IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 5, Ver. I (Sept - Oct. 2016), PP 80-83
www.iosrjournals.org

Study on Hadoop Cluster

J. Alocious Jesintha Mary,
Assistant Professor, Idhaya Arts and Science College,

Abstract: *In today's world, require Data Recovery system is most challenging aspects in the internet or World Wide Webapplications. Now a day's evens a Tera Bytes (TB) and Peta Bytes(PB) of data is not enough for storing large chunks of database(DB). Hence IT industries use concept is known as Hadoop in their applications. This approach has been adopted in Cloud computing environment for unstructured data. Hadoop is an open source distributed computing framework based on java and supports large set of distributed data processing.*

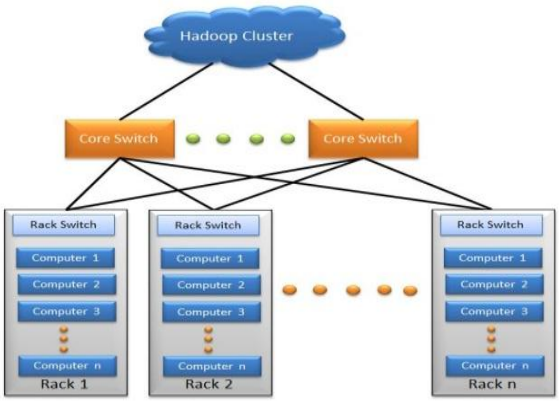
Keywords*: Big data, Hadoop, Hadoop cluster, HDFS, Name node, Datanode, Job tracker, task tracker.*

I. Introduction

Before we should see hadoop we should know the importance of hadoop and how it is related with Big data.Big data means really a big data; it is a collection of large datasets that cannot be processed using traditional computing techniques. Big data is not merely a data; rather it has become a complete subject, which involves various tools, techniques and frameworks. Big data involves the data produced by different devices and applications. Given below are some of the fields that come under the umbrella of Big Data. Black Box Data : It is a component of helicopter, airplanes, and jets, etc. It captures voices of the flight crew, recordings of microphones and earphones, and the performance information of the aircraft. Social Media Data: Social media such as Facebook and Twitter hold information and the views posted by millions of people across the globe. Stock Exchange Data : The stock exchange data holds information about the 'buy' and 'sell' decisions made on a share of different companies made by the customers. Power Grid Data: The power grid data holds information consumed by a particular node with respect to a base station. Transport Data: Transport data includes model, capacity, distance and availability of a vehicle. Search Engine Data: Search engines retrieve lots of data from different databases.

Thus Big Data includes huge volume, high velocity, and extensible variety of data. Big Data Technologies-Big data technologies are important in providing more accurate analysis, which may lead to more concrete decision-making resulting in greater operational efficiencies, cost reductions, and reduced risks for the business. There are various technologies in the market from different vendors including Amazon, IBM, Microsoft, etc., to handle big data. While looking into the technologies that handle big data, we examine the following two classes of technology: Operational Big Data this include systems like Mongo DB that provide operational capabilities for real-time, interactive workloads where data is primarily captured and stored. Analytical Big Data-This includes systems like Massively Parallel Processing (MPP) database systems and Map Reduce that provide analytical capabilities for retrospective and complex analysis that may touch most or all of the data. Hadoop is an open-source software framework for storing data and running applications on

clusters of commodity hardware. It provides massive storage for any kind of data, enormous processing power and the ability to handle virtually limitless concurrent tasks or jobs. Hadoop cluster is a special type of computational cluster designed for storing and analyzing vast amount of unstructured data in a distributed computing environment. These clusters run on low cost commodity computers.

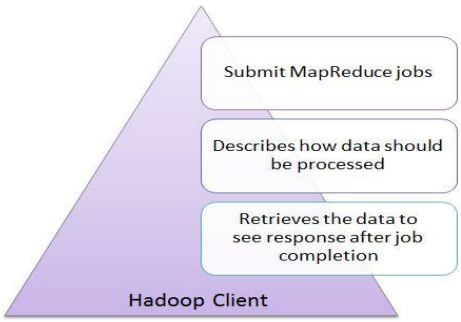


Hadoop clusters are often referred to as "shared nothing" systems because the only thing that is shared between nodes is the network that connects them. Large Hadoop Clusters are arranged in several racks. Network traffic between different nodes in the same rack is much more desirable than network traffic across the racks.

II. Core Components of Hadoop Cluster:

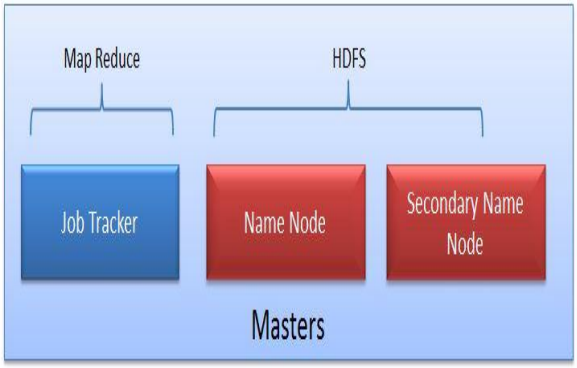
II.1. Client

It is neither master nor slave, rather play a role of loading the data into cluster, submit Map Reduce jobs describing how the data should be processed and then retrieve the data to see the response after job completion.



II.2. Masters

The Masters consists of 3 components Name Node, Secondary Node name and Job Tracker.



II.3 Name Node

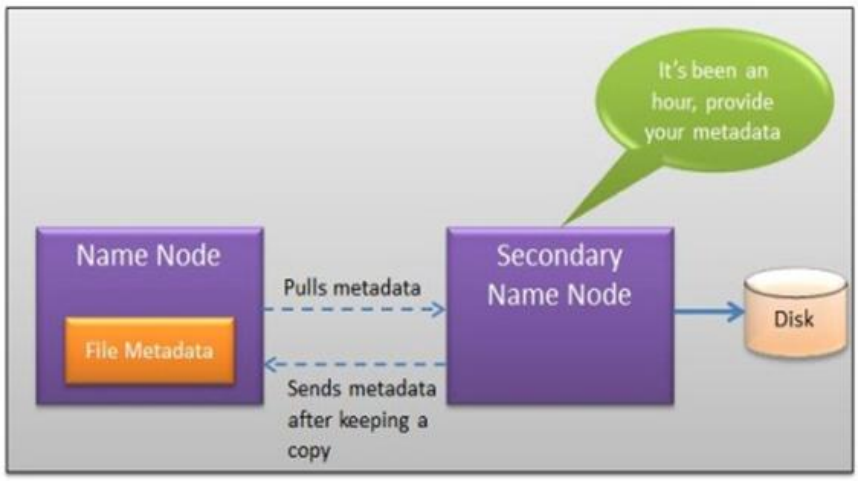
Name Node does NOT store the files but only the file's metadata. The Data Node which stores the files actually. Name Node oversees the health of Data Node and coordinates access to the data stored in Data Node. Name node keeps track of all the file system related information such as to which section of file is saved in which part of the cluster

- Last access time for the files
- User permissions like which user have access to the file

II.4 Job Tracker

Job Tracker coordinates the parallel processing of data using Map Reduce.

Secondary Name Node: Don't get confused with the name "Secondary". Secondary Node is NOT the backup or high availability node for Name node.

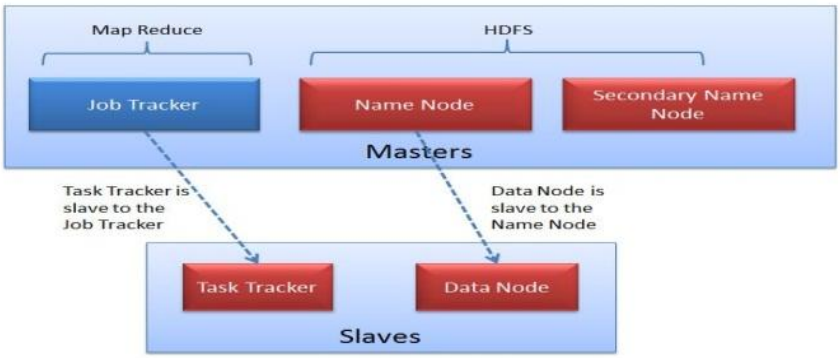


The job of Secondary Node is to contact Name Node in a periodic manner after certain time interval (by default 1 hour). Name Node which keeps all file system metadata in RAM has no capability to process that metadata on to disk. So if Name Node crashes, you lose everything in RAM itself and you don't have any backup of file system. What secondary node does is it contacts Name Node in an hour and pulls copy of metadata information out of Name Node. It shuffle and merge this information into clean file folder and sent to back again to Name Node, while keeping a copy for itself. Hence Secondary Node is not the backup rather it does job of housekeeping. In case of Name Node failure, saved metadata can rebuild it easily.

II.5 Slaves

Slave nodes are the majority of machines in Hadoop Cluster and are responsible to

- Store the data
- Process the computation



Each slave runs both a Data Node and Task Tracker daemon which communicates to their masters. The Task Tracker daemon is a slave to the Job Tracker and the Data Node daemon a slave to the Name Node.

III. A Thriving Ecosystem

Beyond these core components, and as a result of innovation such as YARN, Apache Hadoop has at hriving ecosystem of vendors providing additional capabilities and/or integration points. These partners contribute to and augment Hadoop with given functionality, and this combination of core and ecosystem Provides compelling solutions for enterprises whatever their use case. Examples of partner integrations include: Business Intelligence and Analytics: All of the major BI vendors offer Hadoop integration, and specialized analytics vendors offer niche solutions for specific data types and use cases. Data Management and Tools: There are many partners offering vertical and horizontal data management Solutions alongside Hadoop, and there are numerous tool sets – from SDKs to full ID Eexperiences – for developing Hadoop solutions. Infrastructure: While Hadoop is designed for commodity hardware, it can also run as an appliance, and be easily integrated into other storage, data and management solutions both on-premise and in the cloud. Systems Integrators: Naturally, as a component of an enterprise data architecture, then SIs of all sizes are building skills to assist with integration and solution development. As many of these vendors are already prevalent within an enterprise, providing similar capabilities for an EDW, risk of implementation is mitigated as teams are able to leverage existing tools and skills from EDW workloads.

IV. Conclusion

Hadoop architecture has fully integrated storage and compute frameworks. This key design of Hadoop—collocated storage and compute—lets enterprises meet each data processing need while achieve scale, elasticity, durability, security, and governance demanded by today's big data solutions.

References

[1]. Apache. Hadoop. http://hadoop.apache.org/.
[2]. Luiz André Barroso and UrsHölzle.The Case forEnergy-Proportional Computing.Computer, 40(12), 2007.
[3]. Standard Performance Evaluation Corporation.Specpower_ssj2008. http://www.spec.org/power_ssj2008/.
[4]. Jeffrey Dean and Sanjay Ghemawat. MapReduce: SimplifiedData Processing on Large Clusters. Commun.ACM, 2008.
[5]. Chang Fay et al. Bigtable: A Distributed Storage System forStructured Data. OSDI, USENIX, 2006.
[6]. Xiaobo Fan, W. Weber, and L. A. Barroso. Power Provisioningfor a Warehouse-sized ComputerDhrubaBorthakur.2007[E-book], "The Hadoop Distri- buted File System: Architecture and Design", Availablethrough; http:// hadoop. apache. org/common/ docs/r0. 18.0/hdfs_ design.pdf
[7]. Dean J, Ghemawat S: MapReduce: Simplified data processing on largeclusters. Sixth Symposium on Operating System Design andImplementation:2004;San Francisco, CA Usenix Association; 2004.
[8]. Dean J, Ghemawat S: MapReduce: A Flexible Data ProcessingTool.Communications of the ACM 2010, 53(1):72-77.
[9]. T.White," Hadoop: the definitive guide" 'O' Reilly, 2012.
[10]. D. Borthakur,"HDFS architecture guide. Hadoop ApacheProject"", http://hadoop.apache. org/common/docs/ current/ hdfs_ design. pdf.2008.