HADOOP: A Brief Review

Dr. Kanwal Garg[1],Er. Disha Kangar[2]
[1]Astt. Proff.,K.U.K.,[2]M.tech scholar
[1,2] Kurukshetra University,Kurukshetra
[1]disha.kangar10@gmail.com

Abstract- Information explosion has become the truth on world wide web with the highly active participation and proliferation of Web2.0 companies such as Google and Facebook.The point of concern is how to store and process such a large amount of data for information and knowledge. Hadoop is the answer for this. It has evolved as a superior to its counterpart technologies such as SQL and NFS.Hadoop enables the users to store and analyze the information in newer ways. It has been developed by Doug Cutting.

Keywords-data, Hadoop, HDFS

1. Introduction

Over a decade, many companies like Facebook, Amazon, Google, Yahoo have been using Hadoop as its base technology for storing and processing their vast data repositories. With the evolution of technology, there has been a steep decrease in hardware cost, leading to the opportunity for data scientists to store as much information they want but limitation was imposed by the data traffic that can be put on the network. As a solution Hadoop came into existenceThe entire Apache Hadoop "platform" is now commonly considered to be consisted of the Hadoop kernel, MapReduce and Hadoop Distributed File System (HDFS), as well as a number of related projects –including Apache Hive, Apache HBase, and others.In this paper, there is a brief overview of the technology and other related facets have been described. Section 2 gives a comparison of NFS with HDFS. Section 3 explores the cases in which to use HDFS and when to avoid using it. Section 4 describes the fundamentals of HDFS. Section 5 discusses the advantages of Hadoop. Section 6 explores the possible application areas.

2 . Comparison of NFS with HDFS

Distributed File System serves two purposes – holding a large volume of data and serving a large number of clients over the network. Network File System served the needs of distributed computing for a long time but there were a few limitations associated with it which laid the foundation for development of HDFS.

In NFS, client's remote file system mounts the logical view of the file provided by the server on client locally. So, a limited data can be stored because the constraint is put by the capacity of the machine. While in case of Hadoop, all the data is stored in HDFS and clients are being provided with the processed result.

In NFS, no failure protection feature has been provided. If logical volume goes down, all attached clients will be in hang state. As a remedy to this problem, in Hadoop, replication factor of value 2(by default) or even higher(depending on the criticality of application) is provided.

Overload on the network is also major problem in NFS. In Hadoop, as only queries and their results flow on the network, no overload is there.

3. Just use of HADOOP

To every data processing application, use of the technology can not prove beneficial. There are several situations where right worth of efforts can be gained like when one has to deal with very large files. There can be large log files like that of a power house or a communication provider company. Another significant area is of streaming data access i.e. whenever there is a need to read a bigger volume of data rather than reading a single volume of data from a number of sources. The application may desire the data to be refined in amounts of GBs at a single processing time. Hadoop serves as a good solution where the organization wants to save the I.T. cost.There are also situations where Hadoop must not be used. One of them is the set-up where low latency access is desirable means the need is to read a single record in milliseconds. However in this condition, HBase can provide a solution as it offers a faster access with delayed functionality. Lot of small files can also be problematic when used for Hadoop because it follows the master-slave architecture. The master node stores metadata in RAM and thus the RAM capacity will be unable to handle too many files. Also the situation where parallel write with arbitrary read is seeked, the use of HDFS must be avoided.

4. Fundamentals of HDFS

HDFS like Linux and Windows uses the concept of blocks to store the files but has an added feature to use the available space in a more optimized way. The default block size is 64 MB (may get to 128 MB depending upon the capability of device). If the size of the last block in a file is less than the default size then no padding is introduced, the last block will be stored as such.There are three daemon services provided by HDFS – NameNode, Secondary NameNode and DataNode. NameNode acts as a master node that distributes the work between the DataNodes which acts as slaves and return back the results to the master node. Namenode stores the information about the parameters such as which datanode has which block of file. Thus NameNode is actually a storehouse of metadata. All this information is in RAM. Secondary NameNode comes into action when primary Namenode fails. As status of work is contained in the NameNode so crash of Namenode is accountable. Secondary NameNode is just a copy of NameNode. The concept of checkpoints is used. After every checkpoint, the two NameNodes are made consistent. An additional task of NameNode is the reassignment of blocks to the datanodes in case one of them crashes.DataNodes are arranged in racks. Two network switches are put in the architecture. One switch provides the intra-rack communication while the nother provides the inter-rack communication.

5.Advantages of using HADOOP

Certainly there are a number of factors that boosts the application of Hadoop in industry and the research in academia. Primary advantage of Hadoop is its scalable architecture. It can store and distribute very large data sets across hundreds of inexpensive servers that operate in parallel. Unlike traditional relational database systems (RDBMS) that can't scale to process large amounts of data, Hadoop enables businesses to run applications on thousands of nodes involving thousands of terabytes of data.

Second advantage lies in the fact that Hadoop also offers a cost effective storage solution for businesses' exploding data sets. The problem with traditional relational database management systems is that it is extremely cost prohibitive to scale to such a degree in order to process such massive volumes of data.

In past, in order to reduce the cost, companies used to down-sample data and classify it based on certain assumptions as to which data was the most valuable. While this approach may have worked in the short term, this meant that when business priorities changed, the complete raw data set was not available, as it was too expensive to store. Hadoop, on the other hand, is designed as a scale-out architecture that can affordably store all of a company's data for later use. The cost savings are overwhelming : instead of costing thousands to tens of thousands of pounds per terabyte, Hadoop offers computing and storage capabilities for hundreds of pounds per terabyte.

Flexibility proves to be an important feature in Hadoop by enabling businesses to easily access new data sources and tap into different types of data (both structured and unstructured) to generate value from that data. This means businesses can use Hadoop to derive valuable business insights from data sources such as social media, email conversations or clickstream data. In addition, Hadoop can be used for a wide variety of purposes, such as log processing, recommendation systems, data warehousing, market campaign analysis and fraud detection.

Speed achieved is quite good in Hadoop. It’s unique storage method is based on a distributed file system that basically 'maps' data wherever it is located on a cluster. The tools for data processing are often on the same servers where the data is located, resulting in much faster data processing. If you're dealing with large volumes of unstructured data, Hadoop is able to efficiently process terabytes of data in just minutes, and petabytes in hours

Robustness-The primary objective of HDFS is to store data reliably even in the presence of failures. NameNode failures, DataNode failures and network partitions are the three major failures in a Hadoop architecture. It is gained by approaches like heartbeat and re- replication. The NameNode detects this condition by the absence of a Heartbeat message. The NameNodemarks DataNodes without recent Heartbeats as dead and does not forward any new IO requests to them. In case of re-replication, any data that was registered to a dead DataNode is not available to HDFS any more. DataNode death may cause the replication factor of some blocks to fall below their specified value. The NameNode constantly tracks which blocks need to be replicated and initiates replication whenever necessary.

A key advantage of using Hadoop is its fault tolerance. When data is sent to an individual node, that data is also replicated to other nodes in the cluster, which means that in the event of failure, there is another copy available for use.

6. Applicaton Areas

1. Large Data Sets –HDFS is a successful solution for storing large volumes of unstructured data like tweets.
2. Scalable Algorithms –Minimal inter-process communication and distributed nature of processing makes Hadoop a suitable platfor.
3. Log Management –Hadoop is commonly used for storage and analysis of large sets of logs from diverse locations. Because of the distributed nature and scalability of Hadoop, it creates a solid platform for managing, manipulating, and analyzing diverse logs from a variety of sources within an organization.[5]
4. Extract-Transform-Load (ETL) Platform –Many companies today posses a variety of data warehouse and diverse relational database management system (RDBMS) platforms in their IT environments. Keeping data up to date and synchronized between these separate platforms can be a very problematic. Hadoop enables a single central location for data to be fed into, then processed by ETL-type jobs and used to update other, separate data warehouse environments[5].

7.Conclusion

Hadoop forced with its components such as HDFS serves as a unique platform for a wide variety of applications. Also it has great advantages that it offers to it's users likeflexibility, robustness, cost-effectiveness and so on. Identifying the right application to be run on the platform, it can prove it's worth well.

References

[1] Hadoop Tutorial 11-Limitations of NFS-You Tube
[2] Hadoop Training 2: Deep Dive in HDFS(What is Hadoop)- You Tube
[3] Jimmy Lin and Chris Dyer-"Data- Intensive Text Processing with MapReduce",2010
[4] Joey Jablonski-"Introduction to Hadoop-A Dell technical white paper"
[5] http://www.itproportal.com