RESEARCH ARTICLE OPEN ACCESS

Hadoop Technology

Ankita M.Lahariya
4th year, Department of Computer Science and Engineering, College of Engineering and Technology,Akola.
ankitalahariya24@gmail.com

ABSTRACT
Apache Hadoop is open source software that is freely available from the apache.org source code repository. Hadoop is a free, Java-based programming framework .Big data is an evolving term that describes any voluminous amount of structured, semi-structured and unstructured data that has the potential to be mined for information. When Yahoo, Google, Facebook, and other companies extended their services to web-scale, the amount of data they collected routinely from user interactions online would have overwhelmed the capabilities of traditional IT architectures. So they built their own. Apache Hadoop has emerged as the de facto standard for managing big data.Thus in this way the hadoop technology is abstractly explained.

Keywords: Namenode ,HDFS,PB-PetaByte

I. INTRODUCTION

1.1 What is Hadoop ?

Hadoop is an open source software framework for processing and storing big size data in a distributed fashion on large clusters of commodity hardware. Apache Hadoop is an open-source software framework written in Java platform that offers an efficient and effective method for storing and processing massive amounts of data. Hadoop was created by Doug Cutting and Mike Cafarella in 2005. Cutting named it after his son's toy elephant. It was originally developed to support distribution for the Nutch search engine project. Unlike traditional offerings, Hadoop was designed and built from the ground up to address the requirements and challenges of big data. Hadoop is powerful in its ability to allow businesses to stop worrying about building big-data-capable infrastructure and to focus on what really matters: extracting business value from the data.

II. BIG DATA

A “big” shift is occurring. Today, the enterprise collects more data than ever before, from a wide variety of sources and in a wide variety of formats. Along with traditional transactional and analytics data stores, we now collect additional data across social media activity, web server log files, financial transactions and sensor data from equipment in the field. A new set of technologies has enabled this shift. Now an extremely popular term, “big data” technology seeks to transform all this raw data in meaningful and actionable insights for the enterprise. In fact, big data is about more than just the“ bigness” of the data.

III. HADOOP CLUSTER

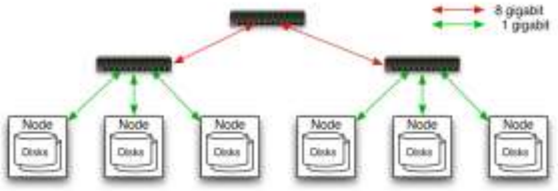

Figure 1: Typical Hadoop Cluster.

A small Hadoop cluster includes a single master and multiple worker nodes. The master node consists of a JobTracker, TaskTracker, NameNode, and DataNode. A slave or worker node acts as both a DataNode and TaskTracker, though it is possible to have data-only worker nodes and compute-only worker nodes. These are normally used only in nonstandard applications .Hadoop requires Java Runtime Environment (JRE) 1.6 or higher. The standard startup and shutdown scripts require that Secure Shell (ssh) be set up between nodes in the cluster.In a larger cluster, the HDFS is managed through a dedicated NameNode server to host the file system index, and a secondary NameNode that can generate snapshots of the namenode's memory structures, thus preventing file-system corruption and reducing loss of data. Similarly, a standalone JobTracker server can manage job scheduling. In clusters where the Hadoop MapReduce engine is deployed against an alternate file system, the NameNode, secondary NameNode, and DataNode architecture of HDFS are replaced by the file-system-specific equivalents.

IV. HADOOP DISTRIBUTED FILE SYSTEM.

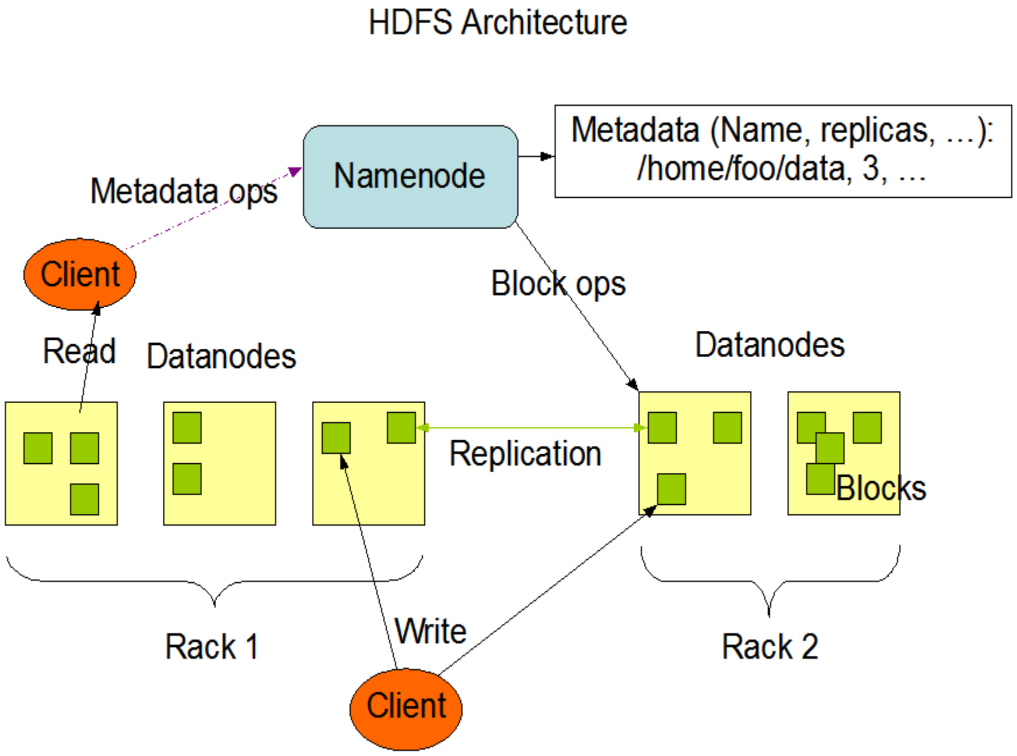

Figure 2: HDFS Architecture

The Hadoop distributed file system (HDFS) is a distributed, scalable, and portable file-system written in Java for the Hadoop framework. A Hadoop cluster has nominally a single namenode plus a cluster of datanodes, although redundancy options are available for the namenode due to its criticality. Each datanode serves up blocks of data over the network using a block protocol specific to HDFS. The file system uses TCP/IP sockets for communication. Clients use remote procedure call (RPC) to communicate between each other.

HDFS stores large files (typically in the range of gigabytes to terabytes) across multiple machines. It achieves reliability by replicating the data across multiple hosts, and hence theoretically does not require RAID storage on hosts (but to increase I/O performance some RAID configurations are still useful). With the default replication value, 3, data is stored on three nodes: two on the same rack, and one on a different rack. Data nodes can talk to each other to rebalance data, to move copies around, and to keep the replication of data high. HDFS is not fully POSIX-compliant, because the requirements for a POSIX file-system differ from the target goals for a Hadoop application. The trade-off of not having a fully POSIX-compliant file-system is increased performance for data throughput and support for non-POSIX operations such as Append.HDFS added the high-availability capabilities, as announced for release 2.0 in May 2012, letting the main metadata server (the NameNode) fail over manually to a backup. The project has also started developing automatic fail-over.

The HDFS file system includes a so-called secondary namenode, a misleading name that some might incorrectly interpret as a backup namenode for when the primary namenode goes offline. In fact, the secondary namenode regularly connects with the primary namenode and builds snapshots of the primary namenode's directory information, which the system then saves to local or remote directories. These checkpointed images can be used to restart a failed primary namenode without having to replay the entire journal of file-system actions, then to edit the log to create an up-to-date directory structure. Because the namenode is the single point for storage and management of metadata, it can become a bottleneck for supporting a huge number of files, especially a large number of small files. HDFS Federation, a new addition, aims to tackle this problem to a certain extent by allowing multiple namespaces served by separate namenodes.HDFS was designed for mostly immutable files and may not be suitable for systems requiring concurrent write-operations.HDFS can be mounted directly with a Filesystem in Userspace (FUSE) virtual file system on Linux and some other Unix systems.File access can be achieved through the native Java API, the

Thrift API to generate a client in the language of the users' choosing (C++, Java, Python, PHP, Ruby, Erlang, Perl, Haskell, C#, Cocoa, Smalltalk, and OCaml), the command-line interface, browsed through the HDFS-UI webapp over HTTP, or via 3rd-party network client libraries.

V. ADVANTAGES AND DISADVANTAGES OF HADOOP

5.1 Advantages

1. Distribute data and computation.The computation local to data prevents the network overload
2. Linear scaling in the ideal case.It used to design for cheap, commodity hardware.
3. Simple programming model.The end-user programmer only writes map-reduce tasks.
4. Fault tolerance by detecting faults and applying quick, automatic recovery
5. Processing logic close to the data, rather than the data close to the processing logic
6. Portability across heterogeneous commodity hardware and operating systems
7. Economy by distributing data and processing across clusters of commodity personal computers
8. Efficiency by distributing data and logic to process it in parallel on nodes where data is located
9. Reliability by automatically maintaining multiple copies of data and automatically redeploying processing logic in the event of failures
10. Ability to rapidly process large amounts of data in parallel
11. Can be deployed on large clusters of cheap commodity hardware as opposed to expensive, specialized parallel-processing hardware
12. Can be offered as an on-demand service, for example as part of Amazon’s EC2 cluster computing service.

5.2Disadvantages of Hadoop

1. Rough manner:- Hadoop Map-reduce and HDFS are rough in manner. Because the software under active development.
2. Programming model is very restrictive:- Lack of central data can be preventive.
3. Joins of multiple datasets are tricky and slow:- No indices! Often entire dataset gets copied in the process.
4. Cluster management is hard:- In the cluster, operations like debugging, distributing software, collection logs etc are too hard.
5. Still single master which requires care and may limit scaling
6. Managing job flow isn’t trivial when intermediate data should be kept

7.Optimal configuration of nodes not obvious. Eg: – #mappers, #reducers, mem.limits.

VI. HARDWARE AND SOFTWARE FOR HADOOP

6.1 Hardware

Hadoop runs on commodity hardware. That doesn't mean it runs on cheapo hardware. Hadoop runs on decent server class machines.

Here are some possibilities of hardware for Hadoop nodes

Table 1: Hardware Specs

	Medium	High end
CPU	8 physical cores	12 physical cores
Memory	16 GB	48 GB
Disk	4disks*1TB=4TB	12disks*3TB=36TB
Network	1GB Ethernet	10 GB Ethernet

6.2 Software

Operating System

Hadoop runs well on Linux.

Java

Hadoop is written in Java. The recommended Java version is Oracle JDK 1.6 release and the recommended minimum revision is 31 (v 1.6.31).So what about OpenJDK? At this point the Sun JDK is the 'official' supported JDK. You can still run Hadoop on OpenJDK (it runs reasonably well) but you are on your own for support .

VII. HADOOP CHALLENGES

This chapter explores some of the challenges in adopting

Hadoop in to a company.

Hadoop is a cutting edge technology

Hadoop is a new technology, and as with adopting any new technology, finding people who know the technology is difficult!

Hadoop in the Enterprise Ecosystem

Hadoop is designed to solve Big Data problems encountered by Web and Social companies. In doing so a lot of the features Enterprises need or want are put on the back burner. For example, HDFS does not offer native support for security and authentication.

Hadoop is still rough around the edges

The development and admin tools for Hadoop are still pretty new. Companies like Cloudera, Hortonworks, MapR and Karmasphere have been working on this issue. How ever the tooling may not be as mature as Enterprises are used to (as say, Oracle Admin, etc.)

Hadoop is NOT cheap

Hardware Cost

Hadoop runs on 'commodity' hardware. But these are not cheapo machines, they are server grade hardware.So standing up a reasonably large Hadoop cluster, say 100 nodes, will cost a significant amount of money.For example, lets say a Hadoop node is $5000, so a 100 node cluster would be $500,000 for hardware

IT and Operations costs

A large Hadoop cluster will require support from various teams like : Network Admins, IT, Security Admins, System Admins. Also one needs to think about operational costs like Data Center expenses : cooling, electricity, etc.

VIII. HADOOP USES CASES AND CASE STUDIES

8.1 Health care

Storing and processing Medical Records

Problem: A health IT company instituted a policy of saving seven years of historical claims and remit data, but its in-house database systems had trouble meeting the data retention requirement while processing millions of claims every day.

Solution:

A Hadoop system allows archiving seven years claims and remit data, which requires complex processing to get into a normalized format, logging terabytes of data generated from transactional systems daily, and storing them in CDH for analytical purposes.

Hadoop vendor:Cloudera

Cluster/Data size**:** 10+ nodes pilot; 1TB of data / day

8.2 Imaging/videos

SkyBox

SkyBox is developing a low cost imaging satellite system and web-accessible big data processing platform that will capture video or images of any location on Earth

Problem:

Analyzing really large volumes image data downloaded from the satellites

Solution:

Skybox uses Hadoop to process images in parallel. Their image processing algorithms are in C/C++. Their proprietary framework 'BusBoy' allows using native code from Hadoop MapReduce Java framework.

Hadoop Vendor: Cloudera and Amazon EC2.

IX. HADOOP DISTRIBUTION

9.1The Case for Distributions

Hadoop is Apache software so it is freely available for download and use.Hadoop in the Cloud Hadoop clusters in the Cloud

Hadoop clusters can be set up in any cloud service that offers suitable machines.

However, in line with the cloud mantra 'only pay for what you use', Hadoop can be run 'on demand' in the cloud Amazon Elastic Map Reduce Amazon offers 'On Demand Hadoop', which means there is no permanent Hadoop cluster. A cluster is spun up to do a job and after that it is shut down - 'pay for usage'.Amazon offers a slightly customized version of Apache Hadoop and also offers MapR's distribution. Google's Compute Engine

Google offers MapR's Hadoop distribution in their Compute Engine Cloud.SkyTab Cloud

SkyTap offers deploy-able Hadoop templates.

X. THE FUTURE OF HADOOP

Hadoop has “crossed the chasm” from a framework for early adopters, developers and technology enthusiasts to a strategic data platform embraced by innovative CTOs and CIOs across mainstream enterprises. These people, who want to improve the performance of their companies and unlock new business opportunities, realize that including Apache Hadoop as a deeply integrated supplement to their current data architecture offers the fastest path to reaching their goals while maximizing their existing investments.

Going forward, Hortonworks and the Apache Hadoop community will continue to focus on increasing the ease with which enterprises deploy and use Hadoop, and on increasing the platform’s interoperability with the broader data ecosystem. This includes making certain it is reliable and stable and more importantly, ready for all and any enterprise workloads.

XI. CONCLUSION

Today, IT organizations and independent users must carefully strategize their approach to dealing with big data to avoid being overrun with data that has no intrinsic value due to the lack of adequate processing tools. Even more importantly, these users need to acknowledge that the right analytic tools, such as Apache Hadoop, present a serious challenge to adoption due simply to the rigorous learning curve. To truly realize the promise of Apache Hadoop and its distributed set of resources for big data analysis, businesses and end-users need to expand their approach by relying on the wealth of resources currently available: access to professional training, commercial platform implementation, and utilizing third-party service providers, such as Cloudera. It's becoming clear that the open-source Apache Hadoop platform changes the economics and dynamics of large-scale data analytics due to its scalability, cost effectiveness, flexibility, and built-in fault tolerance. It makes possible the massive parallel computing that today's data analysis requires.

However, the proper skillset training will be necessary to achieve large-scale data analysis.Organizations can then make appropriate business decisions based on the large amounts of data they accrue by accessing the power of a relatively low-cost, highly scalable infrastructure such as Hadoop to tackle the challenges of big data.Apache Hadoop offer such great value to companies. These integrated management features enable the platform to be implemented by a wide range of users at all levels of skill expertise. Organizations can then make appropriate business decisions based on the large amounts of data they accrue by accessing the power of a relatively low-cost, highly scalable infrastructure such as Hadoop to tackle the challenges of big data.

XII. ACKNOWLEDGEMENT

It is a matter of great pleasure to highlight a fraction of knowledge, I acquired during the time when I was preparing for this topic . This would not have been possible without the guidance and help of many people. This is where I have the opportunity of expressing gratitude from the core of my heart.

Thanks are in order to all the colleagues and friends who knowingly or unknowingly helped me during this work.

REFERENCES

[1] https://twitter.com/HadoopIlluminat
[2] Authors Mark Kerzner, Sujee Maniyam Book Hadoop Illuminated[https:// github. com/-illuminated/hadoop-book]
[3] . http://www.skytap.com
[4] http://en.wikipedia.org/wiki/Hadoop
[5] http://hadoop.intel.com
[6] www.intel.com/bigdata
[7] hadoop.apache.org
[8] http://developer.yahoo.com/hadoop