International Journal of Technical Research and Applications e-ISSN: 2320-8163,
www.ijtra.com Volume 3, Issue 3 (May-June 2015), PP. 114-116

ANALYTICS OF DATA USING HADOOP-A REVIEW

Vineet Sajwan, Vikash Yadav
MGM's College of Engineering and Technology,
Noida Sector-62
Vin94sajwan@gmail.com, 1109510067@coet.in

ABSTRACT- In this paper, we discuss about the Big Data. We analyze and reveals the benefits of Big Data. We analyze the big data challenges and how Hadoop gives solution to it. This research paper gives the comparison between relational databases and Hadoop. This research paper also gives reason of why Big Data and Hadoop.

General Terms
Data Explosion, Big Data, Big Data Analytics, Hadoop, Hadoop Distributed File System, MapReduce
Keywords
Big Data, Big Data Analytics, Hadoop, Hadoop Distributed File System, Map Reduce.

I. INTRODUCTION

We live in the age of data .It is not easy to measure the volume of data. According to the IDC estimate the size of "Digital World" is about 0.16 zettabytes and forecasting the "Digital World" will grow by a factor of 300, from 130 Exabyte to 40,000 Exabyte. This flood of the data is coming from various sources around the world. Society is being surrounded by gadgets and as a results organization have to store huge amount of data .To manage and process that data is big challenge for the organization. The data from the social websites makes a big opportunities for the organization for understand the customer needs. This paradigm is known as Big Data. Big Data, two popular terms are here today for good reason. There is a big market in big future for Big Data. Company are starting to realize there untapped information and unstructured documents siting across the network. Analytics is the solution that mines unstructured and semi structured data and gives insights to the organization from not their privately acquired data but also from large amount of data is publicly available from various sources. We have emails, imagine we can data mine our emails and get meaningful results. The kind of information we get from our emails, documents, textiles, spreadsheet, and pdf. All these unstructured data siting across the network that contains answer which help us to create new products, refining existing product, discover trends, improve customer relation, and understand ourselves and help us to understand even our company better. Hadoop answer many big data challenges that we see today that is how to store terabytes and petabytes of information ,how to access the information quickly , how to work with variety of data and how to do you do all scalable fault tolerant and flexible way. There is a saying "More data usually beats better algorithms." It means for some problem may be your algorithm solve that problem but it can be often be beaten by adding more data simply.

The good news is big data is here and bad news is that we are struggling for storing and processing it.

II. DATA EXPLOSION

If we sum up the state of data in one word it is a lot. There's a lot of data in our world.90% of worlds data is created in last two years that's say big data is here .Through smart phones and tablets one can access data everywhere. Social media is everywhere Facebook, twitter, Instagram, tumbler the list goes on and on generating data with alarming rate which is known as data explosion. Data explosion is the explosion of unstructured and semi-structured data. Since the end of 90's it's been on terror .It is shows exponential growth with no sign of slowing down. Structured data are pretty manageable and predictable curve.Exmaples of the types of the data are:-

- Unstructured Data: - Things like email, pdfs, documents on the hard drive, textiles.
- Structured Data:-Everything has schema associated with it and checked against the schema when we put it into the database.
- Semi-Structured:- Things that have some form of hierarchy or associated with table but nonetheless contain tags. Xml is a tag based format that describe data inside the tag in its hierarchical so got some structure. They don't have scehma associated with it.

III. BIG COMPANIES WITH BIG DATA

A. Google

In starting Google have to make internet searchable for they need to index a billion pages .So they built the technology called mapReduce along with GFS(Google File System) that's what Hadoop is based on .Doug Cutting in 2003 join Yahoo and yahoo gave him the team of engineers. They built Hadoop based on those white papers the GFS and mapReduce. That's the core technology of Hadoop i.e. HDFS (Hadoop distributed File System) and mapReduce.

Google index billion pages using the same technology. Now today 60 Billion pages google indexes to make searchable for the internet in very less seconds.

B. Facebook

Facebook has the largest Hadoop cluster on this planet. They have the cluster that contain of the 100 petabytes of data and they have generated half petabyte of data in one day. Anything everybody does on Facebook from login to clicking to liking and sharing something is tracked on Facebook and thats how they use the data to target views. They are looking what we are clicking what our friends are clicking and find the commonalities them something called collaborative filtering which they use for targeting.

C. Amazon

In Amazon, recommendation system is based on some elements like what a user had bought in the past, items they have in their virtual shopping and item they have rated and liked and what other customer had viewed or purchased. Behind the recommendation system amazon's collaborative filtering algorithm works. This is the reason of Amazon's success.

D. Twitter

400 million a day which equals to 15 terabytes of data every day. They have thousands of Hadoop clusters setup they run thousands of mapReduce jobs every night to discover trends and they sell those trends to the people who makes product so can they target those trends

E. General Motors

General Motors are car manufacture of America .In 2012, they cut off multibillion dollar contract with HP for outsourcing .They are building 220000 square feet warehouses in which Hadoop were installed.

(i) GM combines the Global Information System (GIS) and data analytics to improve the performance of their dealer. These analytics shared are shared with the dealers who now can view local demographics, location characteristics and customer satisfaction.
(ii) With the marketing budget of $2 billion per year GM have lots of potential customer. But instead of mass targeting the public they use big data analytics to create the customer profile.GM knows who the buyers are that want luxury cars and where they are.
(iii) GM focuses on sensors and telematics. For GM Big Data saves 800$ per car in GM telematics .Which means saving lie in the connected cars that can communicate with manufacturer via 4G.

IV. 5 V'S OF BIG DATA

It is the 5 big reason we don't use traditional computing model which is big expensive trickled down machines with lots of process, cords, hard drives read enable to the processes and fault tolerance. It is the hardware solution and we don't use hardware solution and also we can't use the relational world. Relational world is use to design to handle gigabyte of data not petabytes of data. The big data breaks into 5 dimensions that is volume, velocity, variety, value and veracity.

A. Volume

Volume refers to the data generated every second. We are not talking about terabytes but Zettabytes. If we take all the data generated in the world between beginning time and 2008 the same amount of data generated every minute. This makes most data sets too large to store and analyses using traditional database technology. Hadoop uses use distributed system so we can store and analyses data across the database around anywhere in the world.

B. Velocity

Velocity is the speed which we access the data. In traditional computing models no matter how fast our computer is our process is still bound by disk I/O because our disk transfer is having involved in newly pace of processing power. So that's why distributed computing is more useful and Hadoop increases strength of current computing world. It brings computation to the data. It doesn't brings data to the computation. It can process the data locally and when we have cluster of nodes to all working together using and harnessing the power of processor and reducing network bandwidth and navigating the weakness of disk transfer. Yahoo break the record of processing a terabyte of data many times using Hadoop.

C. Variety

The relational model can only handle structure data they can get semi structured and unstructured data. Some engineers they have ETL tool to transform and scrub the data to bring it but that require a lot of extra work .With the help of Hadoop we can now analyses and bring together data of different type such as messages and social media conversations, photos, sensor data, video and voice recording.

D. Value

Having access to the Big data is no good unless we can turn it into the value.
Companies are starting to generate amazing value from their big data.

E. Veracity

Veracity refers to the uncertainty surrounding of data which is due to the inconsistence and incompleteness of data. It is the major challenge to keeping organized data.

V. HADOOP

Hadoop is distributed software solution. It is scalable, fault tolerant distributed system for data storage and processing. There is two component of Hadoop that is HDFS which is the storage and mapReduce which is the retrieval and organising.

- HDFS is self-healing high bandwidth cluster storage. If we put petabyte file in Hadoop cluster. Hadoop would break it up in the blocks and distributed across all the nodes in a cluster where its fault tolerant function come into play. When we setup HDFS we setup replication factor (by default it sets 3).When we put a file in HDFS it's going to make sure there are three copies of every block that make up that file spread across the node in a cluster. If we lose a node it going to self heal because it knows where is data on that node It re-replicate the blocks ever on that node to the rest of the subversion sites of the cluster.
  It has the name node and the data node. Generally one node is name node in a cluster and rest of them are data node. Name node is the meta data server it just hold N memory and knows where are blocks exist on what node on what rack spread across the cluster inside the network.
  That's how Hadoop is fault tolerant.
- mapReduce is the two-step process at the service. Programmer writes the mapper function which will goes out and tell the cluster what data we want to retrieve. The reducer takes all the data and aggregate.

Hadoop is batch processing based system. We working on all of the data. mapReduce is all about working on all of the data inside cluster. Hadoop is flexible because there is no need to understand javato get data of cluster. In Facebook, engineers built Hive because Facebook doesn't to force anybody to learn java for data extraction. Now anybody is familiar with sequel which most of professional are can pull data out of the cluster.
Hadoop is scalable because we keep adding more data into the cluster by add more node which increases the overall performance of cluster.

VI. PROBLEM DEFINITION

As there exist large amount of data, the various challenges are faced about the management of such extensive data like unstructured data, fault tolerance and issues regarding storage of big amount of data.

A. Problem Description

The problem is simple: the storage capacities of hard drivers increased massively over the years-the rate at which data can read from drives have not kept up. In 1990, one typical drive can store 1,370 MB of data and had transfer speed of 4.4 MB/s, so you could read the full data in about 5 minutes. 20 years later, one typical hard drives can store terabytes but speed is around 4.4MB/s so it need half hour to read full data of the disk.

B. Problem Solution

The obvious way to reduce time is multiple disks at once. Imagine if we have 100 drives working in parallel we can read the data in 2 minutes. Only 100 disks is wasteful .But we can store datasets, each of which is one terabyte and provide access to them.

C. Problem to Solve

The first problem to solve is hardware failure:-The chance of failure increases as we connected the drivers. A common way of avoiding data loss is through replication. The Hadoop Distributed Filesystem (HDFS).

The second problem to solve is analysis task need to combine in some way:-The data read to one disk need to be able to combine from any of 99 disks. Various distribution system allow this but doing this is very challenging. MapReduce provides a programming model that abstracts the model from disk read and writes, transforming it into computation over sets of keys and values.

VII. HADOOP VS RDBMS

Hadoop is intelligent as we seen this in the computation of data maximizing the strength of todays computing world and navigating the weaknesses. In a multi rack environment there is a switch which is rack aware. It knows what node belongs to which rack. In a more data locality whenever receive mapReduce job it find the shortest path to the data as possible.

	Traditional RDBMS tools	**mapReduce**
Data Size	In Gigabytes	In petabytes
Access	Batch and interactive	Similar
Update	Read and Write many times	Write once and read many times
Structure	Fixed schema	Unstructured schema
Latency	Low	High
Integrity	High	Low
Language	SQL	Procedural Language (C++,Java)

VIII. CONCLUSION

We live in the data world. There are various challenges in big data .there are various issue too. We came into conclusion that big data is not just huge data it is an opportunity to find new insights and analysis of our future data. We have discuss various parameters of big data. We also have discuss about comparison between RDBMS and MapReduce. There must be support and research on big data for getting new results.

MapReduce is good for analyze huge dataset in a batch function where RDBMS is good for queries and update.

REFERENCES

[1] Shipa, Manjit kaur, "*Big Data and Methodology",* 10 Oct, 2013
[2] Pareedpa**,** A.; Dr. Antony Selvadoss, "*Significant Trends of Big Data*", 8 Aug, 2013
[3] Gurpeet Singh Bedi, Ashima, "*Big Data Analysis with Dataset Scaling in Yet another Resource Negotiator (YARN)",* 5April, 2013
[4] *Hadoop-The Definitive Guide*, Tom White, Edition-3, 27Jan, 2012
[5] Mrigank Mridul, Akashdeep Khajuria,Snehasish Dutta,Kumar N, "*Analysis of Big data using Apache Hadoop and MapReduce"*,Volume 4, May 2014
[6] IBM 2012, What is big data: Bring big data to the enterprise,http://www.01.ibm.com/software/data/bigdata, IBM
[7] Sam Madden, *"From Databases to Big Data",*IEEE computer society,2012
[8] "*Data Mining with BigData"* ,Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Ding , 1041-4347/13/$31.00 © 2013 IEEE
[9] Russom, *"Big Data Analytics"* , TDWI Research,2011
[10] An Oracle White Paper*, "Hadoop and NoSQL Technologies and the Oracle DataBase"*,February 2012