Hadoop: Solution to Unstructured Data Handling

Aman Madaan, Vishal Sharma, Prince Pahwa, Prasenjit Das and Chetan Sharma

Abstract Data is nothing but information of anything and we know it will continue to grow more and more. Unspecified format of data is unstructured data known as big data. 25% of the data that exist is in specified format, i.e., structured data and other 75% is in unspecified format. Unstructured data can be found anywhere. Generally, most of people and organizations pass out their lives working around unstructured data. In this paper, we have tried to work on how one can store unstructured data.

Keywords Big data · Hadoop · Analytics · Unstructured

1 Introduction

The one thing we know about data is that it will continue its pace. Terabytes are old, now we have petabytes, zetta bytes. Unstructured data can be defined as elements within data have no unique structure. The projects under big data offers the potential

A. Madaan (✉) · V. Sharma · P. Pahwa · P. Das · C. Sharma
Chitkara University, Baddi, Himachal Pradesh, India
e-mail: amanmadaan90@gmail.com

V. Sharma
e-mail: sharmavishal660@gmail.com

P. Pahwa
e-mail: princepahwa35554@gmail.com

P. Das
e-mail: prasenjit.das@chitkarauniversity.edu.in

C. Sharma
e-mail: chetan.sharma@chitkarauniversity.edu.in

V.B. Aggarwal et al. (eds.), *Big Data Analytics*, Advances in Intelligent
Systems and Computing 654, https://doi.org/10.1007/978-981-10-6620-7_6

to confront wide range of problems that appear when collecting and working with big data are Variety, Volume, Velocity. Volume (capacity), i.e., Gigabytes to Terabytes to Petabytes. Velocity, i.e., streaming. Variety is texts, images, sounds. Organizations with big data are over 60% more likely than other organizations to have Business Intelligence projects that are driven primarily by the business sector not by IT-sector [1]. According to study, nearly 70–80% of data in a company is unstructured figures which comprise of statistics from docs and different social media technologies. Although it seems difficult to analyze even structured data but Hadoop has made it easy to analyze, store and access unstructured data. According to a study, world is now producing data eight times faster. Analytics-The use of data and related insights developed through applied analytics disciplines. Analytics are of three types: predictive, descriptive or prescriptive. Big firms and business tycoons believe that analytics connected with BD (big data), is going to play a major role in the economy in next 10–20 years. Some suggest today's big data will be the data of future.

Big data is very important nowadays in every sector whether it is scientific, industrial, public or even education sector.

1.1 Few Amazing Statistics

Every minute 121,000 tweets, 671 new websites are created, 3.5 quintillion bytes of data is produced in unstructured format every hour from different data origins like social networking sensors. To handle big data could be easy but it was difficult to handle unstructured large amount of data until Hadoop was developed.

Hadoop: This project was laid by Google and yahoo both. A prime role played by Yahoo for the growth of Hadoop for companies applications.

The success of "Google" is because it always focused on large amount of data, i.e., Big data analytics and social networking sites like Facebook, Twitter are best and successful example of application of big data (as daily there are millions of photos and statuses uploaded daily), so to store that one really needs big data analytics. Hadoop is for situations like clustering and targeting.

1.2 Components that Make up Hadoop

HDFS stands for Hadoop Distributed File System which is expandable as well as genuine storage system which stores each fork in a Hadoop cluster into a unbounded file system [2]. It helps in storing the file in form of big chunks which allows it to store large files on numerous machines efficiently. Chunks of data can be accessed parallely, without scanning complete file into a single computer's memory. By making replica of data on numerous hosts accuracy is achieved; by default, each chunk of data is stored, on three different PCs. If any fork or node fails, the data can be accessed from additional copy of blocks. This approach allows HDFS to dependably store massive amounts of data. For instance, in late 2012, the Apache Hadoop clusters at Yahoo had grown to hold over 350 petabytes (PB) of data across 40,000+ servers. Once data has been loaded into HDFS, we can begin to process it with MapReduce.

1.3 MapReduce

The programming dummy or model by which Hadoop can process great amount of data efficiently is MapReduce and it helps Hadoop to break large data processing troubles into numerous steps, a collection of Maps and Reduce which worked on multiple computers in parallel at the same time.

The purpose of MapReduce is to work with Hadoop. Apache Hadoop automatically optimizes the execution of MapReduce programs so that a given Map or Reduce step runs on HDFS node that contains the blocks of data which is necessary to complete the process. Data processing troubles that once required plenty of time to complete on large and costly computers now can be programmed that can be completed in seconds on machines that are not expensive.

1.4 YARN

As previously noted, Hadoop was initially adopted by many of the large web properties and was designed to meet their needs for large web scale batch type processing. The number of ways to store data in Hadoop expanded clusters and adoption expanded of Hadoop community has broader ecosystem, popular examples being Apache Hive is for querying and Apache Pig for processing of scripted data and Apache HBase for database.

Because of these open source projects, it created door for much richer and wider set of applications and are built on top of Hadoop—but these projects did not address the design limitations inherent in Hadoop; mainly batch-oriented data processing.

YARN is a key piece of Hadoop version 2 which is currently under development in the community. It provides a resource management framework for any processing engine, including MapReduce. It allows new processing frameworks to use the power of the distributed scale of Hadoop. It transforms single application to multi-application data system. This is future of Hadoop.

2 Survey and Motivation

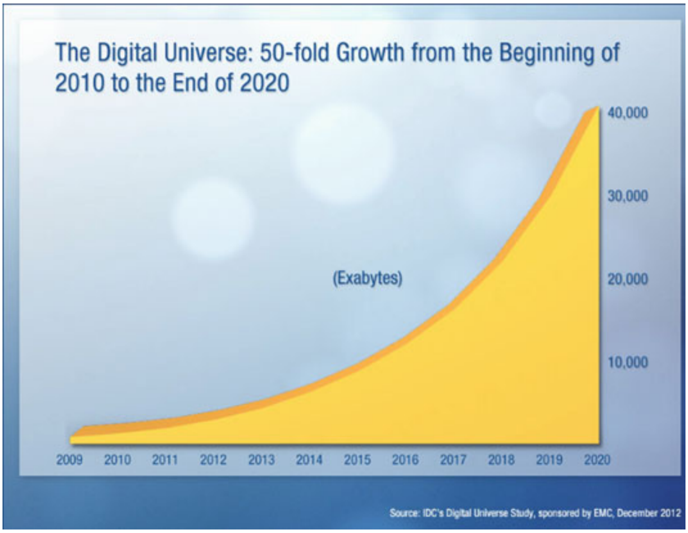
Figure 1 explains that there will be huge increase in data by 2020 and so will be unstructured data. This graph is rising exponentially. According to this rise in graph [3]. Universe will grow from 50-fold up to 2020 and this is need of hour to study how to manage unstructured big data now which will help in future.

Figure 2 depicts that how much amount of unstructured data the world is going to use by 2020 is low as compared to structured data.

BD is defined as unstructured and semi-structured data that can be from all kinds of places, different sources and formats, suitable example would be web content, posts from twitter, contents from Facebook. Insights from these contents were not possible few years back but now it is easy to analyze because of advancement in technology. Decoding the human genome was merely impossible but now it can be achieved in one week only.

3 Architecture of Hadoop

It runs on plenty of machines that do not work on common memory or any storage devices and you can invest on bunch of commodity servers, store them in a structure, and successfully execute the Hadoop software on every one [4]. How you can load Hadoop by your organization's data? Applications break that data into small data



Source: IDC's Digital Universe Study, sponsored by EMC, December 2012

Fig. 1 Digital universe in 2020. 40,000 exabyte's by 2019–2020. 1 exabyte consist of 10^{18} bytes which is similar to 1,000,000 TB

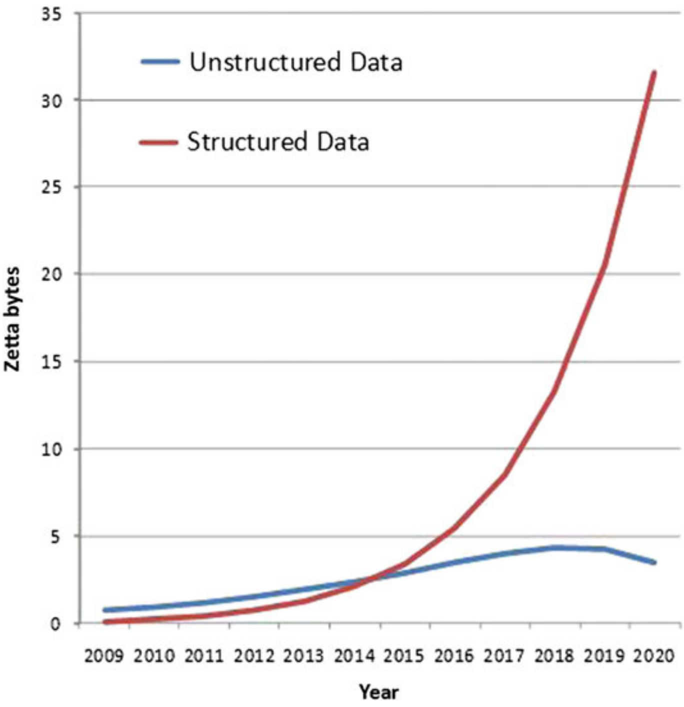


Fig. 2 Unstructured data v/s structured data

and then these data pieces spread across different servers. There is not a single place where you can eye all your data; Hadoop helps to track where your data kept. And as there are numerous copies to single data are stored, data stored on a server can be automatically replaced from a known good copy when it goes offline or dies.

In a unified database system, there are four, eight or big processors are connected to big disks. In a Hadoop cluster, every one of those servers has two or four or eight CPU's. Hadoop breaks the data and stores them in the Hadoop Distributed File System (HDFS), that can go to number of forks on a single cluster and in a single instance can support tens of millions of files. It is then set to be analyzed by the MapReduce framework (Figs. 3 and 4).

4 Proposed Solution

Our only problem is to handle unstructured data and it can be solved by using Hadoop technology. Hadoop acts as a Data warehouse. This helps in getting quality into the data through advanced analytics. It stores data in its native form until we need it. Business analytics and data mining tools are used to retrieve the data required.

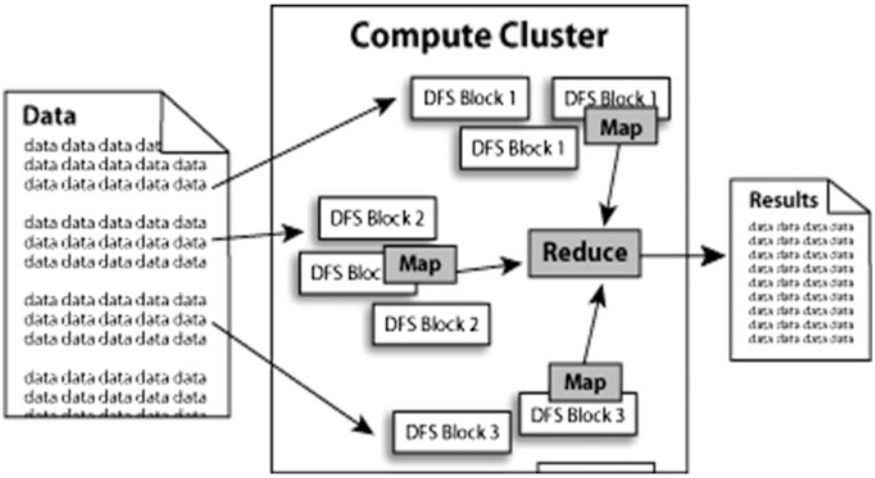


Fig. 3 Hadoop architecture

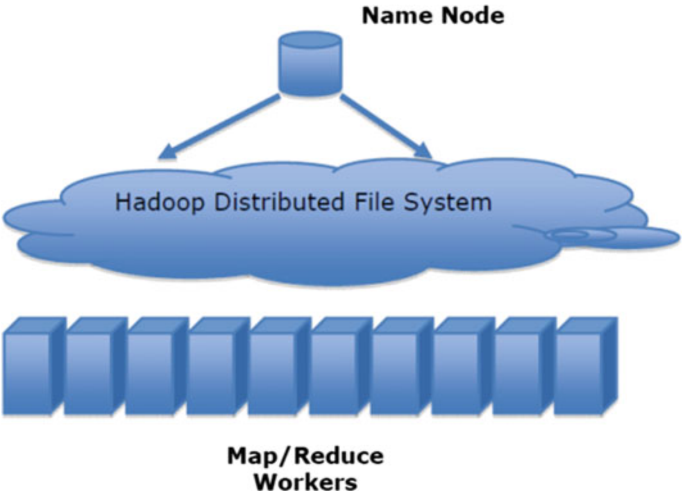


Fig. 4 MapReduce

4.1 Benefits to Use Hadoop

1. Hadoop is open source software so it is very economical to use it.
2. MapReduce is the framework in the Hadoop stack for software programming [5].

3. Existing database and analytics infrastructures are supported by Hadoop. Big data is a big opportunity to those who provides solutions. As an example, $740 million dollars has been invested by Intel into the rising distribution for Hadoop.
4. Quickly processing big data is done through the distributed computing model. As more computing nodes we use, more the processing power we have.
5. Hardware failure does not affect data and application processing.

4.2 Challenges

The most important problem is to collect, combine and analyze the data. Obtaining consent from users is very hard to manage including machine generated sources.

Data integration is time-consuming

In today's world data scientists spend large amount of their time on collecting and moving large amount of data, before it can be explored for useful nuggets. Big data is heterogeneous by nature; as analysts can ask new, different and more instinctive questions of the data, with a focus on mind that what is more profitable.

Real-time analytics is difficult to achieve

Hadoop has been in major use for the exploration of data. Although Hadoop can process large amount of data quickly, it has restrictions. Emerging area for Hadoop is real-time analytics, transferring large amount of data quickly has been a big challenge.

5 Conclusion and Future Work

In this paper, we tried to cover every glimpse of big data, i.e., how unstructured data can be stored and accessed. Hadoop can handle all types of data: structured and unstructured (mixture of data). With no prior need for a schema various kinds of data can be stored in Hadoop. In other words, no need to know how data will be stored. Hadoop reveals hidden relationships and answers many problems which have been hidden from a long time. We can start making decisions based on actual data.

In future we will work on how big data can be used in different fields whether it is educational sector, geosciences, and bio-energy. We will try to use big data analytics in these fields so that these sectors can be uplifted. With the advancement in technology we can perform better in these fields also by knowing some analytics or relationships which were not possible few years ago but now it is need of hour to work on these fields with these technologies.

References

1. http://www.datamation.com/applications/big-data-9-steps-to-extract-insight-from-unstructured-data.html
2. http://www.sas.com/en_us/insights/big-data/hadoop.html
3. http://blog.sqlauthority.com/2013/10/02/big-data-what-is-big-data-3-vs-of-big-data-volume-velocity-and-variety-day-2-of-21/
4. https://beta.oreilly.com/ideas/what-is-hadoop
5. http://www.searchcio.techtarget.com/intelsponsorednews/Five-Things-to-Know-About-Hadoop-for-Managing-Unstructured-Data
6. http://www.digitalreasoning.com/resources/Holistic-Analytics.pdf
7. http://www.dummies.com/how-to/computers-software/Big-Data/Data-Management/Hadoop.html
8. http://www.theglobaljournals.com/gra/file.php?val=February_2013_1360851170_47080_37.pdf
9. http://www.hadoop.apache.org/core/docs/current/api/