ADVANTAGES OF HADOOP

NATHAN C. TOKALA

Abstract— The main focus of computing, for the las few decades, was to increase the computing power of a single system and to create super computers which would have phenomenal processing speed and more RAM. The traditional computing was all process bound which involved significant amount of complex processing on relatively small amount of data. When it comes to present day systems they have to deal with large amounts of data than in the recent past. This is where hadoop offers a whole new approach and also eliminates this bottleneck and offers various advantages over the traditional computing.

Index Terms— DataNodes, Hadoop, Job Track, MapReduce, NameNode

1 Introduction

For decades the main focus of computing was to increase the computing power of a single system and to create super computers which have phenomenal processing speed and more RAM. The traditional computing was all process bound which involved significant amount of complex processing on relatively small amount of data. When it comes to present day systems they have to deal with large amounts of data than was in the past. This data is inherent value and cannot be discarded. For example leading software giants like yahoo, Facebook, eBay have about 170PB , 30PB and 5PB of data generated by their users per day respectively. Many organizations are even gathering terabytes of data per day.

So with all this data to get it to the processor would be a huge bottleneck that might take a lot of time to yield the desired results. This is where hadoop offers a whole new approach which eliminates this bottleneck and offers various advantages over the traditional computing.

Hadoop is based on work done by Google in the early 2000s. They named it google file system (GFS) published in 2003 and the MapReduce published in 2004. This gave rise to a whole new approach to the problem of distributed computing. In this approach the data is not transferred across the network but the processor itself is brought to the data which is distributed across the local data nodes while it is stored. These data nodes communicate as less as possible with each other.

2 High Level Overview:

The power of the Hadoop is in the parallel access to data that can reside on a single node or on thousands of nodes. When the data is loaded into the system it is split into chucks of data which is typically of 64Mb or 128Mb. MapReduce is the process that enables access to run on each of the nodes in the cluster. A master node called the name node allocated the work to nodes such that the MapReduce task runs on those nodes locally. All these jobs which are assigned by the name node are run in parallel, each on their own part of the complete data set that is stored in the file system.

3 Components of hadoop:

Hadoop mainly consists of two core components. The hadoop distributed file system (HDFS) and the MapReduce Software Program. So a set of machines running these two core components is known as a hadoop cluster. More the number of these machines better will be the performance.

HDFS is responsible for storing the data on the cluster. These data files are split into blocks and stored multiple times across the nodes. The default replication factor is three. This ensures reliability and availability.

MapReduce is the system used to process the data in the hadoop cluster. This process consists of two phases: a maples and a reduce phase.

Hadoop is comprised of five specific demons. The NameNode which holds the metadata of the HDFS, Secondary Name Node which performs the housekeeping functions of the nematode, the Detained which stores the actual HDFS data blocks, JobTracker which assigns the map reduce jobs to the datanodes where the processing of the data will take place, and finally the Task Tracker which is on the detained which is responsible for initiating and monitoring individual Map and Reduce tasks.

4 Advantages of Hadoop:

The main advantages of the Hadoop are automation and parallelization. It has a great fault tolerance which partially slows the process but does not have a complete system failure. Even with the failure of some of the data nodes there is no loss in data because we have a replication of the same data on other data nodes. Component recovery is another advantage where a failed node can rejoin the system without restarting the complete system itself. Component failures while running a job does not affect the outcome of the job. This also ensures greater scalability, where increasing the resources will increase the performance.

References

[1] Sammer, E. 2012. *Hadoop Operations*. Sebastopol, CA: O'Reilly Media.

[2] Capriolo, E., D. Wampler, and J. Rutherglen. 2012. *Programming Hive*. Sebastopol, CA: O'Reilly Media.

[3] "HDFS High Availability Using the Quorum Journal Manager." Apache Software Foundation. Available at http://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/HDFSHighAvailabilityWithQJM.html.

[4] Kestelyn, J. "Introducing Parquet: Efficient Columnar Storage for Apache Hadoop." Available at http://blog.cloudera.com/blog/2013/03/introducing-parquet-columnar-storage-for-apache-hadoop/.

Nathan C. Tokala, Junior at Independent School, Wichita KS, USA. His interests lie in financial data analysis and computer programming.