ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

Implementation of Hadoop Operations for Big Data Processing in Educational Institutions

B.Manjulatha[1], Ambica Venna[2], K.Soumya[3]

Assistant Professor, Dept. of CSE, VBIT, Telangana, India[1]

Assistant Professor, Dept. of IT, VBIT, Telangana, India[2]

Assistant Professor, Dept. of IT, VBIT, Telangana, India[3]

ABSTRACT: Education plays an important role in maintaining the economic growth of a country. The objective of this paper is to focus on the impact of cloud computing on educational institutions by using latest big data technology to provide quality education. Our educational systems have a large amount of data. Big Data is defined as massive sets of data that is so large or so complex that it is very difficult to process by using conventional applications and software technologies. This has resulted in the penetration of Big Data technologies and tools into education, to process the large amount of data involved. In this paper we discuss what Cloud and Hadoop is, and its types, operations and services offered. Hence it has an advantage which will surely help the students when used in an appropriate way.

KEYWORDS: Big Data, Learning Analytics, LMS, Educational Data Mining, Hadoop.

I. INTRODUCTION

In education, digital learning technologies such as games and online learning systems collect vast amounts of data as student's progress through the game, test, or activity. This type of incremental information can give a more complete picture of the learning process than traditional measures such as grades and test scores, which only measure outcomes. It can also help educators and researchers gain valuable insight into how to improve and personalize learning for students.

New academic disciplines such as learning analytics and educational data mining are emerging to make sense of this big data in education. The impact of this paper help the students learn, underscoring the need for collaboration between education researchers and those specializing in learning analytics.

Learning Analytics is based on 3 models:
1) Behavioral model is based on observation of student to assess the learning outcome.
2) Cognitive model is purely dependent on teacher.
3) Constructivist model is based on student to acquire the knowledge by their own, which are available for them to achieve a great success in their learning experience.

II. RELATED WORK

In the current learning environments, users like discussion forums, online chats, messages and various Learning Management Systems like OPENOLAT[1].As these learning environments have become accessible anywhere with the help of internet. Big data[2] plays a vital role in educational institutions because lots of information to be stored and retrieved. In addition to the data available from student activities, data are also created by educational institutions which use applications to manage courses, classes and students. The amount of data made available in the above scenarios is so enormous that traditional processing techniques cannot be used to process them. Due to the limitations of the conventional data processing applications, the educational institutions have started exploring "Big Data"

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

technologies to process the educational data.

What exactly is Big Data?

Big data is a term that describes the large volume of data – both structured and unstructured – that inundates a business on a day-to-day basis. But it's not the amount of data that's important. It's what organizations do with the data that matters. Big data can be analyzed for insights that lead to better decisions and strategic business moves.

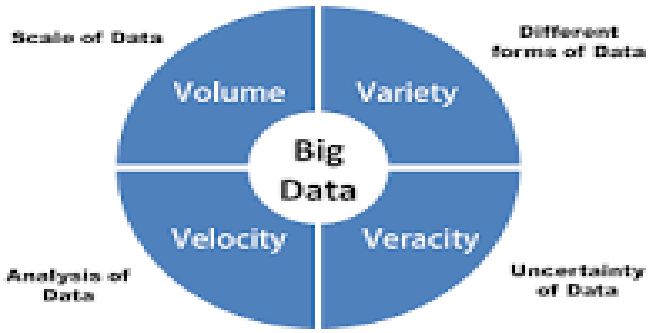
1.1 The characteristics of Big Data include:

Volume: Organizations collect data from a variety of sources, including business transactions, social media and information from sensor or machine-to-machine data. In the past, storing it would've been a problem – but new technologies (such as Hadoop) have eased the burden.

Velocity: Data streams in at an unprecedented speed and must be dealt with in a timely manner. RFID tags, sensors and smart metering are driving the need to deal with torrents of data in near-real time.

Variety: Data comes in all types of formats – from structured, numeric data in traditional databases to unstructured text documents, email, video, audio, stock ticker data and financial transactions.

Veracity: Big Data Veracity refers to the biases, noise and abnormality in data. Is the data that is being stored, and mined meaningful to the problem being analyzed. Inderpal feel veracity in data analysis is the biggest challenge when compares to things like volume and velocity. In scoping out your big data strategy you
need to have your team and partners work to help keep your data clean and processes to keep 'dirty data' from accumulating in your systems.



1.2 Challenges of Big Data

1. SCALE

With big data you want to be able to scale very rapidly and elastically. Whenever and wherever you want. Across multiple data centers and the cloud if need be. You can scale up to the heavens or shard till the cows come home with your father's relational database systems and never get there. And most NoSQL solutions like MongoDB or HBase have their own scaling limitations.

2. PERFORMANCE

In an online world where nanosecond delays can cost you sales, big data must move at extremely high velocities no matter how much you scale or what workloads your database must perform. The data handling hoops of RDBMS and most NoSQL solutions put a serious drag on performance.

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

3. CONTINUOUS AVAILABILITY

When you rely on big data to feed your essential, revenue-generating 24/7 business applications, even high availability is not high enough. Your data can never go down. A certain amount of downtime is built-in to RDBMS and other NoSQL systems.

4. DATA SECURITY

Big data carries some big risks when it contains credit card data, personal ID information and other sensitive assets. Most NoSQL big data platforms have few if any security mechanisms in place to safeguard your big data.

5. COST

Meeting even one of the challenges presented here with RDBMS or even most NoSQL solutions can cost a pretty penny. Doing big data the right way doesn't have to break the bank.

6. MANAGEABILITY

Staying ahead of big data using RDBMS technology is a costly, time-consuming and often futile endeavor. And most NoSQL solutions are plagued by operational complexity and arcane configurations.

7. WORKLOAD DIVERSITY

Big data comes in all shapes, colors and sizes. Rigid schemas have no place here; instead you need a more flexible design. You want your technology to fit your data, not the other way around. And you want to be able to do more with all of that data – perform transactions in real-time, run analytics just as fast and find anything you want in an instant from oceans of data, no matter what from that data may take.

The Big Data technologies overcome these challenges using various techniques.

III.PROPOSED SYSTEM

Open Source Tools:

Several Open source tools exist which help in taming Big Data Some of the top tools are listed below.

MongoDB : It is used because it enables data to build applications faster, handle highly diverse data types, and manage applications more efficiently at scale. MongoDB[3] is an open-source database developed by MongoDB, Inc. MongoDB stores data in JSON-like documents that can vary in structure. Related information is stored together for fast query access through the MongoDB query language. MongoDB uses dynamic schemas, meaning that you can create records without first defining the structure, such as the fields or the types of their values.

Hadoop : Apache Hadoop[4] was born out of a need to process an avalanche of big data. The web was generating more and more information on a daily basis, and it was becoming very difficult to index over one billion pages of content. In order to cope, Google invented a new style of data processing known as MapReduce. A year after Google published a white paper describing the MapReduce framework, Doug Cutting and Mike Cafarella, inspired by the white paper, created Hadoop to apply these concepts to an open-source software framework to support distribution for the Nutch search engine project. Given the original case, Hadoop was designed with a simple write-once storage infrastructure.

Hadoop has moved far beyond its beginnings in web indexing and is now used in many industries for a huge variety of tasks that all share the common theme of lots of variety, volume and velocity of data – both structured and unstructured. It is now widely used across industries, including finance, media and entertainment, government, healthcare, information services, retail, and other industries with Big Data requirements but the limitations of the original storage infrastructure remain.

R Programming: R[5] is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. R is rapidly becoming the leading language in data science and statistics.

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

One of the main attractions of using the environment is the ease with which users can write their own programs and custom functions. The R programming syntax is extremely easy to learn, even for users with no previous programming experience. Once the basic R programming control structures are understood, users can use the R language as a powerful environment to perform complex custom analyses of almost any type of data.

MapReduce[6] : It is easy to scale data processing over multiple computing nodes. Under the MapReduce model, the data processing primitives are called mappers and reducers. Decomposing a data processing application into mappers and reducers is sometimes nontrivial. But, once we write an application in the MapReduce form, scaling the application to run over hundreds, thousands, or even tens of thousands of machines in a cluster is merely a configuration change. This simple scalability is what has attracted many programmers to use the MapReduce model.

2.1 Proprietary Tools:

Splunk[7] : It is an American multinational corporation based in San Francisco,California, that produces software for searching, monitoring, and analyzingmachine-generated big data, via a web-style interface.
Splunk (the product) captures, indexes and correlates real-time data in a searchable repository from which it can generate graphs, reports, alerts, dashboards and visualizations.

ICCube[8]: Server is an in-memory multidimensional online analytical processing (OLAP) server written in Java.
It is typically used as a business intelligence tool to analyze and get insights from a wide range of types of data possibly spread out across multiple data-sources (e.g., RDBMS, Excel files, CSV files, MongoDB, etc...).

2.2. Applications in Learning:

Big Data techniques can be used in a variety of ways in learning analytics are:

1. Data Visualization
Reports on educational data become more and more complex as educational data grow in size. Data can be visualized using data visualization techniques to easily identify the trends and relations in the data just by looking on the visual reports.

2. Attrition Risk Detection
By analyzing the student's behavior, risk of students dropping out from courses can be detected and measures can be implemented in the beginning of the course to retain students.

3. Student skill estimation
Estimation of the skills acquired by the student

4. Behavior Detection
Detection of student behaviors in community based activities or games which help in developing a student model

5. Performance Prediction
Student's performance can be predicted by analyzing student's interaction in a learning environment with other students and teachers

6. Course Recommendation
New courses can be recommended to students based on the interests of the students identified by analyzing their activities. That will ensure that students are not misguided in choosing fields in which they may not have interest.

7. Intelligent feedback
Learning systems can provide intelligent and immediate feedback to students in response to their inputs which will improve student interaction and performance.

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

8. Constructing courseware

9. Grouping & collaboration of students

10. Social network analysis

11. Developing concept maps

12. Planning and scheduling

IV. SKILL ESTIMATION

Skill Estimation refers to the estimation of the skills of the students so that the learning environment can be adjusted to suit the student's skills. Skills were calculated based on the interaction of the student with the system or in the message boards or discussion forums.
Here is a summary of assessment methods described in Brown's, "Assessment: A Guide for Lecturers" (2001), a useful starting point to consider the variety of assessment possible:

Cases and open problems	An intensive analysis of a specific example.
Computer-based assessment	The use of computers to support assessments.
Essays	Written work in which students try out ideas and arguments supported by evidence.
Learning logs/ diaries	Wide variety of formats ranging from an unstructured account of each day to a structured form based on tasks.
Mini-practicals	A series off short practical examinations undertaken under timed conditions. Assessment of practical skills in an authentic setting.
Self-assessed questions based on open learning(distance learning materials and computer-based approaches)	Strictly speaking, a method of learning not of assessment. A process by which an assessment instrument is self-administered for the specific purpose of providing performance feedback, diagnosis and prescription recommendations rather than a pass/fail decision.

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

V. LITERATURE SURVEY

The paper examines various factors like technological, educational and political factors that drive learning analytics such as Big data, Online learning, political and economic concerns. Whereas educational data mining[9] focuses on how to extract useful data from a large learning dataset, learning analytics focuses on optimizing opportunities in online learning environment. Wolfgang Greller et. al. [10] propose a generic framework for Learning Analytics that considers six critical dimensions, namely, Objectives, Data, Instruments, Internal Limitations, External Constraints and Stakeholders. The paper also touches upon the ethical perspective of learning analytics to protect the learners. Erik Duval [11] discusses capturing of the attention data in learning environment in a number of ways such as posts, comments and messages. Data Infrastructure module makes use of Hadoop framework for distributed computation, distributed data storage and Data Broker service. Alyssa Friend Wise et. al. [12] investigates on how students contribute and reciprocate to International Journal of Computer Trends and Technology (IJCTT) – Volume 18 Number 6 – Dec 2014 ISSN: 2231-2803 http://www.ijcttjournal.org Page 261 messages in online discussions in learning environment. A valuable outcome of the findings was the invisible activity validation. Eg: ability to capture listening data such as people who were engaged intensely in discussions but did not post many comments and also the voracious speakers who had a need to improve on their listening efforts.

VI. CONCLUSION

The proposed paper describes the usage of learning analytics [13] is very limited to Higher education institutions in India. In many cases, Higher education institutions in India are not aware of the courses needed by the students. Knowledge from the data mining should be brought out to higher education institutions so that courses could be structured based on the need. The literature review shows that the various research activities are concerned mainly on students after joining into a particular course. This proves to be detrimental if the student has not selected a course properly. Education is the basic need for the developing countries like India.

To increase the number of students continuing higher education, the future research work is towards the design of a system for students to choose courses in the Indian universities using Learning Analytics. An efficient course advisory system can enhance the student performance. Such course advisory system minimizes the drop outs in higher education due to improper course selection. Various Big Data techniques become more and more necessary in learning environments to increase the quality and performance of the students.

REFERENCES

[1] http://www.openolat.com

[2] WAGMOB "BIG DATA AND HADOOP ", KINDLE EDITION.

[3] Karl Seguin "The Little MongoDB Book"

[4] Tom White "The Definitive Guide"

[5] Venables & Smith"A Beginner´s Guide to R by An introduction to R"

[6] Donald Miller and Adam Shook "Map Reduce Design Patterns"

[7] Josh Diakun,Paul R Johnson" Splunk Operational Intelligence Cookbook",Kindle Edition

[8] https://en.wikipedia.org/wiki/IcCube

[9] Alejandro Peña-Ayala, "Educational data mining: A survey and a data mining-based analysis of recent works", Expert systems with applications, Vol. 41, No. 4, pp. 1432-1462, 2014.

[10]Wolfgang Greller, Hendrik Drachsle., Translating Learning into Numbers: A Generic Framework for Learning Analytics, Educational Technology & Society, Volume 3, Issue 5, pp 42-57, ISSN: 1436-4522.

[11] Erik Duval, Attention please! Learning analytics for visualization and recommendation, 1st International Conference on Learning Analytics and Knowledge, 2011, pp 9-17, DOI: 10.1145/2090116.2090118.

[12] Alyssa Friend Wise, Yuting Zhao, Simone Nicole Hausknecht, Learning analytics for online discussions: a pedagogical model for intervention with embedded and extracted analytics, Third International Conference on Learning Analytics and Knowledge, 2013, pp 48-56, DOI: 10.1145/2460296.2460308.

[13] A survey and a data mining-based analysis of recent works", Expert systems with applications, Vol. 41, No. 4, pp. 1432-1462, 2014.

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

BIOGRAPHY

1.B.Manjulatha is an Assistant Professor in the Computer Science and Engineering Department, Vignana Bharathi Institute of Technology, JNTUH. I received Master of Technology (M.Tech) degree in 2012 from LIET, Ranga Reddy, India. My research interests are Network Security, Data Mining, and Internet of Things etc.

2.Ambica Venna is an Assistant Professor in the Information Technology Department, Vignana Bharathi Institute of Technology, JNTUH. I received Master of Technology (M.Tech) degree in 2013 from Tirumala Engineering College, Ranga Reddy, India. My research interests are Computer Network,Web Technologies and Internet of Things etc.

3.kothapalli Soumya is an Assistant Professor in the Information Technology Department, Vignana Bharathi Institute of Technology, JNTUH. I received Master of Technology (M.Tech) degree in 2011 from ATRI,Hyderbad, India. My research interests are Assistive Technologies and Internet of Things etc.