International Journal of Research in Library Science

ISSN: 2455-104X

ISI Impact Factor: 3.723

Indexed in: IIJIF, ijindex, SJIF,ISI, COSMOS

Volume 2,Issue 2 (July-December) 2016,151-154

Received: 25 Oct .2016 ; **Accepted**: 30 Oct. 2016 ; **Published**: 4 Nov. 2016 ; **Paper ID**: IJRLS-1188

Application of Hadoop Distributed File System in Digital Libraries

Kaladhar Arnepalli[1]; K. Somasekhara Rao[2]

Research Scholar JNTUK-Kakinada[1]; Prof.K. Somasekhara Rao,Dept of Lib.and Information Science, Andhra University, Visakhapatnam[2]

librarian@svecw.edu.in[1]; kalepuss.office@gmail.com[2]

ABSTRACT

Generally we access data in digital libraries through client – server architecture. The maintenance charges of these servers and infrastructure required more amount to invest. Limited amount of data only can be stored in the server the amount of data increases replacement of server is also required. All these constraints can be overcome by using a distributed database system. In this paper, I have chosen Hadoop distributed file system to organize digital library operations. Hadoop is a an open-source software framework for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware. It is well known for its flexibility, fault tolerance, scalability and parallel computation.

Keywords— Hadoop, HDFS, Digital Library

1. INTRODUCTION

A digital library is a system for storing massive amounts of data in a binary, digitally accessible format. With the rapid spurt of technology in this era, a lot of filing cabinet databases are switching over to a digital format. Although digital libraries are conceptually simple enough to comprehend and implement, actual implementation involves a large infrastructural cost and investment. Most of this cost is involved in fulfilling the hardware requirements of maintaining a fully scalable, and fault tolerant architecture. With the increase in number of users or growing demand, a library must allow scaling, and appropriate configuration updates.

Hadoop Distributed File System is a platform which allows easy scalability and solid fault tolerance at a very low implementation cost. It is possible to implement a Hadoop based system on multiple mainstream machines using MapReduce parallelism technique. Hadoop has already seen a rapid acceptance amongst multi-national corporations such as Facebook, Amazon, Yahoo, etc. These corporations have fully functioning Hadoop clusters catering to large amounts of data every day.

2. DIGITAL LIBRARY OPERATIONS IN PRESENT SCENARIO

Using of Client-server architecture the significant drawback of such a system is scalability. Though the scalable servers are used, the cost required for scaling is quite high. Another significant issue with server based systems is downtime and fault tolerance. Higher number of clients, usage of higher amount of data on a single server with low band width the server system gets failed. Failed systems will again lead to downtimes while the server is in the process of rejuvenation. Although server systems provide support for RAID (Redundant Array of Independent Disks) setups, these setups tend to be expensive, and are still not fully resistant to faults. There are systems which have also been implemented on cloud architecture using the PaaS (Platform as a Service) approach. This system suffers from scalability issues. PaaS systems are not easy to scale, and require a significantly higher infrastructure cost as compared to client-server systems. Most institutions and universities wouldn't go for cloud setups for this very reason.

3. HADOOP DISTRIBUTED FILE SYSTEM (HDFS) AND DIGITAL LIBRARIES

The reason behind using HDFS is to maintain a distributed database of files. Any number of files need to be uploaded or accessed will be available on a cluster of data-nodes which can be scaled as per requirements. In order to implement an efficient architecture for scalable database systems, we can adopt Hadoop as a base for digital library operations. It provides a robust system which allows scaling and maintenance with zero downtime. The functionality of the system will be continues even if there is any failure in other computer components.

An Apache Tomcat server will be used as a frontend to implement a web-based user interface. Web-based interface allows the system to be platform independent, catering to a wide range of users. The Apache Tomcat server will merely act as an interface between Hadoop and end-users. Using a web-interface also allows us to maintain security while accessing data. It allows multiple users to segregate and manage their data individually without any interference from other users. In its current stage, the system can be implemented on an institution's intranet. All client machines with a web browser are supported. The interface will be designed using HTML5, CSS3, and some elements of jQuery. This will allow an interactive environment for users ensuring an spontaneous user interface.

4. IDEA OF HDFS IMPLEMENTATION

The current operations in digital library system employ a simple approach to database management, systems like MySQL, SQL,RDBMS etc. are using in a closed system. Google Apps also are often being used for this purpose. These systems do not offer the flexibility of searching through shared files or folders. It also does not provide local or intranet based solutions for institutions. Another limitation is the requirement of high-speed internet to upload large volumes of data. This requirement can be eliminated if we deploy the entire system on an internal network. The library system was designed with the idea of using just the internal network of an institution. It will allow all members of the institution to upload, share and search through files on the network. It will also help in maintaining a repository of projects, notes and papers written by all students of the institution.

5. SYSTEM ARCHITECTURE

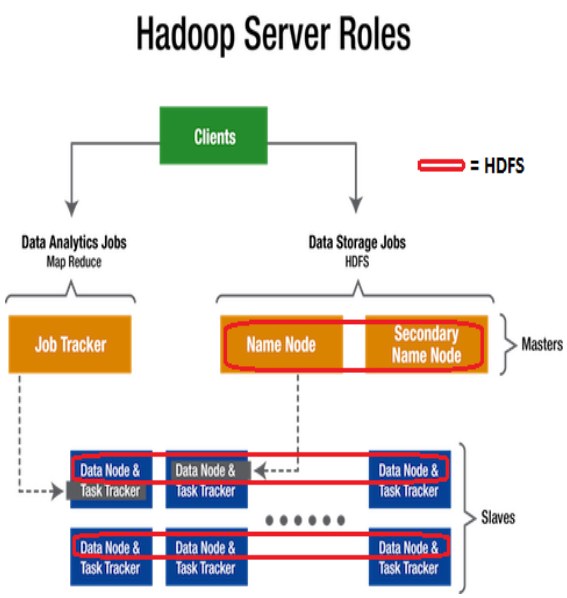

Fig 1: Hadoop System Architecture

Hadoop is designed to increase the performance from single server to thousands of server. There are two major layers:

1)HDFS 2) MapReduce

HDFS that is Hadoop Distributed File System. Itis storage purpose. It has components namely

•Datanode(Slave)

•Namenode(Master)

•Secondary Namenode

Namenode consist the name of data. Datanode consist the Data. Secondary namenode is use for backup purpose whenever the namenode get fails then it retrieve data from secondary namenode. Mapreduce is used to analysis large Dataset in map reduce data is written ones and read many times.It have two functions Mapper and Reducer. It has component namely

•Jobtracker(Master)

•Tasktracker(Slave)

Jobtracker keep the record of each job and assign job to the tasktracker. It also monitoring the jobs. Tasktracker run task or job which is assign by jobtracker and produce the output. Mapper map the input coming from master and given as a output to the reducer. Reducers reduce that output which is coming from mapper and produce the final output.

6. SYSTEM FEATURES

6.1 Availability

The system will be available 24/7 as it will be deployed over an intranet.

6.2 Flexibility

System is highly flexible, since a web user interface is being provided. The user can login from any computer and work from anywhere within an organization's campus.

6.3 Interoperability

Ability of HDFS system to work with other systems is unmatchable. It works very well operating within its own namenode and datanode.

6.4 Reliability

HDFS itself is reliable.

6.5 Robustness

The version hadoop which has been used has a secondary namenode which acts as primary namenode when the namenode fails to work. Hence, providing higher fault tolerance.

6.6 User Interface

The System provides a web interface with the use of languages like HTML5, CSS3, AJAX, Jquery,

CONCLUSION

Installation of Hadoof distributed file system on commodity hardware is an easy task, hence the expenditure on server can be reduced. Establishing a cloud computing facility in libraries is very expensive. The HDFS would be able to supplement most of the facilities of cloud computing.. The very nature of above system operations is on an intranet which eliminates intrusions from outside the network. Using HDFS as a base, the problems like server scalability and high infrastructure cost are limited or negligible. Downtimes can be cut down significantly by the

high availability features present in the system which can serve a large number of users without any interruption. By using a web interface, the need for platform dependent applications is eliminated, allowing users to access the system on any machine within an organization's campus. The functionality of this system can be easily understandable by the users. The HDFS model of digital library operations would resolve most of the basic infrastructural issues of an organization.

References

1. Anam Alam, "Hadoop Architecture and Its Issues", IEEE, 2014
2. Tom White, "Hadoop The definitive guide", O'Reilly
3. Farag Azzedin, "Towards scalable HDFS architecture", IEEE, 2013
4. https://www.google.co.in/search?q=hdfs+architecture+images
5. Kala Karun. A, Chitharanjan. K, "A review on hadoop — HDFS infrastructure extensions",IEEE 2013

