HADOOP the Ultimate Solution for BIG DATA Problems

Papineni Rajesh, Y. Madhavi Latha

B.Tech Final Year, Electronics and Computers Department, KLUniversity
Vaddeswaram, Andhra Pradesh, India

Assistant Professor, Electronics and Computers Department, KLUniversity
Vaddeswaram, Andhra Pradesh, India

***Abstract*— The demand for analysing the data has augmented drastically. Big data is large amount of information present in the digital world around us. The data present in various formats is made accessible and when managing and storing this data and retrieval of data ought to be dynamic and prompt. Hadoop is the technology used for secure, huge amount of data.**

This paper provides an overview of Big Data Analysis and Hadoop technologies used.

***Keywords*— Big Data, Hadoop, Analysis, Security, Visualization.**

I. INTRODUCTION

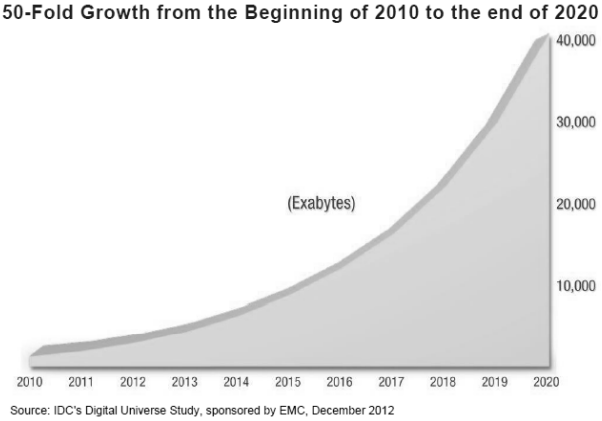
The World of Internet started with data. In the 1st generation web, data was the foremost aim and it is used for display in the webpage of a user. In 2nd generation data with colours are main attraction. Then starting with 3rd generation Images, Videos are used for display of data. In 4th generation we have much more advancement in flaunt of data where transitions and transactions are involved. Now data is large and its handling and management is important. Data is present in diverse forms and its storage space and analyse it for retrieval of data when useful. Smart Data Management must be implemented for better usage of Big Data.

Big means large and Data is the information present in this Digital world. This data is present in the form of structured, unstructured and videos, images, documents of various formats such as .pdf, .docx, .xml etc. The data at hand is in digital form and to store this data we need lot of space i.e., through cloud or servers. The predicament with data arises with managing, storage and security. Data ones moved to another source it may be replicated, and reused. In 2010 the data was grown up to 1.2million petabytes. The storage size of data is getting reduced to smaller and smaller and the amount of data to be stored is getting larger and larger in higher order of magnitude. When actually we necessitate this data for retrieving we need new data techniques and tools.

II. PROBLEMS WITH DATA

A. Data Storage

Data is present everywhere in colossal amount. While storing this huge amount of data in servers and cloud, we need oodles of space, storage devices, storage methods, and cost of manufacturing for this storage devices. Big Data requires big storage that is scalable, efficient, fully automated storage. This data consists of structured and unstructured data that can yield insights in solving business issues and improve customer relations. The next generation storage system is a big issue to today's world. Storage system should support multiple protocols, scale out and up, support business process, security, and flexible architecture and leverage thin and high efficient technologies.

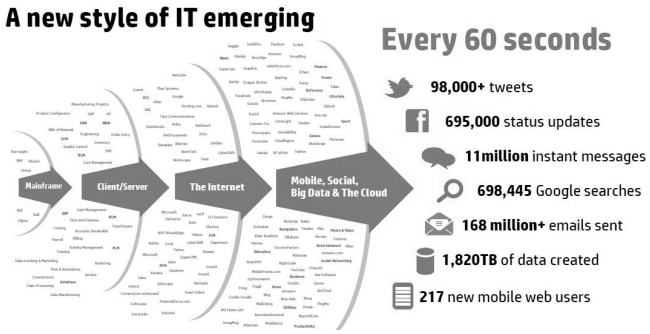


B. Data Management:

Data Management is removing duplicates, standardizing data, incorporating rules to purge incorrect data from inflowing the system in order to create a steadfast source of data. It increases the brunt of your search and research with data citation. By combining data integration, data quality and data management into a unified development and delivery environment, organizations can get the most out of each phase of the data management process. Big data management is the organization, administration and governance of large volumes of both structured and unstructured data.

C. Data Analysis:

Data analytics requires a novel approach to capturing, storing, and analysing data. Business intelligence covers data analysis that relies profoundly on aggregation, focusing on business information. Review the data collection methods, reliable and presentation of data.



Source: http://practicalanalytics.files.wordpress.com/2012/10/newstyleofit.jpg

D. Data Visualization:

Data visualization is the way you amplify, drive traffic, converse information clearly and effectively through graphical means.

E. Data Security:

Data Security alone belongs to one major part of the Big Data. It is the Encryption of data from unauthorized access about corporate data. Mathematical schemes and Algorithms are used to scramble data into unreadable text. Authentication and Backup of the information is to be processed in a secure manner.

III. Facts regarding Data

The thing needed for data to be useful in this digital world:

- New Search Tools.
- Structure to Unstructured data.
- Storage of Data and Information Management techniques.
- Security.

IDC data shows that merely 25% of the data in the world is inimitable. Multiple copies of data are useful at the instance of hardware failure and crashes.

In news bulletin when published information about an incident or a critique various news papers produce data on same incident but amount of storage space for data is more. In social networking an image or video is shared and various shares of equivalent data and replicas are produced. An image by one admin is used for a purpose and same image edited with another name is used by another group. This makes hefty quantity of data of same category. The amounts of money squander on the hardware for storage of data is also mounting.

Today's data is crowd sourced data, with conflicting statements. In a single day there are 2.88billion Google searches. In a single day the data uploaded in YouTube takes 7.8 years to watch all YouTube videos uploaded. In a single day 986million things shared on facebook.

Credit card and Debit card Transactions when done the data is stored online and messages are send to your mobile, mails to your E-Mails, i.e., same data for security purposes is send to all these and data is increasing in exponential rate.
Big data mostly consists of 3 attributes.

- Volume
- Variety
- Velocity

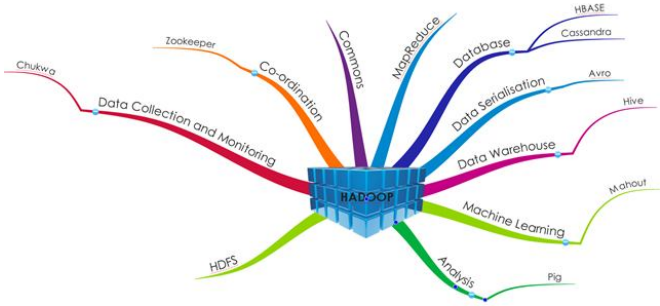
IV. HADOOP - SOLUTIONS TO BIG DATA PROBLEMS

Hadoop is 100% open source software framework that supports reliability and data motion to applications. Hadoop is essentially new way of storing and processing data. Instead of relying on expensive, proprietary hardware and different systems to store and process data, Hadoop enables distributed parallel processing of colossal amounts of data across inexpensive, industry-standard servers that both store and process the data, and can scale without limits.

With Hadoop, no data is too big and in today's hyper-connected world where more and more data is being created every day. Hadoop runs thousands of applications in an assortment of nodes involving thousands of terabytes with high data transfer rate.

Hadoop enables a computing elucidation that is

- Scalable
- Cost Effective
- Flexible
- Fault tolerant



Source:http://stevenimmons.org/2012/02/cio-agenda-big-data-ecosystems/

Hadoop handles all types of data from disparate systems: Structured, Unstructured, log files, images, media, e-mail and with no prior need of schema.

Hadoop, comprised at its core of the Hadoop File System and MapReduce, is very well designed to handle colossal volumes of data across a large number of nodes.

V. Working of Hadoop

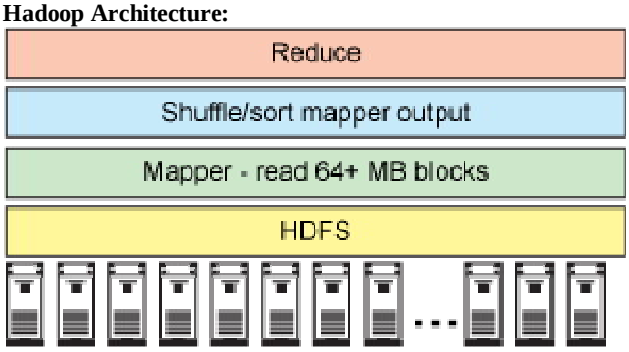
Performing large scale computations is difficult. Multithreading is popular approach for doing parallel computing. Map Reduce is a function used to merge values into a single result.

Before Hadoop:

NFS (Network File System) is the mostly used distributed file system that is still in use. NFS is straightforward and provides remote access to a single logical volume stored on a single machine. NFS is used mainly for its transparency, but limited in power. All volume of data reside on a single

machine does not provide any reliability. It overloads the server when large number of clients tries to retrieve data.

Hadoop Architecture:



In Hadoop data is distributed to all nodes of cluster automatically. The Hadoop Distributed File System will split data files into chunks which are managed by different nodes in cluster. In each chunk is replicated across several machines, so when a single machine fails doesn't result any loss of data. An active monitoring system failure replicates data in response to system failures that results in partial storage. Even chunks are replicated and distributed across single machines they form single namespace, so that they are universally accessible.

Data is distributed across nodes at load time. Input files are broken into other formats specific to application. Files are spread across distributed file system as chunks, and moving computation to the data, instead of moving the data to computation allows Hadoop to achieve high data locality that results in high performance.

In Map Reduce, records are processed in isolation by tasks called Mappers. The output of Mappers is then bought together into second set of tasks called Reducers. Mapping and Reducing tasks run on nodes where individual records of data are present. The flat scalability curve is the major benefits of using Hadoop. HDFS is designed to be robust for various problems. HDFS stores large amount of data that requires spreading of large number of machines. HDFS stores data reliably even if the cluster malfunction happens. HDFS should provide fast, scalable access to data and serves large number of clients by adding more machines to cluster.
Individual machines are referred to as DATANODES. A file can be formed by several nodes and it need not be in the same machine. So the access of file requires cooperation of multiple machines [developer.yahoo.com/Hadoop].

VI. Advantages of using Hadoop solution

- Scalability – When increased in volume also the usage is also scalable.
- Fast response to queries and response times from a database in parallel processing.
- Cost-effective.
- Advanced database analytics.
- Better analytic performance.
- Support of Map Reduce.
- Data Mining Models.
- Sophisticated text analytic capabilities
- Flexibility to implement Hadoop in a Google cloud or on dedicated servers.
- Web interface for data analyst collaboration

REFERENCES

[1] http://www.emc.com/microsites/bigdata/why-big-data-overview.htm
[2] http://www.greenplum.com/industry-buzz/big-data
[3] **A Digital Universe Decade – Are You Ready- IDC paper**
[4] http://www.idc.com/prodserv/FourPillars/bigData/index.jsp#.UUFrEBxTAcM
[5] http://searchdatamanagement.techtarget.com/definition/big-data-management
[6] http://www.sand.com/hadoop-fits-big-data/
[7] http://www.cloudera.com/content/cloudera/en/why-cloudera/hadoop-and-big-data.html
[8] http://stevenimmons.org/2012/02/cio-agenda-big-data-ecosystems/
[9] http://practicalanalytics.files.wordpress.com/2012/10
[10] IDC- Digital universe study by EMC, December 2012
[11] http://developer.yahoo.com/hadoop/tutorial/index.html