International Journal of Reflective Research in Social Sciences
ISSN: 2581-5733, Impact Factor: RJIF 2.38
www.reflectivejournals.com
Volume 1; Issue 1; Sep. 2017; Page No. 20-23

HADOOP: open-source data management technology

Faiqua Shaikh

Research Student, Marathwada College of education, Aurangabad, Maharashtra, India

Abstract

Hadoop is designed to scale up from single server to thousands of machines, each offering local computation and storage. The Hadoop framework application works in an environment that provides distributed *storage* and *computation* across clusters of computers**.** Hadoop is an open source, Java-based programming framework that supports the processing and storage of extremely large data sets in a distributed computing environment.

Keywords: Hadoop, Mapreduce, HDFS

1. Introduction

Big Data is a collection of large datasets that cannot be processed using traditional computing techniques. It is not a single technique or a tool, rather it involves many areas of business and technology. Big data involves the data produced by different devices and applications.

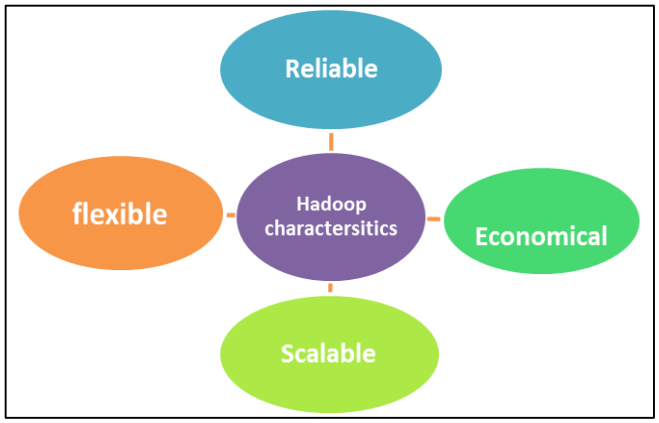
It consists:

- **Social Media Data:** Social media such as Facebook and Twitter hold information and the views posted by millions of people across the globe.
- **Stock Exchange Data:** The stock exchange data holds information about the 'buy' and 'sell' decisions made on a share of different companies made by the customers.
- **Search Engine Data:** Search engines retrieve lots of data from different databases.

Using the solution provided by Googleand his team developed an Open Source Project called HADOOP.

Hadoop runs applications using the Map Reduce algorithm, where the data is processed in parallel with others. In short, Hadoop is used to develop applications that could perform complete statistical analysis on huge amounts of data.

Definition: Hadoop is an open –Source data management technology with extensive Storage and distributed processing



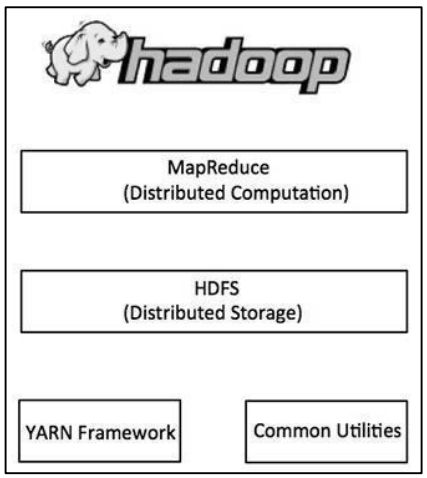
Hadoop Architecture

At its core, Hadoop has two major layers namely:

(a) Processing/Computation layer (MapReduce), and

(b) Storage layer (Hadoop Distributed File System).

Google solved the problem of big data management using an algorithm called MapReduce. This algorithm divides the task into small parts and assigns them to many computers, and collects the results from them which when integrated, form the result dataset.



Map Reduce
MapReduce is a parallel programming model for writing distributed applications devised at Google for efficient processing of large amounts of data (multi-terabyte data-sets), on large clusters (thousands of nodes) of commodity

Hadoop Distributed File System
The Hadoop Distributed File System (HDFS) is based on the Google File System (GFS) and provides a distributed file system that is designed to run on commodity hardware. It has many similarities with existing distributed file systems. However, the differences from other distributed file systems are significant. It is highly fault-tolerant and is designed to be deployed on low-cost hardware. It provides high throughput access to application data and is suitable for applications having large datasets.
Apart from the above-mentioned two core components, Hadoop framework also includes the following two modules:

- **Hadoop Common:** These are Java libraries and utilities required by other Hadoop modules.

- **Hadoop YARN:** This is a framework for job scheduling and cluster resource management.

Importance of Hadoop
Hadoop is changing the perception of handling Big Data especially the unstructured data. Let's know how Apache Hadoop software library, which is a framework, plays a vital role in handling Big Data. Apache Hadoop enables surplus data to be streamlined for any distributed processing system across clusters of computers using simple programming models. It truly is made to scale up from single servers to a large number of machines, each and every offering local computation, and storage space. Instead of depending on hardware to provide high-availability, the library itself is built to detect and handle breakdowns at the application layer, so providing an extremely available service along with a cluster of computers.

Computing power
Hadoop's distributed computing model processes big data fast. The more computing nodes you use, the more processing power you have. Unlike traditional relational databases, you don't have to preprocess data before storing it. You can store as much data as you want and decide how to use it later. That includes unstructured data like text, images and videos

How Does Hadoop Work? OR How Big Data Can be Manage by Hadoop?
It is quite expensive to build bigger servers with heavy configurations that handle large scale processing, but as an alternative, you can tie together many commodity computers with single-CPU, as a single functional distributed system and practically, the clustered machines can read the dataset in parallel and provide a much higher throughput. Moreover, it is cheaper than one high-end server. So this is the first motivational factor behind using Hadoop that it runs across clustered and low-cost machines.

Hadoop runs code across a cluster of computers. This process includes the following core tasks that Hadoop performs

- Data is initially divided into directories and files. Files are divided into uniform sized blocks of 128M and 64M (preferably 128M).
- These files are then distributed across various cluster nodes for further processing.
- HDFS, being on top of the local file system, supervises the processing.
- Blocks are replicated for handling hardware failure
- Checking that the code was executed successfully.
- Performing the sort that takes place between the map and reduce stages.
- Sending the sorted data to a certain computer.
- Writing the debugging logs for each job

Application of Big Data In Education

- Higher education Analytics
- Student Engagement
- Bookstore Effectiveness

1. Higher Education Analytics
Big data enables the maximization of students learning and success in higher education. For example a student is often faced with Questions like, what is the major for me? Similarly instructor wonder how should I customize learning paths so that no student is left behind?
The Tracking of student performance extracurricular interactions and social behavior result in the creation profile which is mapped with student profiles from the institution network to suggest the most relevant major. Teachers are also instant access to students performances based on which they can generate customized learning paths

2. Student Engagement
Data mining can help Universities to get holistic perspective about the student. Big Data Answers Questions like which learning behaviors are associated with better understanding and higher grades or what is the future course selection for this student this can also accommodate a student interest social and extra-curricular activities. This holistic perspective allows institutions to create immersive learning experiences for all students.

3. Bookstore Effectiveness

Big Data is used to improve bookstore profitability using analystic driven application like merchandising effectiveness and text book inventory optimization. Large retailers today find the combination of book bought together and sell them as a bundle making customer experience hassle-free. Additionally, they can also understand the topics in trend based on social media analytic to plan their purchases and inventories.

Current Example of Big Data in Education

Course Smart: Course smart embeds analytics directly into digital textbook. This analystic provide an "engagement Index Score "which measures how much students are interacting with their etextbook this ranges from tracking events such as Text-highlighting, Page views, time spent on each page etc. the resultant engagement score index carries out the accurate predictions of students learning path outcomes before moving ahead with the big data Hadoop use cases in education industry, let's see some of the common challenges with the education sector.

- Developing an industry ready education system
- Improving monitoring and evaluation
- Making system more accountable
- Getting trained teachers for improving quality education
- Knowing the industry demand and making the education system up to the level
- Improving the classroom environment and many others.

The education sector challenges and solutions with the help of big data and Hadoop

We will first list the problem faced and then will provide the possible solution using Big Data and Hadoop ecosystems.

1. Student Acquisition Optimization

Problem: We say that there is unemployment, but if you ask the human resource department of any education industry, they will say they are not getting the right candidates. You can easily find hundreds of candidates participating in a drive for just a few vacancies but at the end of the hardly any vacancies getting filled. Using Hadoop and its ecosystems, the industry can do better sentiment analysis on the students participating in the drive. The same can be done for those whom companies are calling for interviews. Data analysis can be done for the background, interest, capability, college/institution, etc. to know the students better.

2. Course selection

Problem: You will find the majority of students talking about the niche they are studying. Many of those don't like the department they study. This can frequently be experienced in engineering stream. This results in unemployment and depressions.

Possible solution: The best can be done here is, before selecting any branch for study, first do an analysis. What is the interest of the student, in which area he/she like to proceed further?

This can be done by analyzing their social media data like what they are posting, what they share etc. Another thing, take feedback from their teachers and see what kind of interest they showed in previous class, what questions they asked and then take the decision.

3. Teaching Effectiveness

Problem: There are many instructors who are either not giving their best or not able to. Identifying such instructors and improving the performance, rewarding the right teacher are the current challenge.

Possible Hadoop solution: Understanding if the lack of teaching effectiveness is a widely expressed sentiment about individual instructors can enable the institution to take corrective action faster. The institution can also reward those instructors who have widespread positive sentiment and successful students.

4. Increasing Job Oriented Education

Problem: If you see the first Hadoop use case in education I listed above, you will find companies are not getting the right candidates. The main reason behind this is, the current education system is not many jobs oriented. At many places still, that century-old syllabus is processing.

Possible Hadoop solution: Education system should be revised and should include the new courses and technologies. For example, if you talk about the engineering system, still they just teach C and C++. These are good and are the foundation but what about the industry.

So education courses should be prepared in such a way that it can compensate the demands raised by the industries.

Using Big Data, education institutes analyze the jobs being posted by the companies, analyze the trends in the industry and according to that train their students.

5. Student Retention

Problem: Most of the institutes are unable to retain their students for longer. Their students leave the institute/school/college early. This issue exists with schools, colleges or training institutes as well.

Possible Hadoop Solution: Whenever students join any institution, they talk about it. They talk about the likes, dislikes, pros and cons to their family, friends and even post on the social sites or review sites.

Institutes must analyze such posts and find the feedback, trends to serve their students better. Again this will be a part of sentiment data analysis.

Conclusion

One of the most amazing aspects in today's world is educational system using hadoop. Hadoop is being used in every sector now. Here we have discussed Hadoop use cases in education sector for improving the quality. It is efficient, and it automatic

distributes the data and work across the machines and in turn, utilizes the underlying parallelism of the CPU cores. The big advantage of Hadoop is that apart from being open source, it is compatible on all the platforms since it is Java based.

References

1. Retrieved from hadoop.apache.org/
2. Retrieved from https://en.wikipedia.org/wiki/Apache_Hadoop
3. Retrieved from https://hortonworks.com/apache/hadoop
4. Retrieved from www.sap.com/Hadoop
5. Retrieved from https://www.tutorialspoint.com/hadoop/