IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 3, March 2016

Big Data: An Emerging Future Technology

G Rajesh, S Sudarshan, Kareemunnisa, Ch Swetha

Guru Nanak Institutions Technical Campus, Hyderabad

Anasuya Devi Institute of Technology and Science, Hyderabad

Abstract: The era of big data is that exceeds the processing capacity of conventional database systems. Big data is an emerging paradigm applied to datasets whose size or complexity is beyond the ability of commonly used computer software and hardware tools. Such datasets are often from various sources (Variety) yet unstructured such as social media, sensors, scientific applications, surveillance, video and image archives, Internet texts and documents, Internet search indexing, medical records, business transactions and web logs; and are of large size (Volume) with fast data in/out (Velocity). More importantly, big data has to be of high value (Value) and establish trust in it for business decision making (Veracity). Hadoop is an open-source implementation of the MapReduce model, and is usually used for data-intensive applications such as data mining and web indexing. Using Hadoop, large data sets can be processed across a cluster of servers and applications can be run on systems with thousands of nodes involving thousands of terabytes. Distributed file system in Hadoop helps in rapid data transfer rates and allows the system to continue its normal operation even in the case of some node failures. Hadoop enables a computing solution that is scalable, cost effective, flexible and fault tolerant. Hadoop Framework is used by popular companies like Google, Yahoo, Amazon and IBM etc., to support their applications involving huge amounts of data. This article intends to define the concept of Big Data and the importance of Big Data Analytics and Hadoop.

Keywords: Big Data, Hadoop, Big Data Analytics, MapReduce and HDFS.

I. INTRODUCTION

According to the Gartner definition "Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization" [1].

The IBM Big Data Flood Infographic shows that 2.7 Zettabytes of data exist in the digital universe today. Also according to this study there are 100 Terabytes updated daily through Facebook, and a lot of activity on social networks this leading to an estimate of 35 Zettabytes of data generated annually by 2020. Just to have an idea of the amount of data being generated, one zettabyte (ZB) eguals 10^{21} bytes, meaning 10^{12} GB. [2].

The term "Big Data" was first introduced to the computing world by Roger Magoulas from O'Reilly media 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.

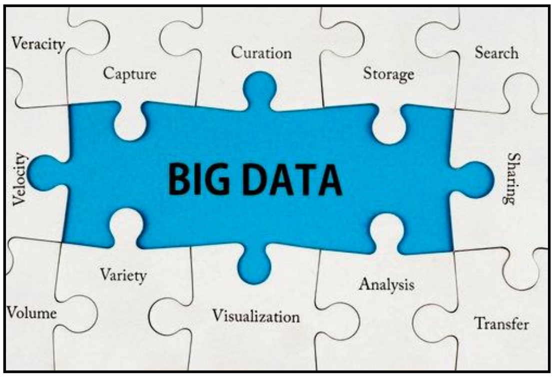

Fig.1Big data Taxonomy [3]

A study on the Evolution of Big Data as a Research and Scientific Topic shows that the term "Big Data" was present in research starting with 1970s but has been comprised in publications in 2008. [4]

According to MiKE 2.0, the open source standard for Information Management, Big Data is defined by its size, comprising a large, complex and independent collection of data sets, each with the potential to interact.[5]

According to Ed Dumbill chair at the O'Reilly Strata Conference, Big Data can be described as, "data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit the strictures of your database architectures. To gain value from this data, you must choose an alternative way to process it." [6]

II. BIG DATA ANALYTICS

Big Data analytics is the process of analyzing and mining Big Data can produce operational and business knowledge at an unprecedented scale and specificity. The need to analyze and leverage trend data collected by businesses is one of the main drivers for Big Data analysis tools[7].

The technological advances in storage, processing, and analysis of Big Data include:

- The rapidly decreasing cost of storage and CPU power in recent years.
- The flexibility and cost-effectiveness of datacenters and cloud computing for elastic computation and storage.
- The development of new frameworks such as Hadoop, which allow users to take advantage of these

DOI 10.17148/IJARCCE.2016.53192

distributed computing systems storing large quantities of data through flexible parallel processing.

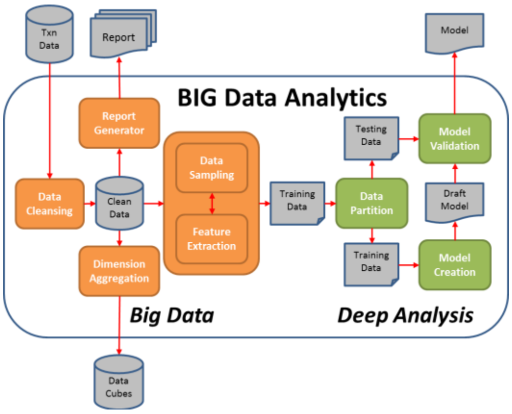

Fig.2. the big data processing part (colored in orange) is usually done using Hadoop/PIG/Hive technology with classical ETL logic implementation.

IV. HADOOP IMPLEMENTATION

Hadoop is an open source framework for processing, storing and analyzing massive amounts of distributed unstructured data. Originally created by Doug Cutting at Yahoo!, Hadoop was inspired by MapReduce, a user-defined function developed by Google in early 2000s for indexing the Web. It was designed to handle petabytes and exabytes of data distributed over multiple nodes in parallel[2].

Hadoop clusters run on inexpensive commodity hardware so projects can scale-out without breaking the bank. Hadoop is now a project of the Apache Software Foundation, where hundreds of contributors continuously improve the core technology.

A.) How Hadoop Works

A client accesses unstructured and semi-structured data from sources including log files, social media feeds and internal data stores. It breaks the data up into "parts," which are then loaded into a file system made up of multiple nodes running on commodity hardware. The default file store in Hadoop is the Hadoop Distributed File System, or HDFS. File systems such as HDFS are adept at storing large volumes of unstructured and sem-structured data, as they do not require data to be organized into relational rows and columns.

Each "part" is replicated multiple times and loaded into the file system so that if a node fails, another node has a copy of the data contained on the failed node. A Name Node acts as facilitator, communicating information such as which nodes are available, where in the cluster certain data resides, and which nodes have failed back to the client.

Once the data is loaded into the cluster, it is ready to be analyzed via the MapReduce framework. The client submits a "Map" job -- usually a query written in Java – to one of the nodes in the cluster known as the Job Tracker. The Job Tracker refers to the Name Node to determine which data it needs to access to complete the job and where in the cluster that data is located. Once determined, the Job Tracker submits the query to the relevant nodes. Rather than bringing all the data back into a central location, processing then occurs at each node simultaneously, in parallel. This is an essential characteristic of Hadoop.

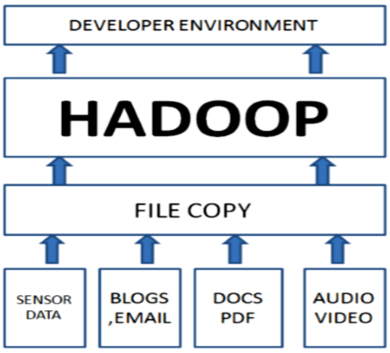

Fig.3 Hadoop supports different types of unstructured data with clustering.

When each node has finished processing its given job, it stores the results. The client initiates a "Reduce" job through the Job Tracker in which results of the map phase stored locally on individual nodes are aggregated to determine the “answer” to the original query, then loaded on to another node in the cluster. The client accesses these results, which can then be loaded into one of several analytic environments for analysis. The MapReduce job has now been completed.

Once the MapReduce phase is complete, the processed data is ready for further analysis by data scientists and others with advanced data analytics skills. Data scientists can manipulate and analyze the data using any of a number of tools for any number of uses, including to search for hidden insights and patterns or to create the foundation to build user-facing analytic applications. The data can also be modeled and transferred from Hadoop clusters into existing relational databases, data warehouses and other traditional IT systems for further analysis and/or to support transactional processing.

B.) Hadoop Technical Components

A Hadoop “stack” is made up of a number of components. They include [2]:

- Hadoop Distributed File System (HDFS): The default storage layer in any given Hadoop cluster;
- Name Node: The node in a Hadoop cluster that provides the client information on where in the cluster particular data is stored and if any nodes fail;
- Secondary Node: A backup to the Name Node, it periodically replicates and stores data from the Name Node should it fail;
- Job Tracker: The node in a Hadoop cluster that initiates and coordinates MapReduce jobs, or the processing of the data.
- Slave Nodes: The grunts of any Hadoop cluster, slave nodes store data and take direction to process it from the Job Tracker.

DOI 10.17148/IJARCCE.2016.53192

NoSQL data stores like Cassandra and HBase are also used to store the results of MapReduce jobs in Hadoop. In addition to Java, some MapReduce jobs and other Hadoop functions are written in Pig, an open source language designed specifically for Hadoop. Hive is an open source data warehouse originally developed by Facebook that allows for analytic modeling within Hadoop.

C.) NoSQL

A related new style of database called NoSQL (Not Only SQL) has emerged to, like Hadoop, process large volumes of multi-structured data. However, where as Hadoop is adept at supporting large-scale, batch-style historical analysis, NoSQL databases are aimed, for the most part (though there are some important exceptions) at serving up discrete data stored among large volumes of multi-structured data to end-user and automated Big Data applications. This capability is sorely lacking from relational database technology, which simply can't maintain needed application performance levels at Big Data scale.

In some cases, NoSQL and Hadoop work in conjunction. HBase, for example, is a popular NoSQL database modeled after Google BigTable that is often deployed on top of HDFS, the Hadoop Distributed File System, to provide low-latency, quick lookups in Hadoop.

NoSQL databases currently available include:

- HBase
- Cassandra
- MarkLogic
- Aerospike
- MongoDB
- Accumulo
- Riak
- CouchDB
- DynamoDB

The downside of most NoSQL databases today is that they traded ACID (atomicity, consistency, isolation, durability) compliance for performance and scalability. Many also lack mature management and monitoring tools. Both these shortcomings are in the process of being overcome by both the open source NoSQL communities and a handful of vendors such as DataStax, Sqrrl, 10gen, Aerospike and Couchbase that are attempting to commercialize the various NoSQL databases.

Steps required for setting up a distributed, single node Hadoop cluster, backed by HDFS[5].
step 1- Install Ubuntu
step 2- Install Java in it
step 3- Add a dedicated Hadoop System user
step 4- Configure SSH
step 5- Now test SSH by connecting to your server
step 6- Disable IPV6
step 7- Hadoop Installations
step 8 -Configuring HDFS
step 9- Configure directory where Hadoop will stores files.
step 10- Format the HDFS file system via in name node
step 11- Start your own single node cluster
step 12- Run a map reduced job
step 13- Copy Local data to HDFS
step 14- Run map reduce job

V. COMPLIMENTARY BIG DATA APPROACHES

Hadoop, NoSQL and massively parallel analytic databases are not mutually exclusive. Far from it, Wikibon believes the three approaches are complimentary to each other and can and should co-exist in many enterprises. Hadoop excels at processing and analyzing large volumes of distributed, unstructured data in batch fashion for historical analysis. NoSQL databases are adept at storing and serving up multi-structured data in near-real time for Web-based Big Data applications. And massively parallel analytic databases are best at providing near real-time analysis of large volumes of mainly structured data.

Historical analysis done in Hadoop can be ported into analytic databases for further analysis and/or integrated with structured data in traditional enterprise data warehouses, for example. Insights gleaned from Big Data analytics can (and should) be productionized via Big Data applications. Enterprises should aim for flexible Big Data architectures to enable these three technologies/approaches to share data and insights as seamlessly as possible.

There are a number of pre-built connectors to help Hadoop developers and administrators perform such data integration, while a handful of vendors -- among them Pivotal Initiative (formerly EMC Greenplum, Cetas, et al.) and Teradata Aster -- offer Big Data appliances that bundle Hadoop and analytic databases with preconfigured hardware for quick deployment with minimal tuning required. Others, namely a start-up called Hadapt, offers a single platform that provides both SQL and Hadoop/MapReduce processing on the same cluster. Cloudera is also pursuing this strategy with its Impala project and Hortonworks via the open source Stinger Initiative.

In order to fully take advantage of Big Data, however, enterprises must take further steps. Namely, they must employ advanced analytics techniques on the processed data to reveal meaningful insights. Data scientists perform this sophisticated work in one of a handful of languages or approaches, including SAS and R.

The results of this analysis can then be visualized with tools like Tableau or operationalized via Big Data applications, either homegrown or off-the-shelf. Other vendors, including Platfora and Datameer, are developing business intelligence-style applications to allow non-power users to interact with Big Data directly.

The bottom-line is that Big Data approaches such as Hadoop, NoSQL and massively parallel analytic databases are complimentary not just to each other but to most existing data management technology deployed in large enterprises. Wikibon does not suggest enterprise CIOs undertake whole-sale "rip-and-replacement" of existing enterprise data warehouses, data integration and other data management technologies for Big Data approaches.

DOI 10.17148/IJARCCE.2016.53192

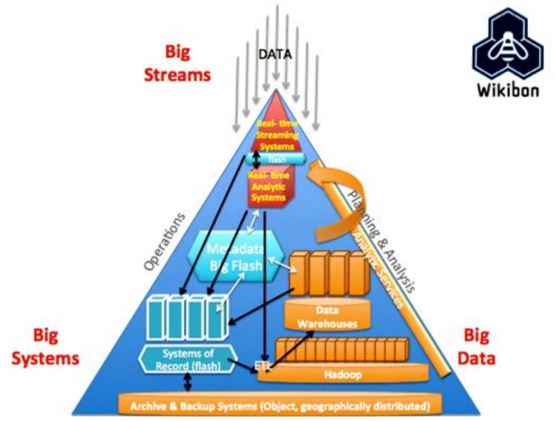
Big Streams
DATA
Wikibon
Real- time Streaming Systems
flash
Real- time Analytic Systems
Operations
Planning & Analysis
Analytic Services
Metadata Big Flash
Data Warehouses
Big Systems
Systems of Record (flash)
ETL
Hadoop
Big Data
Archive & Backup Systems (Object, geographically distributed)

Fig4. Flash and Hyperscale Changing Database and System Design Forever

Wikibon believes CIOs must think like portfolio managers, re-weighing priorities and laying the groundwork toward innovation and growth while taking necessary steps to mitigate risk factors. Replace existing data management technology with Big Data approaches only where it makes business sense and develop plans to integrate Big Data with remaining legacy data management infrastructure as seamlessly as possible. The end goal should be the transformation to a modern data architecture.

VI. CONCLUSION

Big data is the next frontier for emerging and innovative productivity. In 2020, big data finds its place in Industrial leadership. There is need for structuring data in all sector of economy. Cloud computing and Hadoop are in great demand in several business organizations. In future, Hadoop will become one of the most required technology for Cloud computing. IT vendors should help enterprises identify the most profitable and practical Big Data use cases and develop products and services make Big Data technologies easier to deploy, manage and use. Organizations are looking to expand Hadoop use cases to include business critical, secure applications that easily integrate with file based applications. There is need for tools that do not require specialize skills and programmer. Thus, New Hadoop develoments must be easier for users to operate and to get data in and out.

REFERENCES

[1] http://gartner.com/it-glossary/big-data.
[2] G. Noseworthy, Infographic: Managing the Big Flood of Big Data in Digital Marketing, 2012 http://analyzingmedia.com/ 2012/infographic-big-flood-of-big-data-in-digital-marketing/
[3] http://www.thecloudcomputingaustralia.com/2014/09/20/new-big-data-taxonomy-report-by-cloud-security-alliance/
[4] H. Moed, The Evolution of Big Data as a Research and Scientific Topic: Overview of the Literature, 2012, Research Trends, http://www.researchtrends.com
[5] MIKE 2.0, Big Data Definition, http://mike2.openmethodology. org/wiki/Big_DataDefinition
[6] E. Dumhill, "What is big data?", 2012 , http://strata.oreilly.com/2012/01/what-is-big-data.html
[7] http://wikibon.org/wiki/v/Big_Data:_Hadoop,_Business_Analytics_ and_Beyond

DOI 10.17148/IJARCCE.2016.53192