A Digital Resource System on HDFS

Ebenezer Komla Gavua, Seth Okyere-Dankwa, Collinson Colin Agbesi

Abstract: ***The issue of managing digital resources has become a major concern for educational and research institutions expanding with research laboratories, schools and departments. Some of these institutions have a major challenge as to how to manage huge quantities of educational resources required for research, teaching and learning. It is due to this challenge that this paper sought to design and implement a digital resource management system on the Hadoop Distributed File System (HDFS). The main model implemented in the development of the system was the incremental and iterative model. The development of the application took into consideration the user requirements and what the system sought to achieve after implementation. The system's architecture is divided into various levels and these are the user, systems management, data storage and infrastructural levels. The development of this application was focused on the systems management level. The system integration was achieved by connecting MYSQL database server with HDFS through the utilization of sqoop. Security features were implemented on the system to protect the system from attack. The system was tested to ensure that all the modules created were communicating perfectly and the system was producing the expected results.***

Keywords: Digital library; Hadoop distributed file system, cloud storage.

I. INTRODUCTION

The management of digital resources is an important aspect of Information and Communication Technology which is currently given much consideration since almost every of aspect of modern life is controlled by information. Due to the important considerations given to institutional information, many organizations are willing to make funds and infrastructure available for the provision of quality information which can be accessed quickly and reliably. It is due to this premise, that most educational institutions have made the great effort to provide libraries for their teachers and students to promote quick access to quality information. Most educational institutions are making efforts to create digital libraries due to the flexibility provided by Information and Communication Technology. A digital library is a place where various forms of information material such as journals, monograph, visual materials, voice recorder and moving pictures can be retrieve via the Internet (Shiri,2003)

Current digital libraries have well researched user interfaces, architectures that allow ease of use and permit various levels of interactivity including searching and browsing. They are aimed to help users to retrieve useful information easily and quickly (Salim et al.2008).

Revised Version Manuscript Received on June 30, 2017.

Ebenezer Komla Gavua, Koforidua Technical University, Ghana E-mail: mgavua@yahoo.com

Seth Okyere-Dankwa, Koforidua Technical University, Ghana E-mail: Sokyeredankwa@yahoo.com

Collinson Colin Agbesi, Koforidua Technical University, Ghana E-mail: Koliny3k@yahoo.com

The benefits provided by these digital management systems are enormous and as such there are always opportunities to improve their performance through the adopting of current of technologies to promote scalability and interrupted operations.

II. BASIC CONCEPT OF CLOUD STORAGE

Cloud storage is one of the primary use of cloud computing. With the cloud storage, data is stored on multiple third party servers, rather than on the dedicated servers used in traditional networked data storage. When storing data, the user sees a virtual server that is, it appears as if the data is stored in a particular place with specific name. But that place does not exist in reality. It is just a pseudonym used to reference virtual space carved out of the cloud. In reality, the user's data could be stored on any one or more of computers used to create the cloud (Liu et al.2009).

The actual storage location may even differ from day to day or even minute to minute, as the cloud dynamically manages available storage space. But even though the location is virtual, user sees a static location for his data and can actually manage his storage space as if it were connected to his own pc. A typical cloud storage system architecture includes a master control server and several storage servers, as shown in figure 2. At its most basic level, a cloud storage system needs just one data server connected to the internet. A client sends copies of files over internet to the data server, which then records the information.

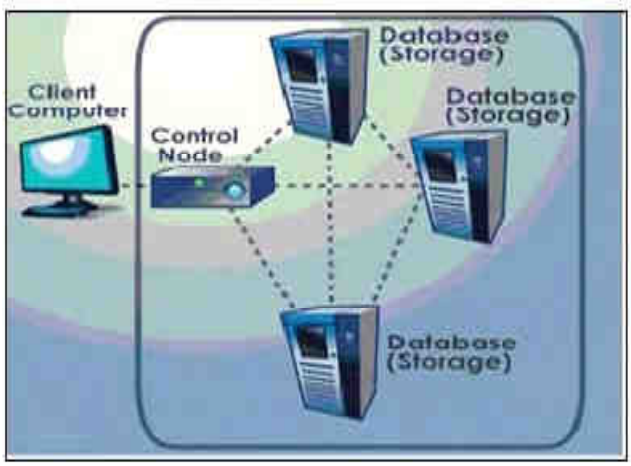

Figure 1 Typical Cloud Storage System Architecture

When client wishes to retrieve the information, he or she accesses the data server through a web based interface. The server then either sends the files back to the client or allows the client to access and manipulate the files on the server itself. Cloud computing is a current trend that considers the Internet as a platform providing on-demand computing and software as a service to anyone, anywhere, and at any time.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

Digital libraries naturally should be connected to cloud computing to obtain mutual benefits and enhance both perspectives (Zhang et al, 2010).

In this model, storage resources are provisioned on demand and are paid according to consumption. Services deployment in a cloud-computing environment can be implemented three ways: private, public, or hybrid. In the private option, infrastructure is operated solely for a single organization; most of the time, it requires an initial strong investment because the organization must purchase a large amount of storage resources and pay for the administration costs. The public cloud is the most traditional version of cloud computing. In this model, infrastructure belongs to an external organization where costs are a function of the resources used. These costs include administration. Finally, the hybrid model contains a mixture of private and public. A cloud-computing environment is mainly supported by technologies such as virtualization and service-oriented architectures.

A cloud environment provides omnipresence and facilitates deployment of file-storage services. It means that users can access their files via the Internet from anywhere and without requiring the installation of a special application. The user only needs a web browser.

The three main types of cloud services are Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS) (Zhou et al, 2010). First, cloud computing offers the ability of libraries to use online software to handle a task like video chat through either Gmail video chat or through Skype. Both of these are free services though there is “little customization or control available with these applications”(Kroski ,2010).

In other words, services you offer through a SaaS’ interface will look like that of your competitors which will not distinguish you from them. On the other hand, since the services and application interfaces are often familiar with users, there would be a decrease in the learning curve for library staff and users. Second, libraries can create applications in an online environment.

III. HADOOP

Hadoop (Padhy et al, 2012) is an open source large scale distributed filesystem which consists of HDFS, Map Reduce, HBase, Hive and Zookeeper and other projects modeled after the Google File System (Borthakur ,2007). Hadoop has two primarily parts: Hadoop distributed file system (HDFS) and MapReduce programming model. MapReduce computing model is divided into two parts, Map and Reduce. Map Reduce model split a calculation job into a number of Maps, and then assign to different nodes to compute. Programming input data to each Map job, after this step will input the intermediate data to Reduce job. Reduce job is to aggregate the Map intermediate data together and output.

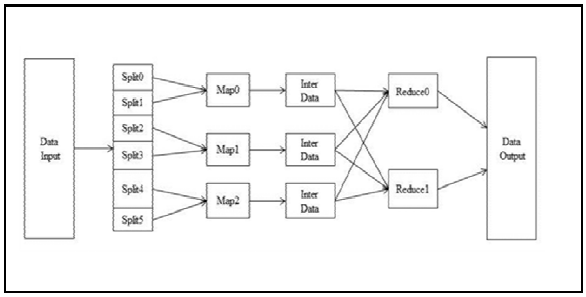

Figure 2 MapReduce Model

Hadoop provides extreme scalability on commodity hardware, streaming data access, built-in replication, and active storage. Hadoop is well known for its extensive use in web search by companies such as Yahoo! as well as by researchers for data intensive tasks like genome assembly (Schatz, 2009).A key feature of Hadoop is its resiliency to hardware failure. The designers of Hadoop expect hardware failure to be normal during operation rather than exceptional (Ghemawat,2003) .Hadoop is a ”rack-aware ” filesystem that by default replicates all data blocks three times across storage components to minimize the possibility of data loss. Large data sets are a hallmark of the Hadoop file system

IV. HDFS SYSTEM ARCHITECTURE

HDFS (Hadoop Distributed File System) adopt Master/Slave architecture (Borthaku, 2008) and can run on the common hardware. HDFS cluster consists of a NameNode and multiple Datanode. Namenode is a center server for managing metadata.It is used to manage the namespace of file system and the access from client to files. One node generally has a DataNode in cluster. DataNode is responsible for managing the attached storage of a node.

HDFS shows the file system namespace by file form. A file is divided into many Blocks and stored in Datanode set. Namenode is used to open, close and rename the files and directories, and establish the mapping between the Block and Datanodes. Datanode is responsible for responding to the demand of customer to read and write. DataNode is used to set up, copy and delete the Block under the command of Namenodes at the same time. The architecture of HDFS can be seen in figure 3.

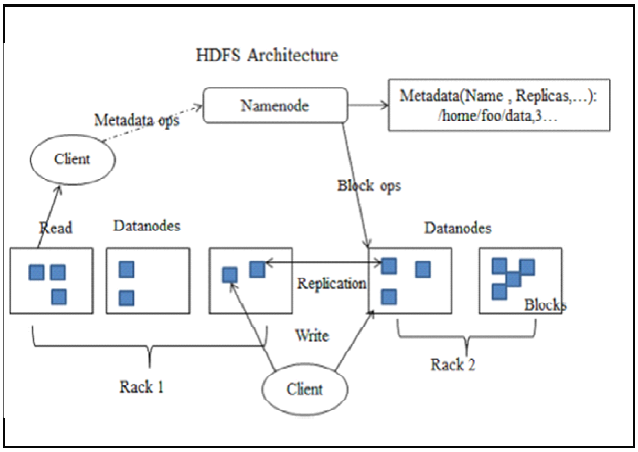

Figure 3. HDFS Architecture

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

The Hadoop Distributed File System (HDFS) library is designed to detect and handle failures at the application layer. As a result, the HDFS is a highly-available service that can be installed on commercially available clusters of computers, which have not been built with any special care for minimizing the failure rate.

The HDFS is a file system designed for storing very large files with a streaming data access patterns, running on clusters of commodity hardware. There are hadoop clusters running today that store petabytes of data (Yu et al ,2011).

Unlike Lustre and PVFS, the DataNodes in HDFS do not use data protection mechanisms such as RAID to a make the data durable. Instead, like GFS, the file content is replicated on multiple DataNodes for reliability. While ensuring data durability, this strategy has the added advantage, that data transfer bandwidth is multiplied, and there are more opportunities for locating computation near the needed data.

The HDFS architecture makes is mostly suitable now for cloud computing as its namespace is a hierarchy of files and directories. Files and directories are represented on the NameNode by inodes, which records attributes like permissions, modification and access times , namespace and disk space quotas. The file content is split into large blocks (typically 128 megabytes, but user selectable file-by-file) and each block of the file is independently replicated at multiple DataNodes (typically three ,but user selectable).The NameNode maintains the namespace tree and the mapping of file blocks to DataNodes(the physical location of file data).

An HDFS client wanting to read a file first contacts the NameNode for the location of data blocks comprising the file and then reads blocks contents from the DataNode closest to the client. The user code does not need to know about the namenode and datanode functionality because for the POSIX-like (Wang et al, 2011) file system interface between the Hadoop client and the file system.

When writing data, the client requests the NameNode to nominate a suite of three DataNodes to host the block replicas. The client then writes data to the DataNodes in a pipeline fashion. The cluster can have thousands of DataNodes and tens of thousands of HDFS clients per cluster, as each DataNode may execute multiple application tasks concurrently.

HDFS keeps the entire namespace in RAM. The inode data and the list of blocks belonging to each file comprise the metadata of the name system called the image. The persistent record of the image stored in the local host's native files system is called a checkpoint. The NameNode also stores the modification log of the image called the journal in the local host's native file system. For improved durability, redundant copies of the checkpoint and journal can be made at other servers. During restarts the NameNode restores the namespace by reading the namespace and replaying the journal. The locations of block replicas may change over time and are not part of the persistent checkpoint.

V. APACHE SQOOP

Apache Sqoop(TM) is a tool designed for efficiently transferring bulk data between Apache Hadoop and structured datastores such as relational databases. A user can use Sqoop to import data from a relational database management system (RDBMS) such as MySQL or Oracle into the Hadoop Distributed File System (HDFS), transform the data in Hadoop MapReduce, and then export the data back into an RDBMS. The connection between hadoop and sqoop is shown below in figure 4.

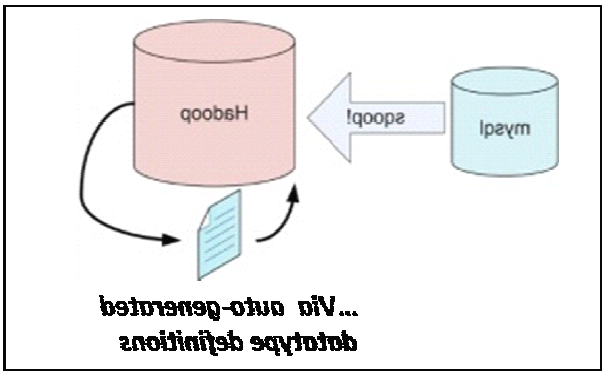

Figure 4. Sqoop

With Sqoop, a user can import data from a relational database system into HDFS. The input to the import process is a database table. Sqoop read the table row-by-row into HDFS. The output of this import process is a set of files containing a copy of the imported table. The import process is performed in parallel. For this reason, the output will be in multiple files. These files may be delimited text files (for example, with commas or tabs separating each field), or binary Avro or SequenceFiles containing serialized record data.

A by-product of the import process is a generated Java class which can encapsulate one row of the imported table. This class is used during the import process by Sqoop itself. The Java source code for this class is also provided to the user for use in subsequent MapReduce processing of the data. This class can serialize and deserialize data to and from the SequenceFile format. It can also parse the delimited-text form of a record. These abilities allow a user to quickly develop MapReduce applications that use the HDFS-stored records in your processing pipeline.

After manipulating the imported records (for example, with MapReduce or Hive) the user may have a result data set which can be exported back to the relational database. Sqoop's export process will read a set of delimited text files from HDFS in parallel, parse them into records, and insert them as new rows in a target database table, for consumption by external applications or users.

VI. SYSTEMS ARCHITECTURE

The whole system consists of various levels. These include the User level, Systems management level, Data Storage level and the Infrastructure level. The various sections can be seen in the diagram below.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

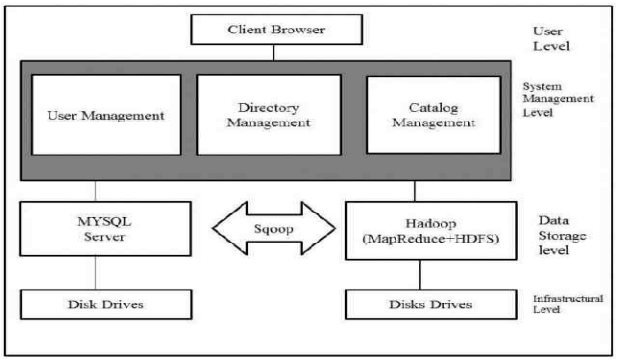

Figure 5 Systems Architecture

i. User level

This is the client / browser used to display the application service interface. The user sends a request to the system via a client and the system return the information to the client. It is at this level that the user can interact with the application. Logging –in, Registration, etc all takes place at this level.

ii. System Management Level

This level is responsible for the connection between the application and the underlying data. It returns the data result set which the user requested the client. It mainly includes user management, directory management and catalog management. Most of the management activities are undertaken at this level.

iii. Data Storage Layer

The implementation of Hadoop, whether stand-alone or cluster works in this layer. It consists of HDFS and MapReduce. It is responsible for data management and task allocation; provide distributed computing and storage for the system. The database management System and Sqoop, which is the link between relation databases and Hadoop, are also utilized in this section. In order to promote efficiency in data storage and retrieval, the data is divided into two categories; transactional data and archival data.

- **Transactional Data**

All transactional data (files that are accessed more often) are stored in the database management system .These includes frequently accessed eBooks, scientific papers and lectures. This will ensure efficiency in data storage and retrieval.

- **Archival Data**

The rest of the big data (archival data) are kept on the Hadoop distributed file system (hdfs) which is sitting on the disks. The disk system used in this application to ensure reliability and fault tolerance is RAID (Redundant Array Independent Disks). RAID offers fault tolerance and higher throughput levels than a single hard drive or group of independent hard drives and it is essential for a broad spectrum of client/server applications. Sqoop is implemented in the system architecture to assist in imported files from the database management system to hdfs and for exporting files from hdfs to the database management system.

- **Infrastructure Level**

It refers to the hardware infrastructure which mainly includes the disk, server etc.

VII. SYSTEMS FUNCTIONAL DESIGN

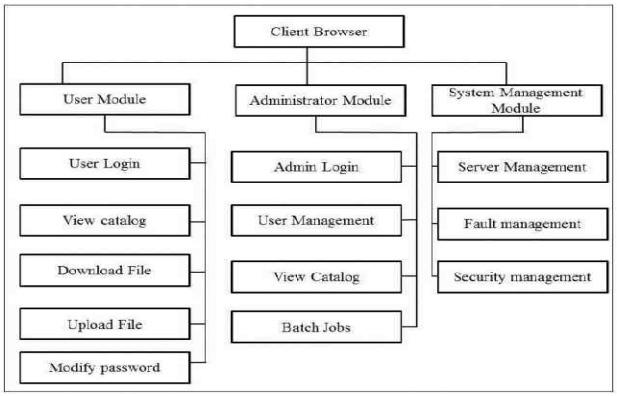

Figure 6 Systems functional Design

- **User Module**

These include the user login, view catalog, download and upload files and modifying of passwords. This module is mainly for the users of the application. The users can be available only after being related and activated by the administrator.

- **Administrator Module**

The administrative module is the module where the initialization processes are enforced. These include a lot of batch jobs. The batch jobs include the upload of files into the database before allowing users to start using the application

- **System Management Module**

These include the fault monitoring, server information management and security management. The provision of server management and maintenance information is provided at this module. This ensures that the server is constantly monitored and tuned up to provide efficient operations. The system is constantly monitored for any fault within the application and error messages are relayed to the administrator. Security management is implemented at this module to protect the system from attacks. The systems security policies are developed and implemented at this module

VIII. CONFIGURING THE HADOOP DISTRIBUTED FILE SYSTEM AND SQOOP

In order to run the Hadoop on a cluster, it is very important that, certain key configurations are done, so that, the system can be activated and utilized.

Configure SSH for cluster in order to log into different machine without using password.

The operation command is as follows:

$ ssh-keygen -t rsa

$ ssh-copy-id -i $HOME/.ssh/id_rsa.pub username@slave

Generate SSH key in every machine, exchange public key and copy public key of NameNode to each DataNode. So they can visit each other without password.

$ cp id_rsa.pub authorized keys

Install JDK1.6 on every machine and configure Java environment.

Hadoop configuration steps:

Configure

Hadoop-env.sh

HADOOP_HOME variables

Environment variables:

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

Export JAVA_HOME=/usr/java/jdk1.6.0_22/
Configure masters file and add IP address of NameNode.
Configure slaves file and add IP address of all DataNodes.
Configure Core-site.xml, Hdfs-site.xml and Mapred-site.xml file.
Copy hadoop1.0.3 which has been configured on NameNode to corresponding directory of other datanodes.
Start Hadoop service. Operate the following command in the directory of hadoop-
0.21.0/bin of NameNode:
$ Hadoop NameNode –format
$ start-all.sh
Alternatively, the file system can be started only by using start-dfs.sh, or start
Map-reduce job by start-mapred.sh. To stop the cluster, use command stop-all.sh

IX. CONCLUSION

Educational Digital libraries are designed to enable users to gain access to unlimited educational resources at real time. The ability to access such materials encourages students, teachers and researchers to continue pursuing their educational goals wherever they are located.

The digital management system was developed on a cloud architecture to enable educational institutions manage their huge data produced daily especially those for library. The Incremental and Iterative process model was employed together with Unified Modeling Language for the systems analysis and design. The strategy design was employed in the implementation stage generate java classes for effective communication of system modules. Apache Sqoop was employed for integrating the application unto the cloud architecture. The implementation of this digital resource on Hadoop distributed file system has enabled a huge amount of educational resources to be provided at the disposal users.

REFERENCES

1. Borthakur, D. (2007) The Hadoop distributed file system: Architecture and design. http://hadoop.apache.org/.
2. Fisher, W.(2002). The electronic resources librarian position: a public services phenomenon.
3. Library Collections, Acquisitions, & Technical Services, 27,1, 3-17.
4. Gantz et al.(2007) “The Expanding Digital Universe: A Forecast of Worldwide Information Growth Through 2010,” An IDC White Paper—sponsored by EMC.
5. Ghemawat, S. Gobioff, H. & Leung, S.(2003) The Google filesystem. In ACM Symposium on Operating Systems Principles.
6. Han, Y.(2010) “On the Clouds: A New Way of Computing,” Information Technology & Libraries 29, no. 2: 87–92;
7. Ipri,T.(2011) “Where the Cloud Meets the Commons,” Journal of Web Librarianship 5, no. 2 (2011)
8. Itani,W., Kayssi, A.& Chehab, A. (2009) “Privacy as a Service: Privacy-Aware Data Storage and Processing in Cloud Computing Architectures”, Eighth IEEE International Conference on Dependable.
9. Jordan, J. (2011)“Climbing Out of the Box and Into the Cloud: Building Web-Scale for
10. Libraries,” Journal of Library Administration 51, no. 1: 3–17, doi:10.1080/01930826.2011.531637.999
11. Kroski, E.(2010) Library Cloud Atlas: A Guide to Cloud Computing and Storage | Stacking the Tech. Retrieved November 5, 2010, from Library Journal.com:
12. Liu, C. & Xiao, H. (2009) The Construction and Operation of Digital Library Service based on Grid Technology Padhy S.C. & Mahapatra,R.K (2012)Cloud Computing : Academic Library in Orissa
13. Salim, J., Yahya,Y., Rashid,N.R.M.,& Othman,S. (2008) "Diglib: Gateway to Digital Libraries and Searchable Databases," The International Journal of Learning, vol. 14, pp. 715.
14. Schatz., M.C.(2009) Cloud Burst: highly sensitive read mapping with MapReduce. Bioinformatics, 25(11):1363–1369.
15. Shiri, A. (2003). "Digital library research: current developments and trends," Emerald Group Publishing Limited, vol. 52, pp. 198-202 .
16. Skibiński, P. & Jakub Swacha, J.(2009) “The Efficient Storage of Text Documents in Digital Libraries,” Information Technology & Libraries 28, no. 3: 143–53.
17. Turner , N.(2009) “Cloud Computing: A Brief Summary”, Lucid Communications Limited,2009.Vaquero, L.M., Merino, L.R., Caceres, J. & Lindner, M. (2009)”A break in the clouds: towards a cloud definition,” ACM SIGCOMM Computer Communication Review, 39(1).
18. Wang, C.C., Pai, W.C. & Yen, N.Y. (2011). “A Sharable e-learning Platform Based on Cloud Computing”,3rd International Conference on computer Research and Development (ICCRD),pp.1-5,2011
19. Wang, Y., Wen, X., Sun ,Y., Zhao, Z., & Yang, T.(2011) “The Content Delivery Network System based on Cloud Storage”, IEEE.
20. Wilder, S.J. (2002) New hires in research libraries demographic trends and hiring priories. ARI, 221, 5.
21. William Stallings, “Network Security Essentials: Applications and Standards”. Third Edition Yu, S., Gui, X., Huang, R., & Zhuang, W. (2011). “Improving the storage efficiency of small files in cloud storage,” His-An Chiao Tung Ta Hsueh/Journal of Xi’an Jiaotong University, vol.45, no.6.
22. Zhang, D., Zhu ,L.,& Zhen-yu, H.,(2010) "Research of cloud storage technology for mobile terminal based on WEB, "Computer Engineering and Applications, vol. 46, No.36,pp. 66--69.
23. Zhou, K., Wang, H. & Li C.(2010) "Cloud Storage Technology and Its Application," ZTE COMMUNICATIONS, vol. 16, No. 04,pp. 24-27

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.