ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 4, Issue 9, September 2017

A Review Paper on the Most Trending Technology: “ Big Data & it’s Processing using Hadoop ”

[1] Harshini Tanuku, [2] Keerthi Bangari
[1] Undergraduate, [2] Assistant Professor
[1][2] Computer Science and Engineering

Abstract - **Big data is defined as a collection of big and complex data sets that have gigantic amount of data, that encompass social media data analytics, data management efficiency and real-time data. Big data analytics is termed as the process in which huge amounts of data are studied for further processing and industrial usage. Big data is processed using a technology known as “Hadoop” that uses the MapReduce paradigm for processing the huge datasets.**

Key Words— **Big data, Hadoop, HDFS, MapReduce Paradigm, Crowd sourcing, “ The 4 Vs”, Operational big data, Analytical big data.**

I. INTRODUCTION

Big data is a collection of large datasets encompassing of huge amounts of data that cannot be processed using traditional computational techniques. It is a term that is used to define large volumes of data that is both “structured” and “unstructured”.“Big data analytics” is defined as the process of analyzing big data. Big data is analyzed so that better business decisions and strategic business moves are made. Despite there being many emerging technologies to handle big data Hadoop is the most preferred tool to process big data. Hadoop uses a paradigm known as the “MapReduce” technique to process the complex and unstructured datasets. Big data ain’t a novice phenomenon, but one that is a part of an ancient method of capturing and storing data from archaic information. Just like the other developments and enhancements in data storage, data processing, the web and the internet, big data is a further step that marks a milestone by setting up new trends in the way we capture heterogeneous data, store it and process it for an efficient running of various business organizations.

It will also act as a bedrock to the many new technologies that are going to take birth and revolutionize in the coming generations.

II. The “4 Vs” of Big data

The tradition of storing large amounts of data for analysis is ages old though the term “big data” is a new one.

The concept of big data gained acceleration during the “Y2K” or the early 2000’s when industry analyst Doug Laney defined the so called “Big data” of this era as the “4 Vs”. The “4 Vs” are namely:

• ***Volume:***

Organizations collect huge amounts of data from various sources like social media, business transactions and information from machine-to-machine data. Storing such huge amounts of data caused troubles in the past, but with the emergence of new technologies (like Hadoop) made it less cumbersome.

• ***Velocity:***

Data streams with an unforeseen speed must be dealt with in a periodic manner. Sensors, RFID tags and smart meetings are driving the need to handle torrents of data in real-time.

• ***Variety:***

Data comes in all types and formats. It can be structured or unstructured. It can also be numerical data for daily base transactions from traditional databases to unstructured text documents, pictures, emails, audio and video files, financial transactions etc.

• ***Veracity:***

Veracity in contrast to its actual meaning (which means germane/valid) is the biases, noises and the abnormalities stored in the data. Veracity questions a very legitimate point that “ Is the data being stored and mined meaningful to the problem being analyzed?

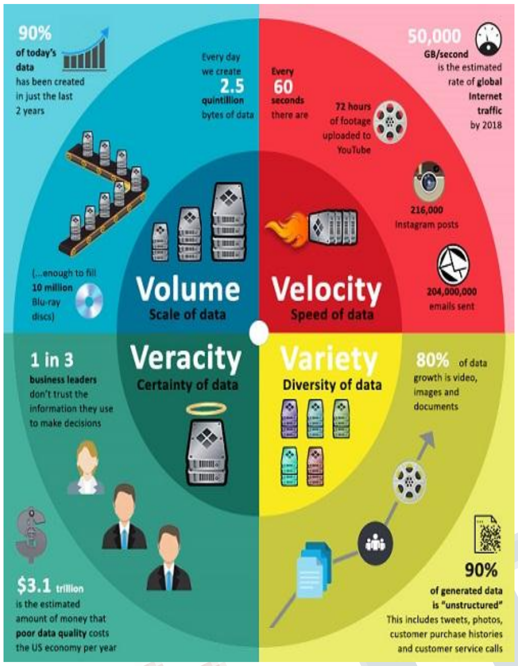
90%
of today's data has been created in just the last 2 years
Every day we create 2.5 quintillion bytes of data
(...enough to fill 10 million Blu-ray discs)
Volume
Scale of data
Every 60 seconds there are
72 hours of footage uploaded to YouTube
216,000 Instagram posts
204,000,000 emails sent
50,000 GB/second is the estimated rate of global Internet traffic by 2018
Velocity
Speed of data
1 in 3
business leaders don't trust the information they use to make decisions
$3.1 trillion
is the estimated amount of money that poor data quality costs the US economy per year
Veracity
Certainty of data
Variety
Diversity of data
80% of data growth is video, images and documents
90%
of generated data is "unstructured"
This includes tweets, photos, customer purchase histories and customer service calls

Figure1: The "4Vs" of Big data.

III. Challenges clouding Big data

A. Heterogeneity of the data

Data required for analytics and query processing must be in a structured format, but the data stored in datasets is usually unstructured. This is called as heterogeneity of data which must be handled by analysts before sending such data for query processing as, data in this format will neither generate proper result sets nor give the apt results for the queries.

B. Data Privacy

One of the most paramount issues looming around big data is its privacy. Information privacy deals with the control over how the personal and private information or data is used. Information privacy is the capacity of an organization or an individual to stop the spreading of confidential information to people other than whom they have given the rights to know about it.

Any kind of security breach t private data can turn out to be a peril to the organization, as they can be malicious usage of this private data by people with a negative intent.

C. Scale of big data

As the name goes, big data consists of many datasets of huge sizes having complex and unstructured data. Handling this type of data has been a problem since decades. But this anomaly was later resolved by the advent of fast processors. This was germane until an extent. But as the data quantities were getting larger day by day, the fast processors became incompatible to process the data. The world started to adopt cloud technologies and it was due to this migration that the data got generated at a very high rate. This high amount of data is causing a major concern to data analysts. Though hard disks can be used to store huge amounts of data, they have a slower I/O performance. Though hard disks are being replaced by solid state devices newer systems must be constructed to handle such voluminous data.

D. Crowd sourcing and Human collaborations of big data

Though there are numerous high end computational models, there are still some patterns which the computer can't recognize. Hence crowd souring is a technique to address such problems. Crowd sourcing is a time and cost effective method for moderating and curating data. It abstains from overhead costs and produces high quality results with little investment. Five benefits that are obtained from crowd sourcing are:

- To obtain real time analytics.
- To save the internal resources.
- To take advantage of scale.
- To save time.
- To capitalize on the human elements.

IV. Tools to handle big data

Big data comprises of large datasets which must be processed using certain technologies. These tools are introduced for manipulating, analyzing and visualizing big data.

Though there are numerous techniques to process big data, "Hadoop" is the most preferred.

Google has given a solution to handle big data using an algorithm known as "MapReduce".

MapReduce algorithm divides a task into small parts and assigns them to many computers and collects the results from them. These results when integrated form the result dataset.

A. Hadoop

Using the solution provided by Google, Doug Cutting and his team developed an open source project called “HADOOP”. Hadoop runs applications using the MapReduce algorithm, where the parallel processing of data is done. In short Hadoop is used to develop applications that could perform a complete statistical analysis on huge amounts of data. Hadoop is an open source Apache framework written in java that facilitates the distributed processing of large datasets across clusters of computers using simple programming models.

B. Hadoop Architecture

Digging to its crust, Hadoop has two paramount layers namely:

•Processing/Computation layer [MapReduce]:

MapReduce is a parallel programming model distributed applications devised at Google for efficient processing of large amounts of data (multi tera byte datasets) on large clusters of commodity hardware in a fault-tolerant, accurate and trust worthy manner.

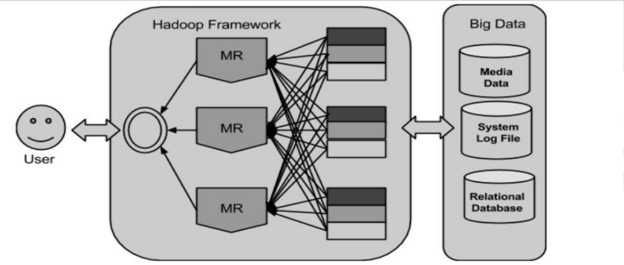

Figure 2: MapReduce Architecture

C. Hadoop Distributed File System [HDFS]

The HDFS collaborates to the Google File system (GFS) and provides a distributed file system that is designed to run on commodity hardware. Some of its features are:

- Highly fault-tolerant.
- Designed to deploy on low cost hardware.
- High throughput access to application data and is suitable for applications having datasets.

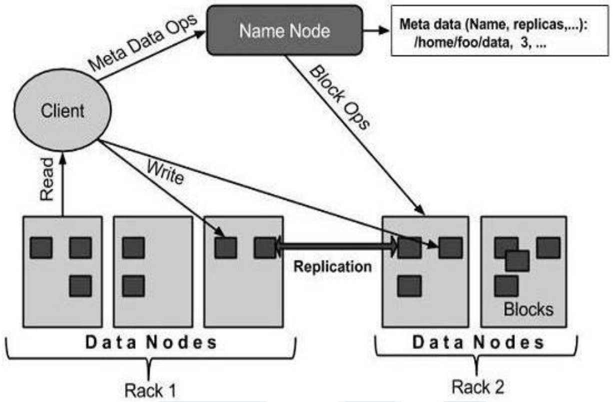

Figure 3:HDFS Architecture

V. How does Hadoop work

Hadoop runs the code across a cluster of computers. The process includes the following:

- Data is initially divided into directories and file. These files are now divided into uniform sized blocks of 128M and 64M.
- These files are distributed across numerous cluster nodes for further processing.
- HDFS, being at the top of the local file system, monitors the processing.
- Blocks are replicated for handling hardware failure.
- Checking that the node was successfully executed.
- 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.

VI. Conclusion

This paper describes an era of big data along with the “4Vs”of it namely Volume,Velocity,Variety and Veracity. This paper also focuses on the challenges of big data and tools like Hadoop intended to process large amounts of data by using map reduce algorithm.

REFERENCES

[1] Karun, A. Kala, and K. Chitharanjan. "A review on hadoop—HDFS infrastructure extensions. "Information & Communication Technologies(ICT), 2013 IEEE Conference on. IEEE, 2013.

[2] Davenport, Thomas H., and Jill Dyché. "Big data in big companies." International Institute for Analytics (2013)

[3]Mayer-Schönberger, Viktor, and Kenneth Cukier. Learning with big data: The future of education. Houghton Mifflin Harcourt, 2014.

[4] Provost, Foster, and Tom Fawcett. Data Science for Business: What you need to know about data mining and data-analytic thinking. " O'Reilly Media, Inc.", 2013.

[5] Bhosale, Harshawardhan S., and Devendra P. Gadekar. "A Review Paper on Big Data and Hadoop." International Journal of Scientific and Research Publications 4.10 (2014):