# Novel Approaches in Big Data

**Manisha Valera, Malvika Vasani**

***Abstract*— Big Data is a term used to understand the datasets that due to their complexity. Big Data are now swiftly growing in all science and engineering domains, including physical, biological and biomedical sciences. The Big Data challenge is becoming one of the most exciting opportunities for the next years. This novel includes the information about what is big data, Technologies and methods, Techniques. Big Data is thus very significant to increase productivity growth in the entire world since it is affecting public segments. Big data refers to capacious data which ranges in Exabyte's (1018) and beyond. The archetype of processing huge datasets has been shifted from centralized architecture to distributed architecture. In this paper, we provide a survey of Big data research, while prominence the specific concerns in Big data world. In this paper we have also discussed the importance and usage of this technology.**

***Index Terms*— Big Data, Data Forms, Technologies and Methods, Techniques.**

## I. INTRODUCTION

In Big data the information comes from multiple, heterogeneous, autonomous sources with complex relationship and continuously growing. Up to 2.5 quintillion bytes of data are created daily and 90 percent data in the world today were produced within past two years [1]. for example Flicker, a public picture sharing site, where in an average 1.8 million photos per day are receive from February to march 2012 [2]. this shows that it is very difficult for big data applications to manage, process and retrieve data from large size of data using existing software tools. It's become challenge to extract knowledgeable information for future use.

## II. WHAT IS BIG DATA?

Big data describes a full information management strategy that includes and integrates many new types of data and data management beside traditional data.
Big data has also been defined by the five V's : Volume, Velocity, Variety, Value, Veracity.

### *A. Volume*

The quantity of generated and stored data, that is being manipulated and analyzed in order to obtain the required results. It is the task of Big Data to convert low-density data into high-density data, that is, data that has value.

***Manuscript received March, 2017.***
***Manisha Valera**, Computer Engineering, Indus University, Ahmedabad, India.*
***Malvika Vasani**, Computer Engineering, Indus University, Ahmedabad, India.*

### *B. Velocity*

At which the data must be processed. As an example, consumer eCommerce applications seek to combine mobile device location and personal preferences to make time sensitive offers.

### *C. Variety*

It represents the type of data that is stored, analyzed and used. The type of data stored and analyzed varies and it can consist of location coordinates, video files, text, audio, data sent from browsers, simulations etc.

### *D. Value*

It is all about the quality of data that is stored and the further use of it. Large quantity of data is being stored from mobile phones call records to TCP/IP logs.

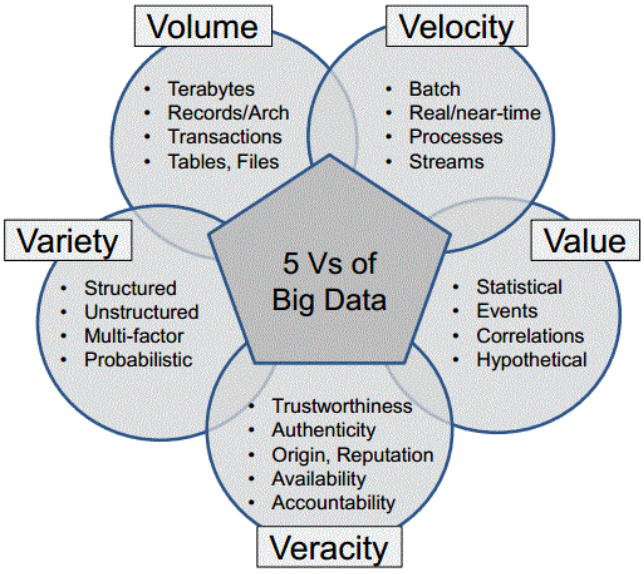


**FIG. 1.** Five Vs

### *E. Veracity*

It is the possible consistency of data is good enough for Big Data. For example, if A is sending an email to B, B will have the exact content that A sent it, if else, the email service will not be reliable and people will not use it.

Current software technologies try to overcome the challenges that "V's" raises. One of these is Apache Hadoop, which is open source software that its main goal is to handle

large amounts of data in a reasonable time.

Growth of Big Data is needed

- Increase of storage volumes
- Increase of processing power
- Availability of data (different data types)

## III. DATA FORMS

### A. *Structured*

The complete data is organized in terms of Entities.

- Relations or Classes (Similar entities are grouped together).
- Attributes (Same descriptions for entities existing in the Same groups).
- Schema (All Entities in the group have a description associated with it.

The complete data are present and follow same order.
All of them have same format and length defined.

### B. *Semi Structured*

- Data is available in many formats,
  - Data Systems
  - File Systems e.g., Web data
  - Data Exchange Formats, e.g. Scientific data
- Data that is not completely structured,
  - Grouping of Similar Entities and semantically organized.
  - Entities may not have same attributes in the group
  - Order of attributes not important & all attributes may not be required

In a Group, size and type of same attributes may differ.

### C. *Unstructured*

- Data – Any type.
- No Format and No Sequence in data
- No Rules in data
- Changeableness is spread across the data.
- Examples – Audio, Images, Video

## IV. STEPS TO APPROACH BIG DATA ANALYSIS

The analysis of big data can be performed as follows:

A. Collection
B. Organization
C. Analysis
D. Visualize
E. Understand, Modify, Restructure

### A. *Collection*

This is the basic step for performing an analytics. The data on which analytics is performed must be collect form different sources. One should select right gear and techniques to acquire records. One can collect data over internets, customer forms, interviews or interrogation, proper feedbacks etc.

### B. *Organization*

The data collected need to be organized in proper forms. The data must be cleaned before organization. The data organization can be like data ware houses.

### C. *Analysis*

The knowledge of business for which analytics is done , can help to guide analytics for desired results. The understanding of commercial enterprise for which analytics is computed, can help to manual analytics for desired consequences. The identification of tools and technique foster the analytics. The tools depending upon the application such as desktop, clouds must effectively selected. The data analyst must cleverly choose the architecture for analysis. For e.g. statics analytics can be carried out with help of R or MATLAB where as dynamic analysis which occurs in changing environment can be carried in HADOOP architecture [8].

### D. *Visualize*

As the data set is very large, one can frequently zoom and note each detail. The table layout doesn't show useful at such example. The visual aids or demo graphs like pie charts. Visualization provide a way to maintain context by showing co-related variable.

### E. *Understand, Modify, Restructure*

The analytics over huge information help us to analyze the causes effects over an occasion. You can come up with decisions and predictions using the result of evaluation.
The new technologies can be used to adapt the upon the results or decision. Accordingly, the technique of collection, organization , and analytics goes on and on. depending upon the consequences or choice , we want to rebuild the strategies [8].

## V. TECHNOLOGIES

### A. *Hadoop*

Hadoop is a structure that can run applications on frameworks with a large number of hubs and terabytes. Apache Hadoop comprises of the Hadoop kernel, Hadoop distributed file system (HDFS), map reduce and related projects are zookeeper, Hbase, Apache Hive.

Hadoop Distributed File System consists of three Components: the Name Node, Secondary Name Node and Data Node. The multilevel secure (MLS) issues of Hadoop by utilizing security improved Linux (SE Linux) convention. In which various sources of Hadoop applications run at different levels.

Hadoop is commonly used for distributed batch index building; it is desirable to optimize the index capability in

ongoing [3].

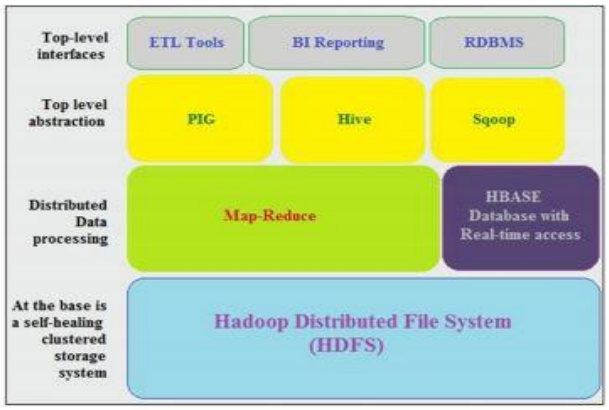


**FIG. 2.** Hadoop Architecture

**Hadoop is:**
**Reliable:** The software is fault tolerant, it expects and handles hardware and software failures
**Scalable:** Designed for massive scale of processors, memory, and local attached storage Distributed: Handles replication [3].

Table 1: The Ecosystem of Hadoop

| Elements\Ecosystems | Hadoop |
|---|---|
| Distributed File Systems | HDFS, FTP File system, Amazon-S3, Windows Azure Storage Blobs |
| Distributed Resource Management | YARN framework |
| SQL Query | **HIVE**: A data warehouse component |
| Machine Learning | **Mahout**: A Machine learning component |
| Stream Processing | **Storm**: real-time computational engine |
| Graph Processing | **Giraph:** A framework for large-scale graph processing |
| Management Interface | **Zookeeper**: A management tool for Hadoop cluster |
| Stream tool | **Flume**: a service for efficiently transferring streaming data into the Hadoop Distributed File System (HDFS). |
| Pluggable to RMDB | **Sqoop**: transfer data between Relational Database Management System (RDBMS) and Hadoop |
| Data Flow Processing | **Pig**: a high level scripting data flow language which expresses data flows by applying a series of transformations to loaded data [ 7 ]. |
| NoSQL database | **HBase**: based on Big Table, and column-oriented |

### *B. MapReduce*

The preparing column in the Hadoop biological system is the MapReduce structure. The system permits the specification of an operation to be applied to a large data set, it divides the problem and information, and run it in parallel. MapReduce is as follows [4]:

**map –** the function takes key/value pairs as input and generates an intermediate set of key/value pairs.
**reduce –** the function which merges all the intermediate values associated with the same intermediate key.

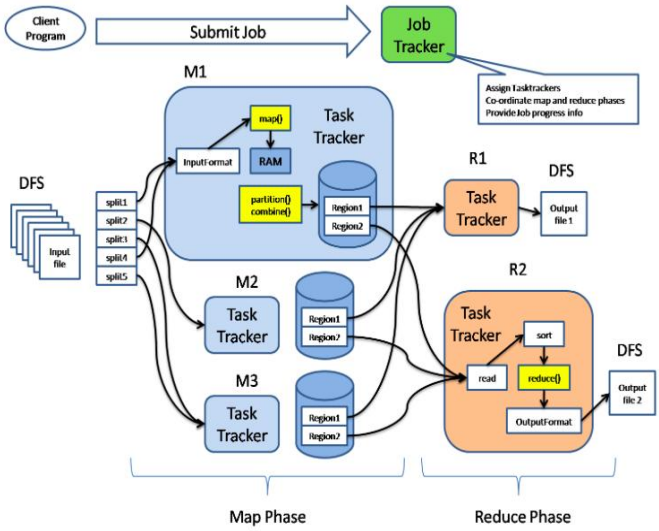


**FIG. 3.** MapReduce Architecture

**Map Reduce Components:**
**A. Nodes:**
**1. Name Node:** manages HDFS metadata, it doesn't directly deal with files.
**2. Data Node:** stores blocks of HDFS—default replication level for each block: 3.
**B. Trackers:**
**1. Job Tracker:** schedules, allocates and monitors job execution on slaves—Task Trackers.
**2. Task Tracker:** runs Map Reduce operations.

### *C. Hive*

Hive is an appropriated operator phase, a decentralized framework for building applications by networking local system resources. Apache Hive data warehousing component, which offers an inquiry dialect called Hive SQL that interprets SQL-like inquiries into Map Reduce jobs automatically.

Applications of apache hive are SQL, oracle, IBM DB2.

Architecture is separated into Map- Reduce-situated execution, Meta information data for information stockpiling, and an execution part that receives a query from user or applications for execution.

The pros of hive is more secure and usage are great and very much tuned.

Hive is a data warehouse infrastructure tool to process structured data in Hadoop. It resides on top of Hadoop to summarize Big Data, and makes querying and analyzing easy.

Hive looks very much like traditional database code with SQL access. However, because Hive is based on Hadoop and MapReduce operations, there are several key differences. The first is that Hadoop is intended for long sequential scans, and because Hive is based on Hadoop, you can expect queries to have a very high latency (many minutes).

### *D. No-SQL*

No-SQL database is a way to deal with information administration and configuration that is valuable for extensive sets of distributed data. These databases are in general part of the real-time events that are distinguished in process sent to inbound channels yet can likewise be viewed as an empowering innovation following analytical capabilities such as relative search applications.

The drawback of No-SQL is Immaturity, No ordering support, No ACID, Complex consistency models, Absence of standardization.

NoSQL (commonly referred to as "Not Only SQL") represents a completely different framework of databases that allows for high-performance, agile processing of information at massive scale. In other words, it is a database infrastructure that as been very well-adapted to the heavy demands of big data. The efficiency of NoSQL can be achieved because unlike relational databases that are highly structured, NoSQL databases are unstructured in nature, trading off stringent consistency requirements for speed and agility. NoSQL centers around the concept of distributed databases, where unstructured data may be stored across multiple processing nodes, and often across multiple servers. This distributed architecture allows NoSQL databases to be horizontally scalable; as data continues to explode, just add more hardware to keep up, with no slowdown in performance. The NoSQL distributed database infrastructure has been the solution to handling some of the biggest data warehouses on the planet – i.e. the likes of Google, Amazon, and the CIA.

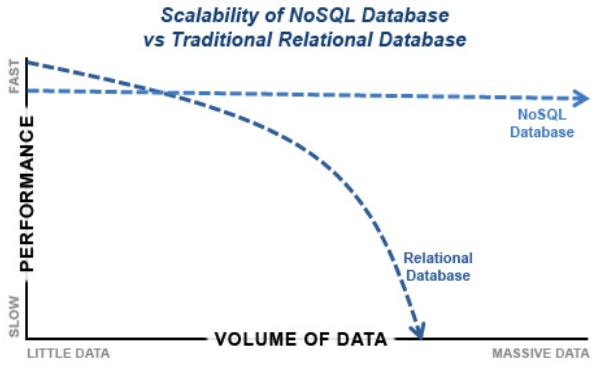


**FIG. 4 Scalability of NoSQL Database vs Traditional Relational Database**

### *E. HPCC*

HPCC is an open source phase utilized for registering and that gives the administration to taking care of big data workflow. HPCC data model is defined by the user end according to the requirements. HPCC system is a single platform having a single architecture. HPCC framework was intended to analyze the gigantic amount of data for the purpose of solving complex problem of big data. HPCC framework depends on big business control dialect which has the decisive and on-procedural nature programming language the main components of HPCC are [5]:

HPCC Data Refinery: Use parallel ETL engine mostly.

HPCC Data Delivery: It is massively based on structured query engine used.

## VI. TECHNIQUES

There are a lot of techniques to start with a project. Some of the tools which have frequent usage are summarized here.

Association rule learning: A set of techniques for discovering interesting relationships, i.e., "association rules," among variables in large databases [6].

**Data mining:** One of the most important terms related to data-driven decision making and describes it as "searching or 'digging into' a data file for information to understand better a particular phenomenon."

**Cluster analysis:** Cluster analysis is a type of data mining that divides a large group into smaller groups of similar objects "whose characteristics of similarity are not known in advance."

**Crowd sourcing:** Crowd sourcing collects data from a large group of people through an open call, usually via a Web2.0 tool. This tool is used more for collecting data than for analyzing it.

**Machine learning:** Traditionally computers only know what we tell them, but in machine learning, a subspecialty of computer science, we try to craft "algorithms that allow computers to evolve based on empirical data.

**Text analytics:** A large portion of generated data is in text form. Text Analytics is the process of converting unstructured text data into meaningful data.

## VII. IMPORTANCE OF BIG DATA

Data is taken from multiple source and integrated across various environments which when analyzed can help us give answers to following [6]:

**1.** Time & Cost reductions.

**2.** Customized & Optimized Market Offerings & New Product Development.

**3.** Strategy Development & Smart Decision Making.

In a Business environment, there are a lot of decisions that are to be taken on the basis of Data & associated analytics and in simple terms, we could define it as Big Data when combined with powered analytics, lot of business related tasks can be accomplished such as:

1. Root Cause Analysis can be conducted in real time for associated defects, failures and issues.
2. POS based generated coupons based on Consumer Behavior.
3. Risk Portfolio – Quick Calculations/Re-Calculations can be conducted in minutes.
4. Conducting Fraud Detection & use of Fraud analytics before hitting organization.

## VIII. USAGE AREAS OG BIG DATA

Big data is used efficiently in numerous fields. Some of them are listed below [6]:

1) Automotive industry
2) High technology and industry
3) Oil and gas
4) Telecommunication sector
5) Medical field
6) Retail industry
7) Packaged consumer products
8) Media and show business
9) Travel and transport sector
10) Financial services
11) Social media and online services
12) Public services
13) Education and research
14) Health services
15) Law enforcement and defense industry

## REFERENCES

[1] Xindong Wu, Fellow, IEEE, Xingquan Zhu, Gong-Qing Wu, and Wei Ding‖ Data Mining with Big Data‖ IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 1, JANUARY 2014
[2] F. Michel, "How Many Photos Are Uploaded to Flickr Every Day and Month?" http://www.flickr.com/photos/franckmichel/6855169886/, 2012.
[3] Jimmy Lin "MapReduce Is Good Enough?" The control project. IEEE Computer 32 (2013).
[4] Jimmy Lin ―Map Reduce Is Good Enough? The control project, IEEE Computer 32 (2013).
[5] Samarati P. Protecting respondent's privacy in microdata release. IEEE Trans Knowl Data Eng. 2001;13(6):1010–27.
[6] Tanvi Ahlawat and Dr. Radha Krishna Rambola "Literature Review On Big Data", International Journal of Advancement in Engineering Technology, Management & Applied Science (Volume 3 Issue 5) May 2016.
[7] Kyuseoks Shim, MapReduce Algorithms for Big Data Analysis, DNIS 2013, LNCS 7813, pp. 44–48, 2013.
[8] Shweta S. Lokhande, Prof. Rahul Patil "A Study on Big Data Analytics, Approches and Challenges", International Journal of Innovative Research in Computer and Communication Engineering (Vol. 4, Issue 11) November 2016.