Big Data – A Brief Study

Ravi Narasimhan, Bhuvaneshwari T

Abstract— In this research article we have provided a brief of the current industry buzzword called Big Data and covered the components of big data from a Hadoop perspective. The study will help to have a thorough understanding of big data and its various components in the hadoop framework.

Index Terms — Big Data, Hadoop, Framework, HDFS, Big Data Components, 3 V's, Big Data Characteristics, Hive.

1 Introduction

As a generic definition, Big Data as we see, is something so huge and complex that it is impossible for traditional systems and traditional data-warehousing tools to process and work on them. Data (Big Data) is generated by machines, generated by humans, and also generated by Mother Nature.

Big data can neither be worked upon by using traditional SQL like queries nor can the relational database management system (RDBMS) be used for storage. Therefore a wide variety of scalable database tools and techniques have evolved. Hadoop, an open source distributed data processing system (which includes HIVE data warehouse system and HBase database) is one of the prominent and well known solutions. The NoSQL has gained prominence as a non relational database with the likes of MongoDB, DynamoDB from Amazon and Cassandra from Apache.

In the year 2001, Doug Laney (then at Gartner) articulated Big Data the definition of Big Data to consist of the 3 V's - volume, velocity and variety.

Day after day, zillions of data is generated all over the universe. Some are machine generated, some human generated and other are generated by nature. Machine generated data can come from Airplanes, Cars, CERN systems (particle physics Lab-Hadron collider) and likewise. Human generated data consists of data from social media like facebook, twitter, etc. Nature generated data comes from whether, genome sequencing, deep ocean data, the universe data, etc.

The research article is divided in the following sequence: Starting with the introduction, we talk about the characteristics of big data (5V's). It is followed with a descriptive note on the various components of Big Data based on Hadoop framework.

Towards the end, we provide a conclusion and future work.

2 Big Data Characteristics – The 5 V's

Big Data is characterized by the 5 V's – Volume, Velocity, Variety, Veracity and Value. In addition we can also think about Variability.

A) Volume - The size of data has increased from Mega Bytes and Giga Bytes to Terabytes & Petabytes and Exabytes. Over the past few years, because of the super low-price in hardware and storage / memory costs, it has become economical to store data, any data. This has led to storage of huge amount of Structured and Unstructured data from various sources, like Social media, from sensors and machine-to-machine data, from airplanes, from digitalization of books, etc. To store data, we have various databases including Greenplum and Hadoop. The storage of data in Hadoop is in the Distributed File System called HDFS which makes Data available to multiple computing nodes. Usage pattern involves the below three steps:

1. Loading data into HDFS
2. MapReduce Operations
3. Retrieving results from HDFS

It is more of a batch processing and hence suited for analytical purposes.

B) Velocity - Data is created at an increasing speed. Users want the full data ASAP, meaning, they want the Data to be given to them almost instantly. Now-a-days flow of data is enormous and continuous. Data is coming to us at an unprecedented speed. Streaming data has made it possible to have a real time provision of data.

With so much of data being generated at very high speed, the terminology called streaming data came into prominence. It's not only about incoming data, but also about insights and decision that needs to be taken at equally high pace.

A common terminology called NoSQL, also called Not Only SQL plays a major role in taking care of the velocity related challenges with big data. Some of the NoSQL technologies include IBM Infosphere Streams, Twitter's Storm, Yahoo S4, etc.

C) Variety - Data for Big Data comes in from various sources. Data is mainly divided into three different types - Structured Data, Semi-Structured Data and Un-Structured Data.

Structured data is what the traditional Data-base management systems work with. It consists of Rows and Columns and resides in fixed fields in a file. Un-Structured data does not have and does not adhere to a pre-defined data like representation. E-mails, video and audio are some of the unstructured data files. Semi-Structured Data does not confirm to a specific arrangement but consists of Ttage to separate the data elements.

Big Data comes from multiple sources and are structured, semi-structured and unstructured. They don't necessarily come in with a nice relational like structure. MarkLogic specializes in document stores encoded in XML or dedicated CML stores. Social Network relations being Graphical in nature, Graph databases like NeoJ4 makes it more efficient.

Other NoSQL Databases also take care of the variety of Data in Big Data.

D) Veracity - As Big Data becomes bigger and the multiple sources of big data are ever increasing, there is lot of chances that there is huge inconsistency and abnormality in the Data. The reliability or the trustworthiness of the data becomes questionable. Data inconsistency and abnormality is one major challenging factor that is a part and parcel of Big Data.

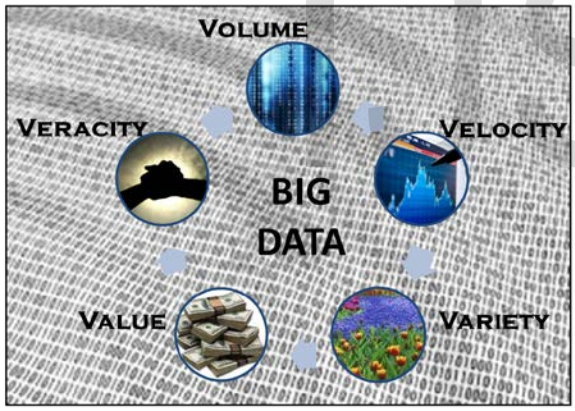

Fig 001- V's of Big Data

E) Value - Cost is one major factor that all organizations need to look into when it comes to big data implementation, or for that matter any software package or framework implementation. The initiative of entering into Big Data is a very critical and the Value needs to be consciously deliberated in the value to price perspective. A thorough understanding of Big Data is a must and the usefulness via Value of big data initiative to the organization needs to be clearly put down on paper.

3 BIG DATA COMPONENTS ON HADOOP FRAMEWORK

Apache Hadoop is an open-source framework that deals with distributed computing of large datasets across clusters of computers using simple programming models. Created by Doug Cutting and Mike Caferella in the year 2005, it is name after a toy elephant. Scaling up from Single servers to thousands of machines with Local storage and computation are the advantage that Hadoop offers. This is one of the major advantages that Hadoop offers as we can use in-expensive hardware.

Hadoop Modules:

A) Hadoop Distributed File System (HDFS)

It is a distributed file system that helps to store large amounts of data in a reliable manner providing fault tolerant file system. HDFS follows a Master/Slave structure, in which we have a one or more devices called as slave devices are controlled by one device known as master device. The master node is called name-node and manages the cluster metadata. Slave node is called data-node and stores data. It is a java based file system.

HDFS provides for rack awareness, fault tolerance, scalability, efficiency, reliability and minimal data motion apart from other utilities and other features.

B) Hadoop YARN / MapReduce

MapReduce / YARN is a cluster recource management and has been built as a programming model in the Hadoop framework to process large amounts of data in a distributed & par allel environment on a cluster. YARN is the latest release of MapReduce in Hadoop 2.0. YARN has been separated into entities including a global ResourceManager; a per-application ApplicationMaster; a per-node slave NodeManager and a per application container running NodeManager. Yarn is the heart of Hadoop and ResourceManager and NodeManager being help to manage the applications in a distributed manner.

C) HBase

HBASE is a Hadoop database which is a non relational (NoSQL) database that runs on top of HDFS. HBASE allows for random, real time read/write access for the big data, is columnar, provides fault tolerant storage and fast access. It also provides for transactional kind of capabilities by allowing updates, inserts, deletions etc. It is modeled after Google's BigTable and supports billions of rows with million of columns. Tables in HBase can serve as inputs or outputs for jobs running in MapReduce.

D) Pig

Apache PIG is a scripting language enabling users to write complex MapReduce transformations including summarizing / aggregation, joining, sorting etc. PIG provides for data operations including ETL, research and iterative data processing. Apache PIG's properties include Ease of programming, optimization opportunities and extensibility so that users can write their own functions to do specific processing as per requirements. One of the main features of PIG is parallel processing enabling it to handle very large datasets.

E) Hive

Hive is a Data-Warehouse software tool used for managing, querying, summarizing and analyzing large data sets. HiveQL is a SQL like language is used for querying petabytes of data in hive. It is used to analyze data in HDFS and provides full support for map/reduce. The advantage that Hive offers is that it is very similar to the traditional SQL language, it is fast over big datasets, it is scalable and extensible and provides for various reporting.

F) Sqoop

Sqoop is a software tool designed to transfer bulk data between Hadoop and relational databases. Sqoop is used to import data from external databases into HDFS or HBASE or HIVE. It can also be used to extract data from HDFS or HBase or Hive into relational data stores. Sqoop allows for data imports from and data exports to external relational databases and parallel data transferring. It uses simple SQL query as well as saved jobs that can run number of times for importing the updates regarding the data between Hadoop and relational databases.

G) ZooKeeper

Zookeeper offers operations services in the Hadoop framework. It is a centralized service used for maintaining configuration information, named registry, provides data synchronization and group services that are used by distributed applications. Zookeeper's architecture supports high-availability through redundant services. It allows for the various distributing processes to coordinate between themselves through a shared hierarchical name space of registers called znodes.

H) Avro

Avro is a data serialization system, which serializes data in a compact binary data format and provides for rich data structures and a container file for storing persistent data. Avro relies on schemas to read and write data. It uses JSON (Java script open notation) for defining data types & protocols. It makes use of wire format for communicating between hadoop nodes, and between client program & services.

I) Cassandra

Apache Cassandra is a high availability, highly scalable and high performance open source distributed database management system having capability of handling huge amount of data across multiple servers. It provides for fault tolerance and is decentralized.

J) Mahout

As per Wikipedia, Apache Mahout is a project of the Apache Software Foundation to produce free implementations of distributed or otherwise scalable machine learning algorithms focused primarily in the areas of collaborative filtering, clustering and classification. It contains a library of machine learning algorithms Mahout has the data science inbuilt to find meaningful patterns and insights in the data (in HDFS). Mahout is used to produce scalable machine learning algorithms. Mahout supports collaborative filtering, clustering, classification and frequent itemset mining.

K) Spark

Apache spark is a fast data analytics and machine learning algorithmic engine used for processing data at a large scale. Spark is integrated with hadoop and has an advanced analytical engine which makes it 100 times faster them hadoop map reduce by utilizing in-memory processing.

L) Flume

Flume is a reliable distributed service for efficiently collecting aggregating and moving large amount of LOG Data. It helps the users make most use of valuable log data. It allows for streaming data from multiple sources, collecting high volume real time web logs.

Some of the BI / Analytics Applications includes of others:

M) R

'R' is a open source free software which allows for statistical computing and graphics. It offers efficient data handling and storage facilities, a suite of operators for various calculations, well developed simple and effective programming language and innumerous user defined and developed packages which can be downloaded and used.

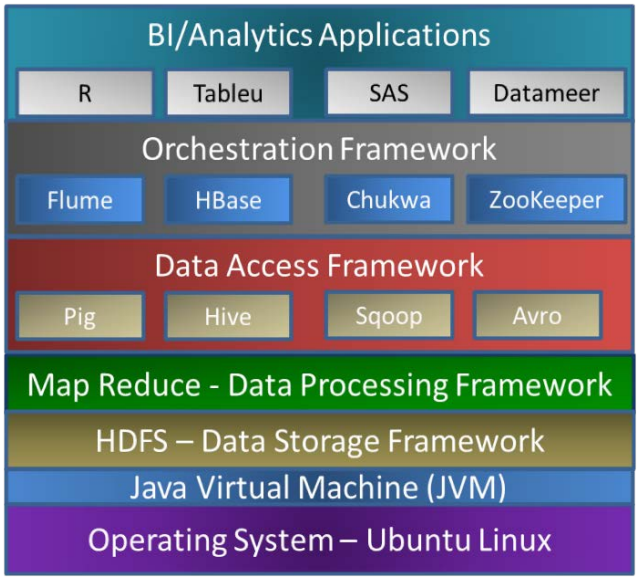

Fig 002: Hadoop Framework Components

N) Tableau

Tableau software is an visual analytics software providing business intelligence and analytics on the cloud. It has been

positioned at a leader position in Gartner research's 2014 magic quadrant research on Business intelligence and analytics platforms.

O) SAS

SAS is a statistical software providing for statistical, analytical reporting and ETL processing on large amounts of data. It is widely used across various industries and provides enterprise wide solutions

P) Datameer

Datameer is a data analytics and data visualization software for big data providing data integration, dynamic data management and self service analytics. It offers quick and effective reporting facility for big data with neat visual features.

4 CONCLUSION AND FUTURE WORKS

Big Data is here to stay and is living up to its hype. In this research article, details about Big Data has been discussed taking the Hadoop Framework as a base. We have dwelt into the characteristics of Big Data and provided deep information on the various components of big data from a Hadoop perspective.

Future work would involve a detailed study on challenges and issues with big data and the use cases in various industries.

5 REFERENCES

[1] www.cio.com
[2] Big Data basics from oreilly: http://strata.oreilly.com/2012/01/what-is-big-data.html
[3] Apache Software Foundation. Official website www.Apache.hadoop.org
[4] Hadoop Big Data Sandbox provider: Official website http://hortonworks.com/hadoop/
[5] Open source statistical package. Official website http://www.r-project.org/
[6] Data Analytics and Data Visualization provider: Official website. www.datameer.com
[7] Visual Analytics Business Intelligence Software provider. www.tableausoftware.com
[8] SAS Institute, official website www.sas.com
[9] http://www.bigdatalandscape.com/
[10] White, Tom. Hadoop The Definitive Guide 2nd Edition. United States : O'Reilly Media, Inc., 2010.
[11] A. Vailaya, "What's All the Buzz Around "Big Data?"", IEEE Women in Engineering Magazine, December 2012, pp. 24-31,
[12] S. Madden, "From Databases to Big Data", IEEE Internet Computing, June 2012, v.16, pp.4-6