragmented in to number of blocks and the after receiving the requests from the network host to the server the IP address will be fetched and Domain Name Server will execute the access made by the user / client and communication will be established.

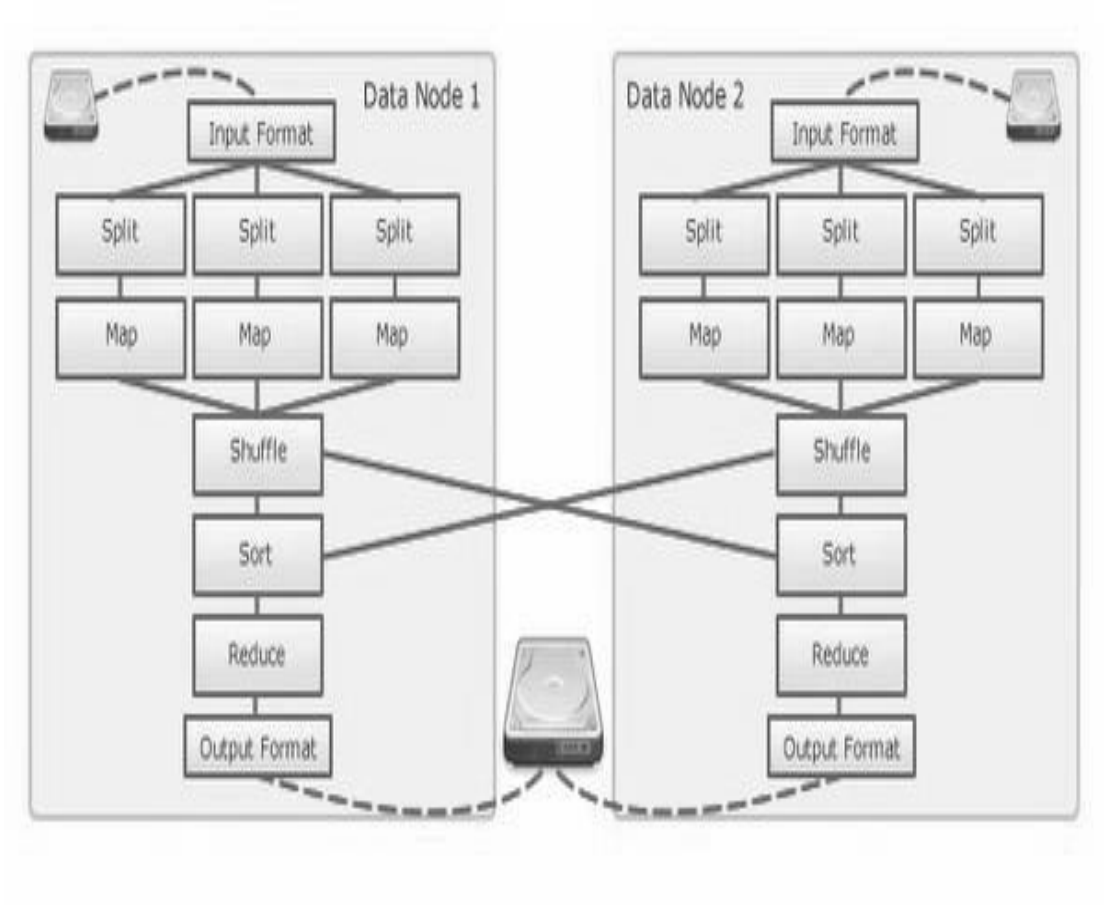

Fig.6: Internal functioning of Map Reduce Data

Fig.6 represents the internal functioning of the Map reduce 2.0 in Hadoop classification of the data. Here it has been shown that two data nodes consisting of the Inputs in terms of data blocks are used which gets further splits into number of sections as per their categories. Such data splitting gets further mapped as per the requests and the time/ space required prior to execution. Then sequential analysis and execution has been made to get an output in a proper format such that no duplicity will be made.

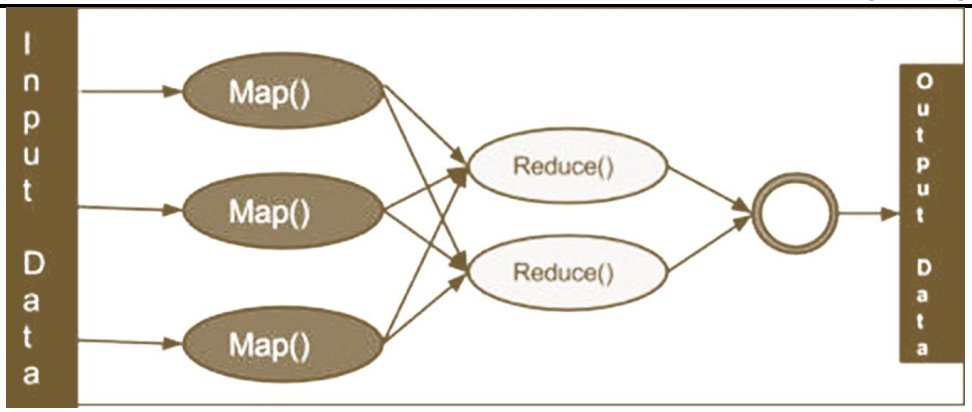

Fig.7: Map Reduce I/O

Fig. 7 Shows the basic functioning of the Map reduce data in Hadoop analysis and performance efficient process.

CONCLUSION:

The study results in huge variation in the growth of big data since 2014 to 2020 during evaluation of the performance and analyzing the Big Data. It is due to the social networking site s like Instagram, Facebook, WhatsApp and many more applications. To reduce the data from huge platform Map Reduce data is the best to analyze the performance and detecting the execution time of the requests made by user / client. Resource manager plays a vital role in making the connection with the help of application master. For this Containers are used for respective application master for saving the data and at last to analyze this huge amount of data in a specific manner.

Reference:

[1] Shvachko, Konstantin, et al., "The hadoop distributed file system", Mass storage systems and technologies (MSST), 2010 IEEE 26th symposium on IEEE, 2010.

[2] Lee, Yeonhee, and Youngseok Lee, "Toward scalable internet traffic measurement and analysis with Hadoop", ACM SIGCOMM Computer Communication Review 43.1 (2013): 5-13.

[3] Lee, Youngseok, Wonchul Kang, and Hyeongu Son, "An internet traffic analysis method with MapReduce", Network Operations and Management Symposium Workshops (NOMS Wksps), 2010 IEEE/IFIP. IEEE, 2010.

[4] Big Data Analytics with R and Hadoop, Vignesh Prajapati-Packt Publishing Ltd - 2013

[5] Navonil Sarkar, Jitin Michael, Vivek Kumar, Rahul Chaurasia, "Hit Count Analysis Using BigData Hadoop", International Research Journal of Engineering and Technology (IRJET), Volume: 03 Issue: 11 | Nov -2016, pp: 592-595, e-ISSN: 2395 -0056, p-ISSN: 2395-0072.

[6] J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Cluster", OSDI, 2004.

[7] Shvachko, Konstantin, et al., "The hadoop distributed file system", Mass storage systems and technologies (MSST), 2010 IEEE 26th symposium on IEEE, 2010.

[8] Scscc J. Shafer, S. Rixner, and Alan L. Cox, "The Hadoop Distribution Filesystem: Balancing Portability and Performance", in Proceedings of the 10th ACM SIGCOMM conference on Internet measurement ACM 2010.