Paper ID: CSEIT38

INFORMATION STORAGE AND MANAGEMENT SYSTEM (ISMS) USING HADOOP

Vishal Mali1,
Vijay Bhosale
Kiran Asawale
Pooja Mokashe
Manoj A. Patil
Rajarambapu Institute of Technology, Rajaramnager Maharashtra, India,

***Abstract*— In today's era data management becomes difficult task for organization. In day to day life, data increase rapidly which results in loss of data and its accuracy. This kind of increasing data is called as Big Data. Big Data contains large amount of data which need to manage efficiently and accurately. So the motivation behind this system is to manage the data efficiently and accurately and we have chosen the recent technology, called "Hadoop". Apache Hadoop is an open-source software framework written in Java for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware. Hadoop allows the data to be processed faster and more efficiently. Hadoop makes it possible to run applications on systems with thousands of nodes involving thousands of terabytes. Distributed file system facilitates rapid data transfer rates among nodes and allows the system to continue operating uninterrupted in case of a node failure. This approach lowers the risk of catastrophic system failure, even if a significant number of nodes become inoperative. RIT is one of the reputed and well-known colleges in Kolhapur University and many committees like National Board of Accreditation (NBA), National Assessment and Accreditation Council (NAAC), Local Inquiry Committee (LIC) etc. visit to RIT for accreditation. These committees need historical data, which is 3-5 years old data to assess college performance. This data is having variety and is kind of a big data and to manage the big data, we are using Hadoop Technology. This technology is highly scalable technology which stores structured and unstructured data. By using Hadoop, the data storage and management is easier and it is also an open source technology. Data storage is done using Hadoop Distributed File System (HDFS) and manipulation is done using Map and Reduce Technique. Mapper maps the given data as key-value pair then data is shuffled across and sorted in an order and Finally it is reduced by reducer technique. The proposed system provides separate portal to admin and faculty to update their information.**

Index Terms — Big Data, Hadoop, HDFS, NameNode, DataNode, Map Reduce, Hive

I. INTRODUCTION

Rajarambapu Institute of Technology College is one of the respected engineering colleges in Shivaji University. Many Committees visit to RIT for accreditation at a regular interval like LIC, NBA, NAAC etc. These Committees assess college performance in a given period of time and they require huge amount of Information. Now this historic Information about RIT is increasing day-by-day because it is an autonomous institute. Its increasing ratio becomes exponential high and it is not only in the form of structured data, as it contains audio, video, image, plain text and numeric data. The conventional database management systems are unable to process this kind of data in integration. This data falls under the category of Big Data and requires one of the new technologies to handle that data. Big Data is a term that describes the large amount of data both structured and unstructured which is hard to manage by conventional Database handling systems. This kind of data contains structured like simple database files and non-structured data like videos, PDF's and excel, word etc type of files and this kind of variety, that amount of data and the increasing rate of data needs to be handled by some intelligent technology.

Because of the reasons stated above we are using Hadoop Technology to manage the data.

II. LITERATURE SURVEY

The related work done in storage and management of data by making use of various technologies or strategies is mentioned below.

In First Survey, the current or existing system, faculty needs to submit all the data to the data operator then data operator puts that data in

respective folders. Whenever committee visits our collage data operator will bring hard copy of that data and submits to the committee.

Following are the Advantages:

 This system provides modularity by storing specific data in a particular folder.

Following are the Drawbacks:

 Tedious work.

 Time consuming work.

 Data management was difficult

In Second Survey, an organization, 'Managing of Database and Information System' is done using DBMS (Database Management System). Efficient management of database and providing access control to the database is the motive of the system.

Following are some advantages of this system:

 Improved data sharing

 Improved data security

 Improved decision making

Following are some drawbacks of this system:

 Database systems require hardware and software which is sophisticated and personnel with good skills.

 To maximize the database efficiency, you must keep your system current/updated.

 It is hard to manage the complex database system

In Third Survey 'The High Performance Storage System' is a system proposed by National Storage Laboratory and IBM Federal Systems Company. The HPSS mainly focuses on scalable systems for parallel storage for parallel computing as well as traditional supercomputers and workstation clusters

Following are the advantages of this system are as follows:

 HPSS is designed using network-connected storage devices which having transfer rate of 100 Mb/S or beyond that.

 They include mechanisms for efficient organization like clustering, partitioning and placement of data under applications control.

 Multiple storage system located in different geographical areas are integrated to one single logical system accessible by client systems.

Following are some disadvantages of this system:

 High level system i.e. requires more specific resources.

 Cost for implementation is high.

 The size and complexity of the systems is increased to process large amount of data in minimum time which results in higher staff and management overhead.

In Fourth Survey, another solution for data handling i.e. 'Managing and Analyzing Large Data Sets' proposes a fact that another biggest challenge of data handling ahead, lies in managing gained data effectively and efficiently.

Following are the advantages of this system that will provide more elegant approach for managing the data in an organization.

 Information stored by means of the data acquisition and processing life cycle makes it easier to trace and track the data. Also it makes data more usable and logical to store.

 Data mining mechanism is used to search or locate the data.

 It makes it easy to process the data: by making report analysis, it is easy to process the data when size of data is still growing regularly.

Following are some disadvantages of this system:

 Data acquisition through selecting proper data mining technique is time consuming.

 Need to select proper data processing tools and techniques.

III. PROPOSED SYSTEM

RIT is one of the reputed and well-known, an autonomous institute in Shivaji University. Many committees like NBA, NAAC, LIC etc. are visit to accreditation. At the time of accreditation, it require more man power to give different kind of information like placement, academic achievements, student information, conference attended by student/staff or paper/journals presentation etc. which results in time consuming activity. As RIT is an autonomous institute, proper documentation should be maintained. But it requires more man power to maintain the information related to institute as department wise. This task carried out by doing paper work and then it filled by them manually and stored at central computer which can be available centrally. Then by accessing centrally stored information, the required reports are generated. This is time consuming job and overhead of paperwork.

The proposed system aims to automate the task of making information available centrally and one person who is admin of the system can generate required reports in minimum time. Previous institute data is in GB's, but now a day's data grows rapidly and it is in TB's/PB's because of an autonomous institute and its accuracy need to maintain. So the proposed system can work with such Big Data and manage it efficiently and accurately. Data mining is useful for searching and generating required reports from these kinds of data.

The objectives of the proposed system are:

 To reduce paperwork.

 To reduce operational time.

 To get work done in less efforts

 To increase accuracy and reliability.

 To increase operational efficiency.

The proposed system will provide following benefits:
 Reduce the paperwork and efforts taken to make that paperwork
 Easy solution to generate flexible reports
 Make all the data available centrally at one place.
 User of the system can fill data from where he/she wants to.
 Data Management will become easier.

Scope of Project:
This system aims to produce efficient reports required as per the need of committees. Proposed system has limited scope i.e. for departments in different colleges. Scope of proposed system can be extended for the university.

Modules:
This proposed system contains 3 main modules:
 Admin module

Admin module contains 3 sub modules, which are staff management in which staff profile is managed, report generation which generates required report and data filling like consultancy, student information, conference attended, staff information etc.

Staff Management		
SrNo	Activities	Description
I.	Add staff Profile	Admin can add new staff profile here.
II.	Remove staff profile	Admin can remove staff profile.
III.	Make changes in staff profile	Admin can make changes in the staff profile as needed.

Report Generation		
SrNo	Activities	Description
I.	Consultancy reports	Generate reports on consultancy information.
II.	Student Training Programs	Generation of reports on different Training programs organized for students.
III.	Remedial lectures for students	Generate reports on Remedial lectures taken for students
IV.	Paper published at conference	Generate reports on Paper published at conference
V.	Award received	Generate reports on different awards received by faculties

.

VI.	National/international conference attended	Generate reports on National/international conference attended by the faculties.
VII.	Training Programs attended by staff	Generate reports on different Training Programs attended by staff
VIII.	Guest/Expert lectures for outside colleges/other departments	Generate reports on Guest/Expert lectures that are arranged for outside colleges/other departments
IX.	Industrial Training visit details	Generate reports on detailed information about Industrial Training visits.

III.	Award received	When a faculty receives any award he can mention it here.
IV.	National/international conference attended	In this section, staff can fill the information about National/ international conference attended by him.
V.	Guest/Expert lectures for outside colleges/other departments	Here staff have to fill the information when he/she takes a guest lecture for other college or other department in our college
VI.	Industrial Training visit details	Staff needs to give detailed information about training visits.

Data Filling		
SrNo	Activities	Description
I.	Consultancy	Admin can fill the consultancy information which contains training programs conducted.
II.	Research projects sponsored by college to students	Here admin need to fill the information about different research projects sponsored by college to students
III.	Training program organized by department for staff	Admin can fill the information regarding different training programs organized by the department to staff
IV.	Training and placement related information.	Admin can fill here the information related training and placement department like no of students placed.

• Staff module
Staff module contains provision for staff to fill information related to their activities and academic reports.

Staff Module		
SrNo	Activities	Description
I.	Training program organized for student.	Here staff need to fill the information as he arranges a training program for students
II.	Paper published at conference	Staff will fill the information about Research paper publication here

• Database module
Database module is used to store and retrieve the data while working with it.

Database module		
SrNo	Activities	Description
I.	Store	Store information about Consultancy, paper published at conference, awards received, national/international conference attended, training program attended by staff, training program organized for student, professional membership, industrial/training visit details, guest/expert/remedial lectures detail, different scan copy photos etc.
II.	Retrieve	Retrieve information about unique id, number of appointments, appointment sequence, reports of the patient. Consultancy, paper published at conference, awards received, national/international conference attended, training program attended by staff, training program organized for student, etc.

Data Flow diagram for ISMS is shown as follows which shows interfacing details of ISMS.

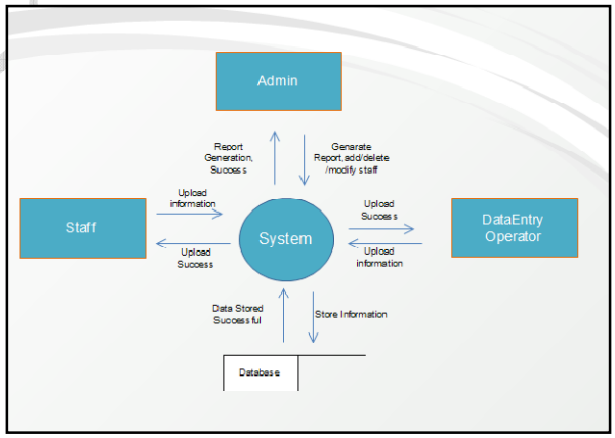

Fig 1 level: 0 Data Flow Diagram

The flowchart for ISMS is as shown below

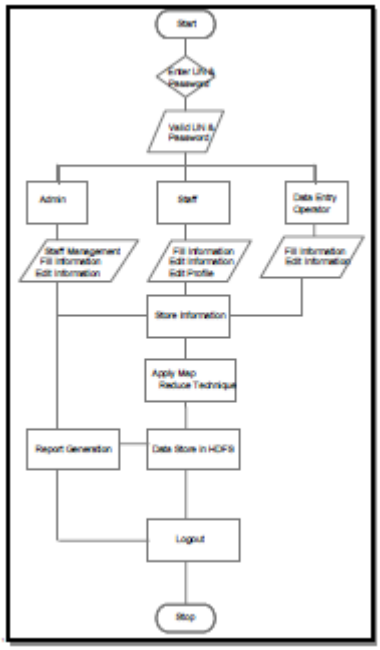

Fig 2 Flowchart of ISMS

.

IV. IMPLEMENTATION

ISMS system is a system in which Big Data that is historical data of institute is collected and stored centrally. As historical data is large so it is managed efficiently and accuracy is maintained. The user interface has provided to the user to work with the system. User can be admin, staff or data entry operator. User must have login to the system and then he/she can access the system. Here we have used HTML for front end designing and for querying, we are using PHP scripts. For data processing and storage, we are using Hadoop File System (HDFS).According to the user's credentials; they can fill the data and make changes to it. To manage these data Hadoop is used, so Map Reduce Technique is applied on it and data stored in HDFS system. Then as per the requirement user can generate specific reports. For report generation, data need to analyze, summarize and querying it. It is done by using Apache Hive, as queries of Hive are used to work on Hadoop that is queries are translated to MapReduce.

V. CONCLUSION

Information Storage and Management System (ISMS) is an efficient system for an institute. It overcomes drawback of paper work and speedup the data access. It provides central database for collecting information related to departmental activities. In future we will build this Information Storage and Management System (ISMS) for the university.

REFERENCES

[1]IEEE Explore (Vittal, A and Shivraj K.: Role of Information Technology and Knowledge Management in improving project management.)
[2] https://equizine.wordpress.com/2012/10/22/ advantage-ans-disadvantages-of-database-management-system/
[3] https://www.mapr.com
[4] https://hadoop.apache.org
[5] Bershad, B. et al, "The Scalable I/O Initiative",
[6] Development and Performance Improvement of Enterprise Information Management System (IEEE journal)
[7] http://www.michael-noll.com/tutorials/running-hadoop-on-ubuntu-linux-single-node-cluster/
[8] http://www.tutorialspoint.com/hadoop/hadoop_ enviornment_setup.htm
[9] https://www.youtube.com/watch?v=Mo KW5eY5yVY
[10] https://www.youtube.com/watch ?v=hJsaChh2Yhk