

Big Data and Advanced Analytics Tools

Rahul Kumar Chawda¹

Maulana Azad National Institute of Technology,
Bhopal, India 462003
rahul.chawda3@gmail.com

Dr. Ghanshyam Thakur²

Maulana Azad National Institute of Technology,
Bhopal, India 462003
ghanshyamthakur@gmail.com

Abstract: Nowadays a big data analytics is a very broad area for both academia and industry. Big data analytics has attracted intense interest from all academia and industry recently for its attempt to extract knowledge, information and wisdom from big data. Big data and cloud computing, two of the most important trends that are defining the new emerging analytical tools. Big data analytical capabilities using cloud delivery models could ease adoption for many industry, and most important thinking to cost saving, it could simplify useful insights that could providing them with different kinds of competitive advantage. Many companies to provide online Big Data analytical tools some of the top most companies like Amazon Big data Analytics Platform, HIVE web based Interface, SAP Big data Analytics, IBM InfoSphere BigInsights, TERADATA Big Data Analytics, 1010data Big Data Platform, Cloudera Big Data Solution etc. Those companies analyze huge amount of data with help of different type of tools and also provide easy or simple user interface for analyzing data. This paper deals with the study of big data 5V's definition, Analysis requirements, tools, frame works and different type of cloud based big data analytics tools provide by different companies and functioning of Hadoop or MapReduce Process.

Keywords: *Big data; Advanced Analytic Tool; HIVE; SAP Big Data; Hadoop; Teradata Analytics.*

I. INTRODUCTION

Nowadays big data and cloud has been evolved into a buzzword from academia to industry all over the world. Most of the companies such as Google, Facebook, Amazon, Twitter and Baidu already built many systems. Governments including USA, China, India and Japan etc. provide a lot of funds to support the big data research. While big data and cloud is extremely hot, it also comes with challenges.

User requirement of minute to minute information increases the size of data. Data Storage capacity is increasing. Storage capacity gives huge size of data and but less or know information. It is big question what will a data collector do with this massive size of data? Answer is simple person wants knowledge from this data which satisfies users need.

Data Analytics tool help in understanding the data by applying various tool and methods. This paper presents some analytics tool which is widely used in the industry.

II. BIG DATA

Big Data, according to Wikipedia(Dec 2015), is “Big data is a broad term for data sets so large or complex that traditional data processing applications are inadequate. Challenges include analysis, capture, data curation, search, sharing, storage, transfer, visualization and information privacy.[1] The term often refers simply to the use of predictive analytics or other certain advanced methods to extract value from data, and seldom to a particular size of data set. Accuracy in big data may lead to more confident decision making. And better decisions can mean greater operational efficiency, cost reduction and reduced risk” [1]. The data set contain structured database, unstructured database and semi-structured database such as social network, search engines or GPS (Global Positioning System) data. According to gather 5V's definition, big data has five characteristics: volume, variety, velocity, variability or value as show in figure I. [2]



Figure I Big Data V's Model

A. Volume (Large amount of data):

Volume means adatasets that are huge. This data can be generated every second Ex. Images, Video, Audio, emails and sensor data share every time. We are talking about zettabytes, but Yottabytes or brontobytes of data. Defined datasets that are too large and easily amassed into Zettabytes, even Petabytes of information. Below Table – I show datasets volume size. Large volume of datasets is not only an analysis issue, but also a storage issue.

Table – I Dataset Volume Size

Value	Name	Abbreviation
1000 ¹	Kilobytes	KB
1000 ²	Megabytes	MB
1000 ³	Gigabytes	GB
1000 ⁴	Terabytes	TB
1000 ⁵	Petabytes	PB
1000 ⁶	Exabytes	EB
1000 ⁷	Zettabytes	ZB
1000 ⁸	Yottabytes	YB
1000 ⁹	Brontobytes	BB
1000 ¹⁰	Geopbytes	GEB

B. Velocity (fast processing velocity):

It means fast dataset has being produced and data move around. For example, post comment, image, video, audio file on facebook; watching and uploading videos on YouTube; Big data technology now allows us to analyze the data without store data ever putting it into databases.

C. Variety (many type of data and source):

This refers to the different types of datasets that contain structured, unstructured and semi-structured data, such as emails, audio files, documents, video, images, log files, click streams, call records, or financial transactions. Many different attributes in multiple dimensions in the datasets provides more and more information for traditional database management tools or application to handle.

D. Veracity (Correct – meaning useful data)

This basically refers to the messiness or trust worthiness of the data. With many forms excellence and accurateness are less controllable. Ex. Facebook posts with asterisk (*), hash tag (#), underscore (_), tiled (~), smiles, strikers, abbreviation, typos and colloquial speech. Big data analytics tools and technology now allow us to work with these types data. The huge volumes often make up for the lack of excellence and accurateness.

E. Value (low – density data value):

Big data tends to have a relatively low value density, as compared to the data we manage in traditional system. For Example logistics industry

have best mode to transport for goods based on weight and value or ratio of business relevance to the size of the data.

Big data frequently involves a development of cultural and technical changes throughout your business or provide new business opportunities to expanding your sphere of inquiry to exploiting new acumens as you combine both traditional and big data analytics. It involves making “sense” out of huge volumes of mixed data that in its raw from absences a data model to define what each element means in the context of the others. There are some new issues like discovery, iteration, flexible capacity, mining and predicting or decision management. For Ex. Many organization have no idea whether or not social data huts light on sales trends or challenge comes with figuring out which data element relate to other data elements, and in what data storage capacity. This process of discovery not only involves sightseeing the data to understand how we can use it but also formative how can relate to traditional database.

III. BIG DATA ANALYSIS REQUIREMENTS

Big data analyzing requirements are some of methods you can use to find meaning and discover unseen relationships in big data. There are three major significant requirements is:

- Minimize data movement is basically all about conversing computing resources. In traditional analysis scenarios, data is brought into the computer system, processed and then sent to next destination. ForEx. Filpkart sales data might be extracted from e-system, transformed into a relational data type, and loaded into an operational data store structured for generating report. The volume of data increase very time, this type of ETL (Extract, Transform and Load) architecture become increasingly less well-organized or just more and more data to move around. It makes more intelligence to store and process the data in the same place.
- Use existing skills means is new data or new data sources comes the need to obtain new skills. Most of the time existing skill sets will determine where analysis can and should be done. Mostly organizations have more and more people who can analyze data using SQL or MapReduce, it is very

important to be able to support both type of processing.

- Attend to data security is essential part for many business application. Basically data warehouse users are habituated not only to with judgment defined metrics, dimensions and attributes, but also to a reliable set of management policies and security controls. These rigorous processes are often poor with unstructured data source and open source analysis tools.

IV. BIG DATA ANALYZING TOOLS

Big data analyzing tools have five key approaches to analysis data and generating analysis data or reports. [3]

- Discovery tools:** This tool is very useful throughout the information lifecycle for frequently, intuitive exploration and analysis of information form any combination of structured, unstructured and semi – structured source. These tools permit analysis alongside traditional BI source systems. With help of BI tools users can draw new reports, come to meaningful or useful conclusion, and make informed decisions quickly.
- BI (Business Intelligence) tools:** This tool is very useful for analyzing, reporting and performance management, primarily with transactional data form data warehouse and information systems.
- In-Database Analytics:** These analytical techniques that allow data processing are applied directly with in the database. In-Database Analytics include a variety of technique like credit scoring, fraud detection, trends or finding patterns and relationship in your data.
- Hadoop (High-availability distributed object oriented platform):** This is a most popular open-source platform for scalable, reliable and distributed computing. These are very useful for pre-processing data to identity macro trends or find chunks of information, such as out of rang values. All organization mostly use Hadoop as an ancestor to advance forms of analysis.
- Decision Management:** consist of predictive modeling, self-learning, and business rules

to take informed action based on the existing context. Mostly this type of analysis allows distinct recommendations across multiple channels, maximizing the value of every client to communication.

All of these methodologies have a role to play uncovering hidden relationship. Most of the organization provide analyzing tools and framework which are available to explore big data and are capable to perform analytics. We look at these all tools and frameworks provide by various organization.

HIVE Web based interface [4]

This is an alternative option to using the Hive command line interface. The Hive web based interface, abbreviated as HWI or Hive WebUI, is a simple graphical user interface (GUI). This is an interactive web interface which is basically designed for administration purpose of the Hive and as well as for querying the database. In Hive user can create tables and delete tables and also browse the database schema. Mostly more user can execute the queries by supplying it from the webUI. It is not actually analytical platform but Hive easy and interactive by using the webUI. Hive webUI to work, the user should have Hive configured and deployed on the computer system which requires Hadoop as well for processing.

IBM InfoSphere BigInsights[5]

IBM InfoSphere BigInsights is a Big Data analytics platform provided by IBM, which support not the same type of analytics under one roof. InfoSphere is built on top of Hadoop to improve its capabilities and provides an interactive UI on it for analyzing the big data. This has built-in analytics capability, with social data analyzer for analyzing social media data, text analytics for in receipt of insights from hug amount textual data, machine data analytics for analyzing machine data like sensors and GPS related data and InfoSphere also the integration with other big data technologies. InfoSphere provides a SQL interface, namely as BigSQL, Jaql a Declarative query language and spreadsheet interface called Bigsheets for all tools analyzing and exploring big data easily. BigSQL, Jqal and Bigsheets modules of InfoSphere are some of the core components of the system. BigSQL offer the facility to user to

sightsee the database schema and analysis the data using structured query language. Jqal a Declarative query language to facilitate analyzing of structured, unstructured and semi-structured data. BigSheets a web-based analysis and visualization tool using familiar, excel spreadsheet interface that enable to analyzing hug amount of data and long running data collection jobs. This system can upload data from multiple sources such as web crawlers, database, text files, Json, csv files etc. and can store it in the distributes file system for processing.

SAP Big Data Analytics [6]

SAP is one of the most leading provider of big data analytics platforms. SAP is one of the 1st company to introduce in-memory database for analytics. This aims to make Event Stream Processor (EPS) an excellent in SAP HANA (High Performance Analytic Appliance). It is building in-memory database was to provide a single database for both transactional and analytical data processing, mostly refer to as Online Transaction Processing and Online Analytical Processing systems. SAP's ESP is a standalone, overall streaming analytics platforms that has a long, rich history as one of the original complex event processing event processing. It has broad base of customer's services in telecommunication, financial, energy, retails, manufacturing, transportation and logistics, and all public or private sectors. SAP's provide significant road map item for EPS is to integrate it as a service that can run within its SAP HANA in-memory database. This will afford streaming analytics capabilities to SAP HANA's strong analytics capabilities.

TERADATA Big Data Analytics [7]

The Aster large data analytics appliance is solution from TERADATA for big data analytics. Aster is essentially a database developed by TERADATA supporting row and column based storage. It's the key element of their big data analytical platform which contains of Aster SQL-MapReduce, Aster Database, which is mainly an interfacing Structure Query Language for Aster Database. It is a combination of hardware and software by connecting several nodes with Infniband in its place of running

hadoop on commodity hardware with traditional network connections.

Cloudera Big Data Solution [8]

Cloudera was formed in 2008 to help enterprise companies use Apache hadoop to get more valuable output of all of their data. Cloudera's open source platform, is the most popular distributed big data technologies. The big contributions of Cloudera to big data world is the abstraction over different big data technologies such as Hadoop, Hive, HCatalog etc.it is provides easy to use technologies without going into technical details and also provides the system management, deployment, configuration, security management, diagnostics and repots generation etc.

HortonWorks [9]

This technology funded in 2011 by the ex-Yahoo engineers. The similar concept as cloudera to provide different big data technologies for enterprise computing. HortonWork data platform for enterprise computing which builds on Hadoop and also provide different type of analysis like real time, batch and interactive. Its supports like different type of technologies for data integration and data flow control. Thus technologies are Apache Falcon, Sqoop and Flume this are the part of the platform and provides easy and systematic access for handling data in and beyond Hadoop. If we are deploying and configuring HortonWork data platform entails substantial expertise and training to use it along with other technologies experience.

Amazon Big data Analytics Platform [10]

Amazon big data analytics platform provide to analyzing very huge data sets requires momentous compute capacity that can differ in size based on the volume of input data and analysis required. Big data workloads in an ideal world suited to pay as meter cloud computing model and as per your requirement you can easily change or resize your environment on AWS to meet your wants without having to delay for further more hardware, or being required to over invest to provision sufficient capacity. Hadoop or Hive both can be accessed via the amazon exposed API's or via launching an Elastic Compute Cloud by Amazon instance.

[illegible]

In the Above Table - II Comparison of different Big Data Platforms/Tools we mention features of the companies. The following features have short definition is:

Analysis Web Interface: Many companies provide web interface. This interface to provide that will connect to data or display content to the client. Client analysis data with help of analysis tool in this interface.

Hadoop Support: A wide range of companies to support Hadoop. Hadoop use many companies and organization for researchers and production. Client also boost companies to provide online Hadoop supports.

Spark Support: Spark is a very powerful open source processing engine built around speed, ease of use and sophisticated analytics. It was originally developed at UC Berkeley in 2009 [18]. Today many Companies rapidly adopt Spark or provide Spark Support.

Analytics: Now days many company to offer Analytics-as-a-Services. This AaaS is very useful for researcher and organization because they needs big data management and analysis without capital expenditure necessary to keep those tasks on -premise. Company know the finish up for valuable, actionable insights that client can use to work better and faster. Company must make clear which type of data are most likely to support analytics that client can use and know which data they are received [19].

Database Support: company are offer to analysis data from many source like data warehouse, RDBMS, Hadoop, NoSQL DBMS.

Data Management (DM): Most of the company to offer new range of analytics environment. The high quality data across multiple analytical data stores including data warehouse, RDBMS, Analytical RDBMS appliances, Hadoop clusters, Hadoop appliances and NoSQL DBMS. Data Management that variety of different data management tools for individually different platform to a common suite of tools supplying data to all platforms.

Non-expert users Support: It means some people are not sound technical like some scientists, researcher etc. So some of the company to offer Non-expert user Support.

Business users Support: All company to provide business users support because all organization need to analysis our data.

V. HADOOP [14]

Hadoop means High-availability distributed object oriented platform is constructed on the Hadoop distributed file system and MapReduce Programming Model originally developed by google in 2004. Now days it's maintained by an Apache software foundation as an open source software. Hadoop is most widely used big data technology for analyzing the huge data sets. Hadoop batch and parallel processing nature make it perfect for crunching huge amount of data very easily and in cost effective manners.

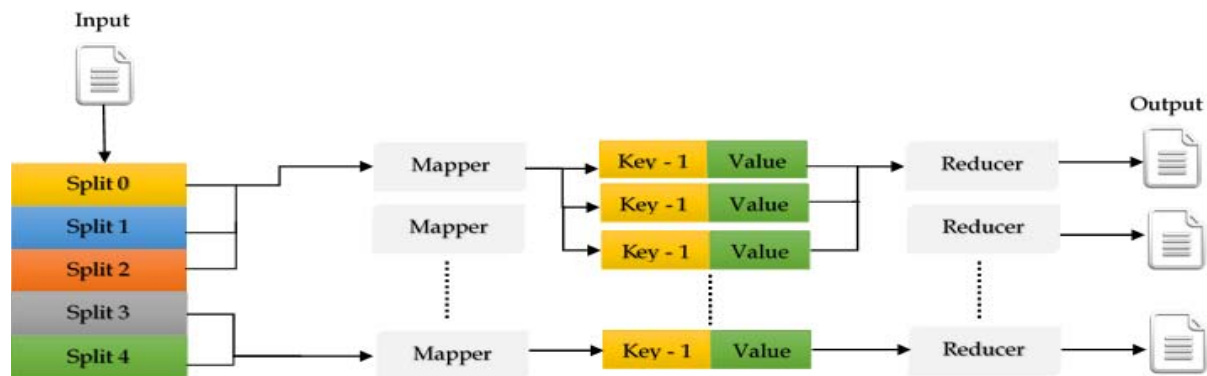



Figure 2. Hadoop MapReduce Programming Model

MapReduce for data processing, sharing, clustering data because it has the capability to take a query over a dataset, divide it into various small flotsam and jetsam, and run on parallel to the cluster. This MapReduce programming model based on the concept of key/value pairs as shown in the Figure II, where the user inputs data files which split into many large piece of 128 MB in size and passed to the mapper where different large pieces of the same file assigned to other mappers for parallel Processing. Mapper break down the input file and convert it to key / value pairs for processing and generates in-between key / value pairs after processing as show in the figure II. Reducer then loads all these key / value pairs and sort it in order to group the related keys together, reducer then process these group to produce the output file. Each reducer usually has zero or one output file at the end. We can describe the map and reduce processes as are:

Mapper (Key, Value) → Data (Intermediate Key, Value)
Reduce (Intermediate Key, List (Value)) = 
Final Output

Hadoop basically based on the MapReduce programming model, which have need different modules for parallel processing. Hadoop have a multiple nodes, NameNode, DataNode, Job Tracker, and Task Trackers all this handle parallel processing.

NameNode: It's basically manages the file system in the form of tree data structure, metadata and DataNodes that store the actual data and responsible for all jobs on the file system like move, delete, or add file. NameNode executes periodic checkpoints of the namespace and to help keep the size of the containing log of HDFS modification within certain limits at the NameNode. Before a user requests the operation on the file system, the NameNode alert the concerned DataNode to process the request and sends the result back to the user.

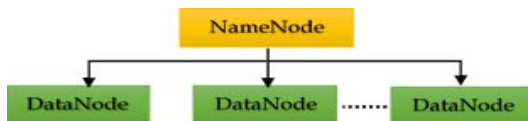


Figure 3. NameNode Manage DataNodes

JobTracker: This is a responsible for taking in requests form a user and assigning the tasks to TaskTracker on the DataNode where the data is

locally present. It's also responsible for periodic checking of TaskTracker that is it up and running by pinging the node so that in case of node die, processing can be transferred to other live node.

TaskTracker: A TaskTracker is a daemon that accepts jobs for the mapper, reducer and shuffling or sorting tasks from JobTracker. The TaskTracker continue sending a heartbeat message to the JobTracker to alert that it is active mode. It is start and monitor the Map and Reduce Tasks and sends progress status information back to the JobTracker.

Hadoop Distribute File System: HDFS is used for storing and retrieving data within Hadoop architecture. HDFS have some characteristics like File Storage, Large Files, Streaming Data Access and Designed for Commodity Hardware, Client Interface.

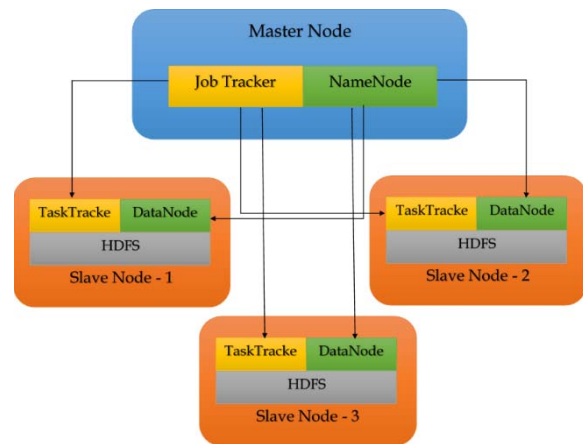


Figure 4. Hadoop Architecture with three Nodes

The above figure IV shows the Hadoop architecture in a cluster setup which has three nodes. Form the above figure we can see JobTracker is communicating with the TaskTracker for assigning jobs such a Map, Reduce and Shuffling. NameNode is responsible for handling all the data nodes in a cluster and DataNode is assigned to each node in the cluster. In addition to most important NameNode there may be a less important NameNode in the Master node which is typically used to store the replica of metadata of the NameNode for handing Letdowns [14][15][16][17].

VI. CONCLUSION & FUTURE WORKS

This paper, introduces a big data 5'Vs definition. The basic big data analyzing requirements is to find meaning and discover unseen relationships in big data. The purpose of big data analyzing tools with

five key approaches to analysis data and generate analysis data or reports. We look at all tools and frameworks provide by various organization and comparison table. InfoSphere is best tool so far. For the future work, we plan to develop our own Hadoop system Comparing Apache Spark, storm and Map Reduce with Performance Analysis using K-Means.

REFERENCES

- [1]https://en.wikipedia.org/wiki/Big_data
- [2]<http://www.ibmbigdatahub.com/blog/why-only-one-5-vs-big-data-really-matters>.
- [3] <http://www.oracle.com/technetwork/database/options/advanced-analytics/ds-oracle-advanced-analytics-1510025.pdf>.
- [4] <https://cwiki.apache.org/confluence/display/Hive/HiveWebInterface>
- [5]<http://www-03.ibm.com/software/products/en/Infosphere-biginsights-standard-ed>
- [6] <http://fm.sap.com/data/UPLOAD/files/Big%20Data%20Analytics%20Guide.pdf>.
- [7] <http://bigdata.teradata.com/>
- [8] <http://cloudera.com/>
- [9] <http://hortonworks.com/hdp/>
- [10]<https://aws.amazon.com/training/course-descriptions/bigdata/>.
- [11] <https://www.1010data.com/>
- [12] <http://www.ibmbigdatahub.com/blog/why-only-one-5-vs-big-data-really-matters>.
- [13]<https://www.hpe.com/in/en/solutions/big-data.html>
- [14] Tom White, Hadoop; The Definitive Guide, 3rd ed., Mike Loukides and Meghan Blanchette, Eds.: O'Reilly, 2012.
- [15] Jeffrey Dean and Sanjay Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," in OSDI'04: Sixth Symposium on Operating System Design and Implementation, 2004.
- [16] Aditya B. Patel, Manashvi Birla, and Ushma Nair, "Addressing Big Data Problem Using Hadoop and Map Reduce," in 2012 Nirma University International Conference on Engineering (NUICONE), 2012, pp. 6-8.
- [17] Hadoop Wiki. [Online]. <http://wiki.apache.org/hadoop>