BIG DATA DIGITAL REFERANCE TECHNOLOGY

Under The Guidance Of :- Mrs. Pratiksha R. Deshmukh

CHAUDHARI TEJASVINI P.
Computer Department,
Government College of Engineering
and Research,Awasari(Pune).
Pune,India.

MULAY AJINKYA S.
Computer Department,
Government College of Engineering
and Research,Awasari(Pune).
Pune, India.

LOHAKARE AVINASH K.
Computer Department,
Government College of Engineering
and Research,Awasari(Pune).
Pune,India.

NAWARE CHETAN G..
Computer Department,
Government College of Engineering
and Research ,Awasari(Pune).
Pune, India.

ABSTRACT

Flexible digital library systems need to be able to accept or import documents and meta- data in a variety of forms, and associate metadata with the appropriate document. It analyses the requirement the requirements of import process for general digital libraries. Some of the open ended requirements must be extensible with the existing architecture that facilitates the additional new features with the ”HADOOP”.
An experimental digital library gains more wider acceptance and more significant user bases, it become more important to investigate the way in which the user interact with the system in practice along with the transaction logs. Even though there are many advantages of existing digital libraries like ”Green Stone” which is used widely along with features like handling the advanced digital data. But the most prominent feature of the system that is going to be proposed is the use of new technologies like use of Hadoop clustering to handle the ”Big Data” in an easy way.

I. 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.

II. EXISTING SYSTEM

Most digital library systems today use a client-server architecture. The most significant drawback of such a system is scalability. Although servers are considered to be scalable, the cost required for scaling is quite high. This also includes maintaining the server on a regular basis. Another significant issue with server based systems is downtime and fault tolerance. Server systems consisting of a single server unit, have a higher tendency to fail in case of higher loads or bandwidth surges. Failed systems will again lead to downtimes while the server is in the process of rejuvenation. Although server systems provide

ISSN: 2278 – 1323

support for RAID 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.

III. PROPOSED SCHEME

In order to implement an efficient architecture for scalable digital library systems, we can use Hadoop as a reliable foundation. Hadoop provides a robust system which allows scaling and maintenance with zero downtime. Fault tolerance of such a system can be configured dynamically allowing more than 50% of the system to fail while maintaining full functionality. Keeping these features in mind, we have proposed the use of Hadoop Distributed File System for maintaining a distributed database of files. All files which need to be uploaded or accessed will be present on a cluster of data-nodes which can be scaled as per requirements. 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 intuitive user interface.

Features Of Proposed System

1.Use of HADOOP to handle big data.
2.Auto suggestion facility using Ajax.
3.Use of MD5 algorithm for user authentication and security purpose.
4.Online video play capability.
5.Online book printing facility.
6.Ability to handle unstructured data

IV. PROJECT IDEA

- Basically to give all related information to the user for education.
- To facilitate the the user with easy central authorized access .
- To digitalize reading.

Project Scope

The Digital Services has a combined focus on technical imaging and visual information access and retrieval, enabling us to develop optimal methods for creating and providing access to experimental electronic les of visual information, as well as studying its use in the digital environment. These project addressed the following critical issues that face both research and practice in the area of digital imaging.

1.Digital Preservation.
2.Access to Collections.
3.Formats, Metadata and Vocabulary Standards.
4.Search and Retrieval of Visual Information.
5.Visual Resources for Instruction and Research.

V. Working and Implementation Methodology

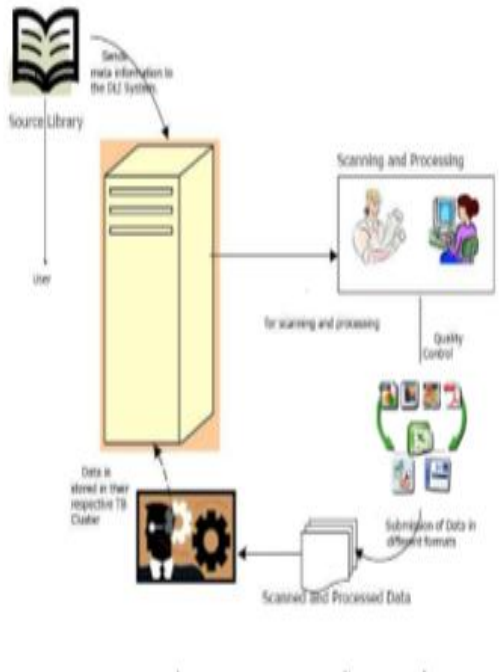

Figure 5.1: Architecture and System Overview

Our main objective is to set up a Digital Library Engine on HDFS, thus creating an environment

ISSN: 2278 – 1323

where users can upload their files, or retrieve the stored files from the system using a Search Engine provided with the system. Deployment of Digital File Library on HDFS provides higher scalability, reliability and speed data transfer or retrieval.

Platform Support :

- Ubuntu.
- IDE: Eclipse, Tomcat Server, Apache Hive,
- JSP,

VI. SRS (SYSTEM REQUIREMENT SPECIFICATION):

Functional Requirements

Admin Prerequisite for all requirements below

- Title- Uploading library specific data

Description-This action is done to add new book to library book collection.

- Title- Delete / modify library data

Description-This event is to delete an existing data or modify its information.

- Title- Validate user account

Description-When a new member sign up then he should wait for acceptance by Administrator according to library policies.

- Title delete member

Description-Admin can delete a member due to some specific rules.

- Validating the library data

Description Admin can validate the data being uploaded by users.

Common Functions

- Title-login

Description both Admin and members must be logged in before they modify any information.

- Title-search for book

Description-User or admin wants to search some book by name, author or subject etc.

Non-functional Requirements

- Error handling

WLMS product shall handle expected and non-expected errors in ways that prevent loss in information and long downtime period.

- Performance Requirements

The system shall accommodate high number of books and users without any fault. Responses to view information shall take no longer than 5 seconds to appear on the screen.

- Safety Requirements

System use shall not cause any harm to human users.

Minimum Hardware Requirements

- System : Intel i3 or above.
- Ram : 4Gb or more at server node and 2Gb or more at client node.

Minimum Software Requirements

- Platform :Windows/Linux
- Coding Language :JSP/Java/Ajax
- Database : HDFS

VII. CONCLUSION

Digital libraries are thoroughly needed world wide to handle the large amount of data. But for this purpose not only the SQL database should be used , for overcoming some problems related to SQL-based systems the new Hadoop-based systems are introduced and this satisfies the need of handling the big data. And hence named as “BIG DATA DIGITAL REFERANCE TECHNOLOGY”.

REFERENCES

[1] Farag Azzedin, “Towards scalable HDFS architecture”, IEEE, 2013

[2] Kala Karun. A, Chitharanjan. K, “A review on hadoop — HDFS infrastructure extensions”,IEEE 2013

[3] Anam Alam, “Hadoop Architecture and Its Issues”, IEEE, 2014

ISSN: 2278 – 1323

[4] Weiming Lu, Liangju Zheng, Jian Shao, Baogang Wei, Yueting Zhuang, "Digital Library Engine: Adapting Digital Library for Cloud Computing", IEEE, 2013
[5] Tom White, "Hadoop The definitive guide", O'Reilly

ISSN: 2278 – 1323