Review Article | Impact Factor: 0.621 | ISSN: 2319-507X
Sachin Chavhan, IJPRET, 2014; Volume 2 (8): 870-879 | IJPRET

INTERNATIONAL JOURNAL OF PURE AND APPLIED RESEARCH IN ENGINEERING AND TECHNOLOGY

A PATH FOR HORIZING YOUR INNOVATIVE WORK

OVERVIEW ON BIG DATA SYSTEMATIC TOOLS

MR. SACHIN D. CHAVHAN[1], PROF. S. A. BHURA[2]

1. M.E 1st Year, Department of Computer Science and Engineering, B. N. C. O. E., Pusad, India.
2. Assistant Professor, Department of Computer Science and Engineering, B. N. C. O. E., Pusad, India.

Accepted Date: 15/02/2014 ; Published Date: 01/04/2014

Abstract: Data mining environment produces a big amount of data ,information that need to be analyzed, patterns have to be extracted from that to gain knowledge and more information. Due to increase in use of social media sites, email, document, pdfs and sensor data etc., data is generated at exponential speed. The growth of records has affected all fields, whether it is the business area or the world of science. A larger amount of data gives a better output but also working with it can become a challenge due to processing limitations. Achieving the full use of data in this increasingly digital world requires not only new data analysis algorithms but also a new generation of systems and distributed computing environments to handle increase in the volume, lack of structure of data and the increasing computational needs of massive-scale analytics. In this new era with boom of data both structured and unstructured, it has become difficult to process, supervise and analyze patterns using traditional databases and architectures. This paper, try to review different big data analytical tools for analyzing the data. Also try to cover a variety of platforms for big data.

Keywords: Big Data, Enterprise, Open Source, SQL, HDFS Big Data, Data Mining.

PAPER-QR CODE

Corresponding Author: MR. SACHIN D. CHAVHAN

Access Online On:

www.ijpret.com

How to Cite This Article:

Sachin Chavhan, IJPRET, 2014; Volume 2 (8): 870-879

INTRODUCTION

The term “Big Data” was first introduced to the computing world in 2005, in order to define a great amount of data that traditional data management techniques cannot manage and process due to the complexity and size of this data. Big data is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications. The challenges include capture, duration, storage, search, sharing, transfer, and visualization. The trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data allowing correlations to be found to "spot business trends, determine quality of research, prevent diseases, link legal citations, combat crime, and determine real-time roadway traffic conditions." Big data can have large amount of data i.e. Exabyte’s (EB) equals 1018 bytes, meaning 109 GB.

2. Features of Big Data

There are four features of Big Data: Volume, Velocity, Variety and Veracity, variability, complexity.

• Volume: Many factors contribute to the increase in data volume. Transaction-based data stored through the years. Unstructured data streaming in from social media. Increasing amounts of sensor and machine-to-machine data being collected. Approximately, 800,000 petabytes of data were stored in the world every year and it is going on. Social networking sites generate around 7 TB and 10 TB of data every day respectively. As implied by the term “Big Data”, organizations are facing large volumes of data.

• Velocity: Data is streaming in at unprecedented speed and must be dealt with in a timely manner. RFID tags, sensors and smart metering are driving the need to deal with torrents of data in near-real time. So organizations must be able to analyze this large and varied data in real-time or near real time to find insights in this data.

• Variety: Different types of data called as variety. Data today comes in all types of formats. Structured, numeric data in traditional databases. Information created from line-of-business applications. Unstructured text documents, email, video, audio, stock ticker data and financial transactions.

• Veracity: Refers to the degree in which a leader trusts the used information in order to take decision. So getting the right correlations in Big Data is very important for the business future.

• Variability: In addition to the increasing velocities and varieties of data, data flows can be highly inconsistent with periodic peaks. Daily, seasonal and event-triggered peak data loads can be challenging to manage.

• Complexity: Data comes from multiple sources, and it is still an undertaking to link, match, cleanse and transform data across systems.

Big data should matter to you, because for instance, by combining big data and high-powered analytics, it is possible to:

• Determine root causes of failures, issues and defects in near-real time, potentially saving billions of dollars annually.

• Optimize routes for many thousands of package delivery vehicles while they are on the road.

• Recalculate entire risk portfolios in minutes.

• Quickly identify customers who matter the most. Big Data used as,

• Organizations uses big data to make decisions to get a competitive advantage. For example, cellular companies can analyze call records data to know their quality of service and to initiate the necessary improvements.

• Customer transactions can help a credit card company to detect frauds. Relation between Source big data, Managing and Analyzing Big Data given by following diagram,

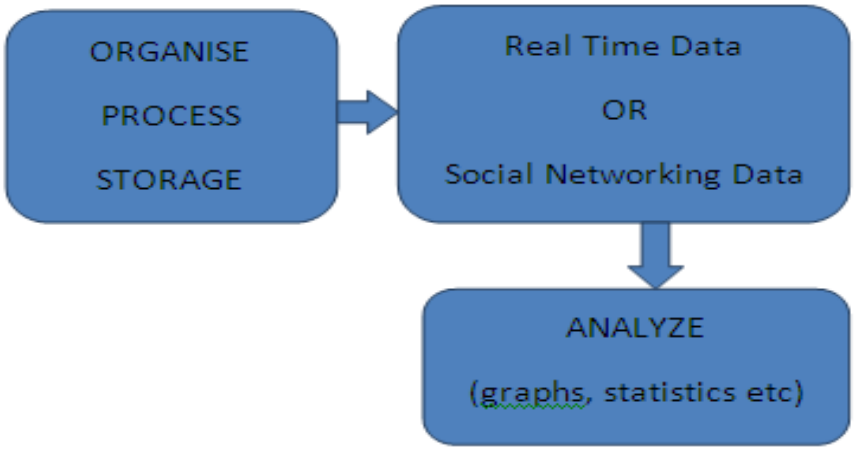


Fig1.Source Big data and Managing and Analyzing big data

2. BIG DATA TECHNOLOGIES AND TOOLS

In this we are going to discuss the tools that are used for solving big data from technology standpoint – Hadoop (HDFS, Map Reduce) which is an open source computing framework and NoSQL which is non-relational database.

2.1 HADOOP

High-availability distributed object-oriented platform or "Hadoop" is a software framework which analyses structured and unstructured data and distribute applications on different servers. Hadoop is a free, Java-based programming framework that supports the processing of large data sets in a distributed computing environment. It is part of the Apache project sponsored by the Apache Software Foundation. Hadoop makes it possible to run applications on systems with thousands of nodes involving thousands of terabytes. Its distributed file system facilitates rapid data transfer rates among nodes and allows the system to continue operating uninterrupted in case of a node failure. This approach lowers the risk of catastrophic system failure, even if a significant number of nodes become inoperative. The Hadoop framework is used by major players including Google, Yahoo and IBM, largely for applications involving search engines and advertising. The preferred operating systems are Windows and Linux but Hadoop can also work with BSD and OS X.Below is an overall Hadoop architecture,

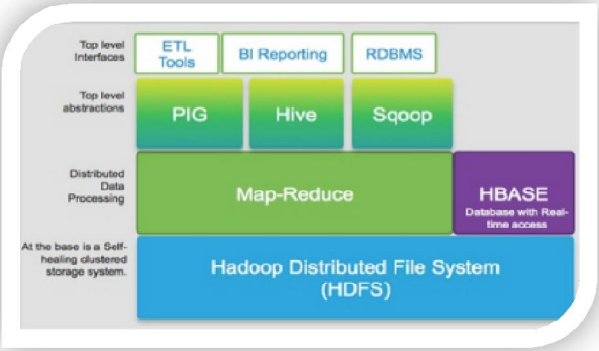


Fig.2.HadoopArchitecture

3. Basic Application of Hadoop

Hadoop is used in maintaining, scaling, error handling, self healing and securing large scale of data. These data can be structured or unstructured. What I mean to say is if data is large then

traditional systems are unable to handle it. Thus, Hadoop comes in the picture. Below are some basic features of Hadoop -

- Hadoop maintains and secures the data by storing and keeping its replica.
- It is focused on scaling according to data usage.
- It can detect and delete the failed task and as well as failed transaction of data.
- It not only recovers the data but also automatically restores the data at its place

3.1 HDFS- (Hadoop Distributed File System)

Is part of Hadoop and is known as a special file system which deals with distribution and storage of large set of data. HDFS stores file as sequence of same size of block except the last block. It also deals with hardware failure and smoothen the data handling.

3.2 Hive –

Hive was initiated by Face book. Apache Hive is an open-source data warehouse system for querying and analyzing large datasets stored in Hadoop files. Hadoop is a framework for handling large datasets in a distributed computing environment. Hive is data warehouse tool which is based on Hadoop and converts query language into MapReduce jobs. It deals with the storage , analysis and queries of large set of data. Query language in hive used as HQL statement. Hive Query Language is similar to standard SQL statement. Hive has three main functions: data summarization, query and analysis.

It supports queries expressed in a language called HiveQL, which automatically translates SQL-like queries into MapReduce jobs executed on Hadoop. In addition, HiveQL supports custom MapReduce scripts to be plugged into queries. Hive also enables data serialization/deserialization and increases flexibility in schema design by including a system catalog called Hive-Metastore.Hive supports text files (also called flat files), Sequence Files (flat files consisting of binary key/value pairs) and RCFiles (Record Columnar Files which store columns of a table in a columnar database way.)

3.3 Hbase –

Hbase is a Hadoop application which runs on top of HDFS. Hbase system represents set of table but Hbase is column oriented database management system i.e. different from the row oriented database management system. Generally if we talk about database then we think of

relational database system but unlikely Hbase is not relational database at all and also it doesn't support Structured Query Language like SQL. Java is preferred language use for Hbase application. One most important feature of Hbase is to real time read or write to large set of data. HBase is not a direct replacement for a classic SQL database, although recently its performance has improved, and it is now serving several data-driven websites, including Facebook's Messaging Platform.

4.4 Pig –

Initiated by Yahoo, became open source in 2007. **Pig** is a high-level platform for creating MapReduce programs used with Hadoop. Do you know why it is named as Pig? It is because it can handle any type of data!! Strange but true. Pig is a high level procedural programming platform developed for simplifying large data sets query in Hadoop and MapReduce. Pig has two components- one is Pig Latin which is programming language and the other is run time environment where PigLatin programs are executed. Pig Latin abstracts the programming from the Java MapReduce idiom into a notation which makes Map Reduce programming high level, similar to that of SQL for RDBMS systems.

4.5 Map reduce -

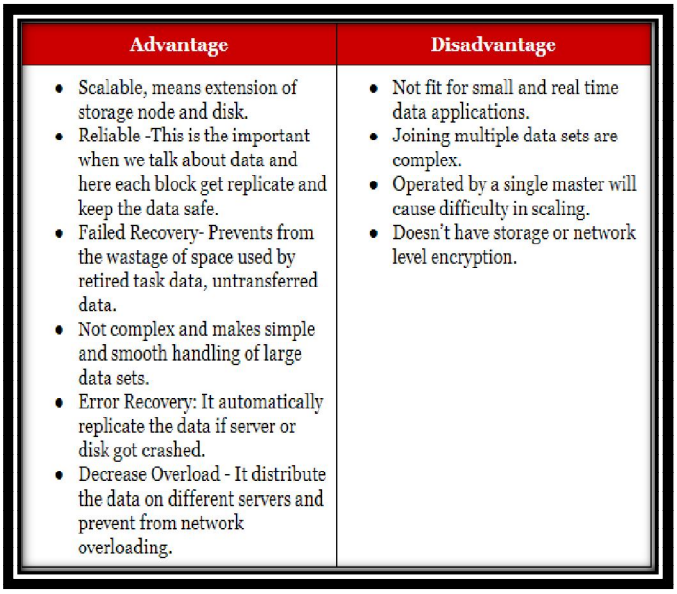
MapReduce is a software framework that allows developers to write programs that process massive amounts of unstructured data in parallel across a distributed cluster of processors or stand-alone computers. It was developed at Google for indexing Web pages and replaced their original indexing algorithms and heuristics. it is developed in 2004.

The framework is divided into two parts:

- **Map**, a function that plots out work to different nodes in the distributed cluster.
- **Reduce**, another function that collates the work and resolves the .

5. Advantage and Disadvantage of Hadoop:

Hadoop having advantage of, platform which provides Distributed storage & Computational capabilities both and having disadvantage of Security which is also one of the major concern because Hadoop does offer a security model But by default it is disabled because of its high complexity.



Advantage	Disadvantage
• Scalable, means extension of storage node and disk. • Reliable -This is the important when we talk about data and here each block get replicate and keep the data safe. • Failed Recovery- Prevents from the wastage of space used by retired task data, untransferred data. • Not complex and makes simple and smooth handling of large data sets. • Error Recovery: It automatically replicate the data if server or disk got crashed. • Decrease Overload - It distribute the data on different servers and prevent from network overloading.	• Not fit for small and real time data applications. • Joining multiple data sets are complex. • Operated by a single master will cause difficulty in scaling. • Doesn't have storage or network level encryption.

Fig.3. Advantage and Disadvantage of Hadoop

6. **NoSQL**

As the term says NoSQL, it means non relational or Non-SQL database, NoSQL database, also called Not Only SQL, is an approach to data management and database design that's useful for very large sets of distributed data.

NoSQL, which encompasses a wide range of technologies and architectures, seeks to solve the scalability and big data performance issues that relational databases weren't designed to address. NoSQL is especially useful when an enterprise needs to access and analyze massive amounts of unstructured data or data that's stored remotely on multiple virtual servers in the cloud. .

Contrary to misconceptions caused by its name, NoSQL does not prohibit structured query language (SQL). While it's true that some NoSQL systems are entirely non-relational, others simply avoid selected relational functionality such as fixed table schemas and join operations. For example, instead of using tables, a NoSQL database might organize data into objects, key/value pairs or tuples.

It refers to Hbase, Cassandra, MongoDb, Riak, CouchDB. It is not based on table formats and that's the reason we don't use SQL for data access. A traditional database deals with structured data while a relational database deals with the vertical as well as horizontal storage

system. NoSQL deals with the unstructured, unpredictable kind of data according to the system requirement.

NoSQL Technologies HBase (part of the Hadoop ecosystem),Cassandra,MongoDB, Riak,CouchDB.

6.1 Cassandra–

Database is used to handle the large set of data when we need to scale the database with high performance. Linear scalability and proven fault-tolerance on commodity hardware or cloud infrastructure make it the perfect platform for mission-critical data. Cassandra's support for replicating across multiple datacenters is best-in-class, providing lower latency for your users and the peace of mind of knowing that you can survive regional outages.

Cassandra's data model offers the convenience of column indexes with the performance of log-structured updates, strong support for demoralization and materialized views, and powerful built-in caching.

Cassandra deals with the fault tolerance and replication of the data. With this we can go deeper in columns, super columns and more. It is a partial relational database system, supports best query capability but don't have joins feature. It follows the column family model map with two dimensional and 3 dimensional. 2D model includes column family with some column in it, while 3D model created by associating super column in column family.

6.2 MongoDB–

Is an agile NoSQL document database, unlike the traditional database which store the data in rows and column, MongoDB stores the document data in binary form of JSON document which is also known as BSON format. It is used for high scalability, availability and performance. MongoDB eschews the traditional table-based relational database structure in favor of JSON-like documents with dynamic schemas (MongoDB calls the format BSON), making the integration of data in certain types of applications easier and faster.

Released under a combination of the GNU Affero General Public License and the Apache License, MongoDB is free and open source software. In MongoDB dynamic schemas are the unit of database, which found in document where set of documents are found in collection while set of collection makes the database. Written in C++.

6.3 Riak-

Is open source NoSQL database system which is designed for availability, fault tolerance, scalability and high performance. It provides three kind of storage key/value store, document oriented store and web shaped store. It also stores documents in the JSON format. When we talk about data modeling, we will see that there is no 'Master', only nodes are there. All nodes are same and don't have different responsibility.

6.4 CouchDB–

Is open source NoSQL database ,distributed, and schemaless. It stores the document data in the JSON format. It also provides feature related to web, like access of document from the web browser through HTTP. It is a NoSQL database that uses JSON to store data, JavaScript as its query language using MapReduce, and HTTP for an API.[1] One of its distinguishing features is multi-master replication. CouchDB was first released in 2005 and later became an Apache project in 2008. JavaScript can also be use to modify the document. In CouchDB document is combination of strings "keys" and "values".

Advantage and Disadvantage of NoSQL:

Advantage	Disadvantage
• Open source. • Scalable - No need to expand server size. • Automatic Repair - Perform automatic repair of failed task data. • Multiple storage system - Data can be store as key value or document. • Simple and Easy layout - Simple and easy to create architectural layout.	• Maintenance - Since it's open source have so less probability of support assistance. • Setup issue - To install and setup it needs skill and system with specific configuration. • No backup - It is mainly for storage and less efficient for data backup.

7. CONCLUSION

To conclude, after the analysis of both Hadoop and NoSQL Big Data Tools, It's all about the usage and needs of an individual or the company. It is impossible to meet the expense of a few tools at a personal level because of the prices and complications; while using open

source systems might pose an outdating and modifications problem. We live in the data age. In the last decades, the continuous boost of computational power has produced an overwhelming flow of data. According to IDC, the size of the digital universe was about 0.18 zettabyte in 2006 and it is forecasting a tenfold growth by the end of 2011 to 1.8 zettabyte (a zettabyte is one billion terabytes).

The result of this is the appearance of a clear gap between the amount of data that is being produced and the capacity of traditional systems to store, analyze and make the best use of this data .There are also security issues involved in choosing the tool. Open source promotes development and innovation and supports developers. As organizations continue to collect more data at this scale, formalizing the process of big data analysis will become paramount. This paper describes different tools associated with managing big data, used to handle such large data sets.

REFERRENCES

1. Manjula M Ramannavar, Mr. Mahesh G Huddar," A Survey on Big Data Analytical Tools", International Journal of Latest Trends in Engineering and Technology (IJLTET),2013.

2. Chanchal Yadav,Shuliang Wang ,Manoj Kumar," Algorithm and approaches to handle large Data", International Journal of Computer Science and Network, Vol 2, Issue 3, 2013.

3. Sherif Sakr, Anna Liu, Daniel M. Batista, and Mohammad Alomari," A Survey of Large Scale Data Management Approaches in Cloud Environments", IEEE COMMUNICATIONS SURVEYS & TUTORIALS, ACCEPTED FOR PUBLICATION,2011.

4. http://www.searchdatamanagement.techtarget.com/definition/NoSQL-Not-Only-SQL

5. http://www.searchcloudcomputing.techtarget.com/definition/MapReduce

6. http://www.searchdatamanagement.techtarget.com/definition/Apache-Hive

7. http://www.sas.com/en_us/Home/Insights/Big Data.html

8. http://www.floost.com/VentureHire-post--5671864

9. http://www.gcn.com/microsites/2012/snapshot-managing-big-data/01-big-data techniques.aspx