Big Data technologies and its opportunities and challenges in present scenario

Mr. Shantanu Pradyut Chakraborty[1]

Miss Sreya Bhattacharjee[2]

Miss Sayanti Chakraborti[3]

TECHNO INDIA UNIVERSITY, DN-20

SALT LAKE

And Tata Consultancy Services

[1]Mr. Shantanu Pradyut Chakraborty and [2]Miss Sreya Bhattacharya, [3]Miss. SayantiChakrabarti, Assistant Professor, Techno India University, Data Science and Tata Consultancy Service Kolkata

Big Data technologies and its opportunities and challenges in present scenario

ABSTRACT

Big Data, a **contemporary popular adoption** in academia and industry, is the biggest **trump card for a data-driven world**. When fathomed and processed till the best depths, the Hercules amount of data that is generated on a momentary basis, which intertwines structural and functional units at the core, can provide a **plethora of consolidated information**. Such information is skeletal to any action at individual or organisational level for the **decision-making ball game**, the acumen of our life-cycle. The **amplified hustle** that hovers Big Data is entailed to the **sperm of data flow** and **cost cordial, resource optimized management** of the namesake. The mushrooming of data compounds to an annual approximate of 40%. Prospects are very high that by the next fiscals, in 2020, the Internet will have about 50 billion devices corded to it leading to an estimated data production escalation by 44 times to 2009, reaching nearly to a 45 ZB; thus stipulating the volume growth rate of business data to a doubling in every 1.2 years. Such explosion entails along with, **a myriad of challenges** in terms of data **collection, curation, processing, storage, management, maintenance, security, analysis, transfer, visualization, retention, flexibility**; which need to be addressed carefully and efficiently, as inherent to Big Data Analytics. This study encircles the arena of Big Data aiming to collectively unearth and delve into the terminologies, attributes, definitions, characteristics, tools, technologies and components related to Big Data. There is a parallel focus on SWOT Analysis of Big Data Analytics, addressing in a jiffy the advantages, challenges, future research scopes and open issues and limitations, associated with Big Data and its components. Lastly, this study also intends to survey and summarise the **step-wise Big Data processing cycle**, in association with the functionalities of the different constituent tools and technologies in anutshell.

Keywords: big data, technologies, terminologies, opportunities, volume, variety, velocity, business intelligence, data warehouse, data lake, growth, exponential

Data, Information,Knowledge/Insights,Actionable Intelligence, Informed Decision Making,Enhanced BusinessValues

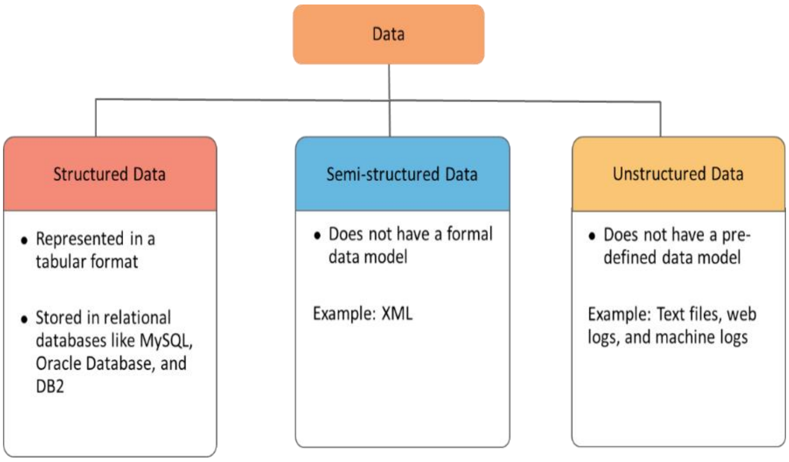
INTRODUCTION

“Data is the new science. Big Data holds the answers.” ~ Patrick Paul Gelsinger, (CEO – VMware; Ex-President and Ex-COO, EMC Information Infrastructure Products)

The Internet of Things (IoT), the World-Wide Web (WWW) and their uber-connected nature have broken dormancy to a burgeoning rise in data, in turn, leading to the coining of the term: Big Data. Big Data, hence, refers to **capacious data volumes, diverse, dynamic and mammoth for acquisition and wrangling exploiting conventional resources**. As on date, people worldwide, to the tune of 2 billion are connected to the Internet, and over 5 billion individuals can be hailed as mobile phone owners. Prospects are very high that by the next fiscals, in 2020, the Internet will have about 50 billion devices corded to it. Data production will hence undergo an estimated surge of 44 times greater than in 2009, reaching nearly to a 45 ZB; thus **stipulating the volume growth rate of business data to a doubling in every 1.2years.**

As on date, Big Data, holds magnanimous importance, **validating the real time cascade of greater data generation, analysis of the former; sharpened accuracy of analysed data; robust consequential decision making and process defining.** The need of the hour is hence, to gather, store, process, understand and further analyse the data at hand, both in an effective and efficient way, so as to make proper and informed business decisions, and also to explore data furthermore to exploit and leverage the value ofdata.

Data arises from a multitude of sources, characterised always by composition, condition and context, can be classified broadly at birth as follows: [15]


Data
Structured Data
• Represented in a tabular format
• Stored in relational databases like MySQL, Oracle Database, and DB2
Semi-structured Data
• Does not have a formal data model
Example: XML
Unstructured Data
• Does not have a pre-defined data model
Example: Text files, web logs, and machine logs

- Unstructured Data – non conformity topre-defined data models/processing formats; originates from social media feeds/posts, log files, e-mail, multimedia data, text messages/chats, flat files, memos and a host of others; no or little metadata; constitutes approximately 80% of the total data generated globally.
- Semi-structuredData–notabidingtoaformaltabularstructure;inherentstructural tags to segregate semantic elements and denote hierarchy; e.g. mark-up languages like XML, HTML; self-describing XML, HTML, JSON.

- Structured Data –organised format; arranged relations of tuples and attributes in data modelascertainingcardinality&degreesrespectively;convenientworkingfeatures like easy storage, retrieval, insert/update/delete, indexing, scalability, integrity constraints, transaction processing.

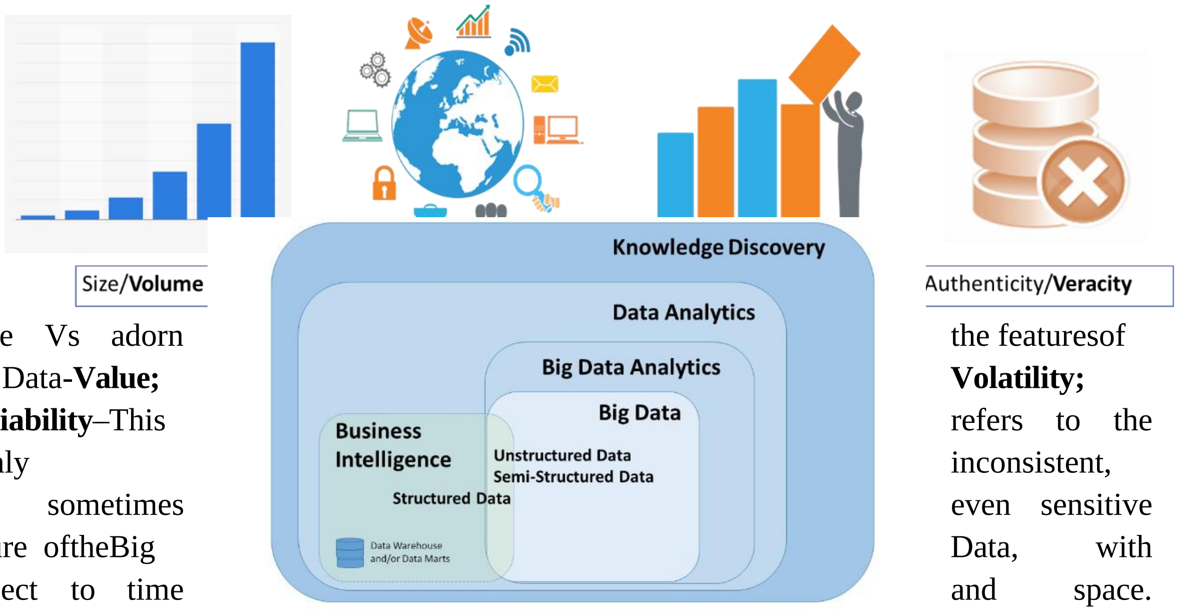
Much paradoxically to the pre-era Small Data, Big Data, is complex in the truest sense, in aspects i.e. multiple and unknown data sources, exploding volumes, data generation and processing speed, countless ways and purposes of data analysis, innumerable insights drawn from the data and several pros and cons associated with the tools, technologies and components used, defining the 4Vs jargon of its properties: Volume, Variety, Velocity, Veracity.

- **Volume** – humongous proportions from bits to bytes to currently zettabytes.
- **Variety** – various sources of origin owing to versatility of file formats andstructures.

Bits ☾ Bytes ☾ Kilobytes ☾ Megabytes ☾ Gigabytes ☾ Terabytes ☾ Petabytes ☾ Exabytes ☾ Zettabytes ☾ Yottabytes
Batch Processing ☾ Periodic Processing ☾ Near Real-Time Processing ☾ Real-Time (Online) Processing

- **Velocity** – turbo-speed data generation, super-fast data acquisition, provisioning, usage and processing, shifting paradigm from Batch Processing to Real-Time Processing.
- **Veracity**–authenticityofdatageneration&processingmeasuredinaccuracy, integrity, correctness and absence of biasedness, errors and noise in the data.

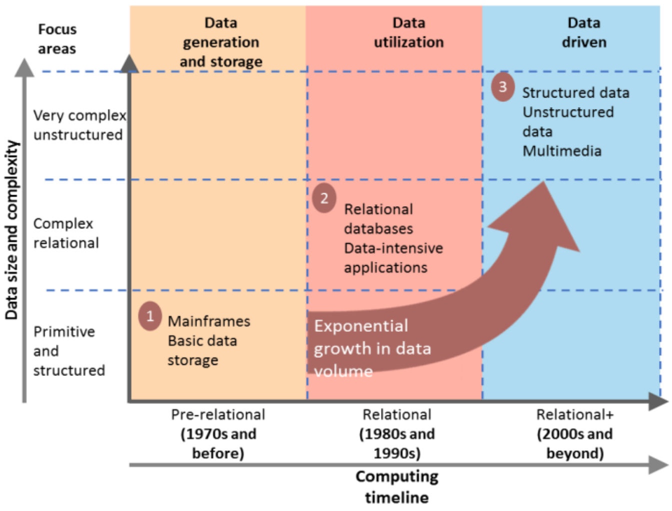
3



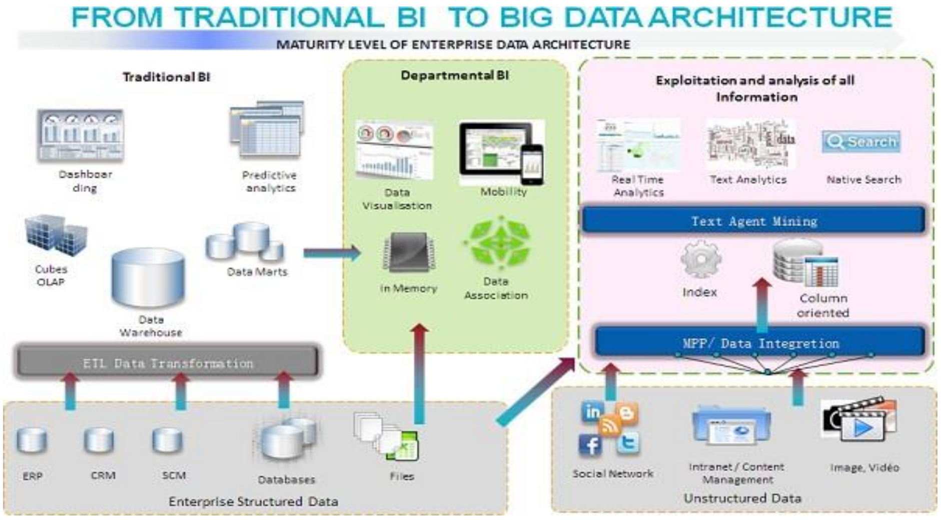
more Vs adorn the featuresof Big Data-**Value; Volatility; Variability**–This refers to the highly inconsistent, and sometimes even sensitive nature oftheBig Data, with respect to time and space. How best such features canbe utilized to carve the best out of available resources is seen further with the progression of this work.

LITERATURE REVIEW

In a nutshell, the advent of Big-Data, as grided to the right, **evokes the quintessential urge to explore more about the Big Data, along with Hadoop**, the most popular open-source framework/platform designed for Big Data Analytics; for **cost-effective, highly coherent capturing/ingestion, storage, processing, persistence, integration, visualization, analysis** of the Big Data, in order to **derive deeper insights about the data**, to yield **speedy and cognizant decisions**, bestowing **competitive advantage** oforganisations.

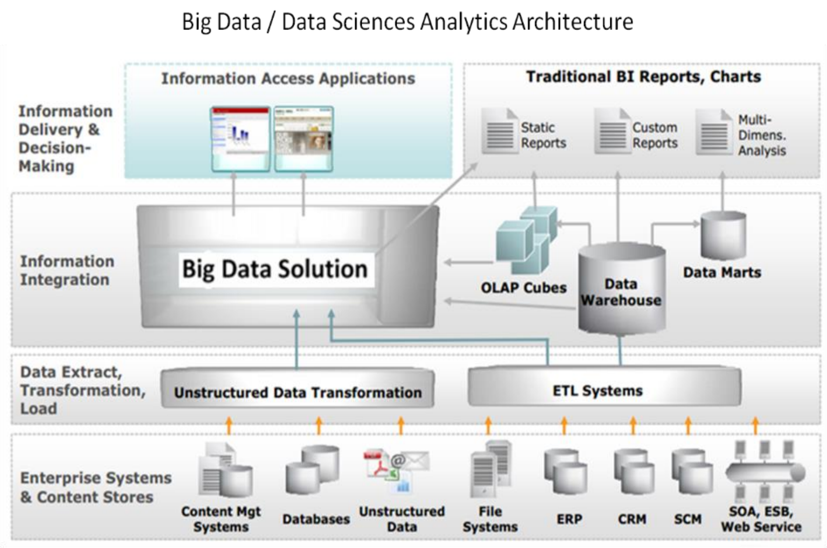

The Evolution of Big Data
Focus areas
Data generation and storage
Data utilization
Data driven
Data size and complexity
Very complex unstructured
Complex relational
Primitive and structured
1 Mainframes Basic data storage
2 Relational databases Data-intensive applications
3 Structured data Unstructured data Multimedia
Exponential growth in data volume
Pre-relational (1970s and before)
Relational (1980s and 1990s)
Relational+ (2000s and beyond)
Computing timeline

Big Data, being **manifold and various in framework and origin**, i.e. structured, semi-structured or unstructured, and being databases, data warehouses, multimedia data, social media, public web, log files, documents, application and sensor data, therefore, lays ground for **higher operational efficiencies, cost-effectiveness, potential time management, ideation and automation, optimized offerings and suggestions of new product lines/services**; consequently being an indispensable part of the business enterprises dealing with the enormous amount of data.



However, Big Data is **not a replacement for traditional practices** like Relational Databases, Data Warehouses, Data Marts, Data Lakes, and the more recent Business Intelligence practices along with OLAP (On-Line Analytical Processing) and OLTP (On-Line Transaction Processing) Systems. They rather **co-exist and cross-elevate capabilities** in terms of data processing and analysis, and data mining practices. Databases, Data Warehouses, Data Marts, Operational Data Stores (**ODS**) all serve as basic data storage systems, that retrieve, consolidate and store data obtained from homogeneous and/or heterogeneous sources leading to **Knowledge Discovery** processes of prudence. Such systems also can be further analysed using the **OLAP** sub-systems, contrary to the Relational Databases which are storage area of atomic day-to-day transactional data, arising the **OLTP** systems. The **Data Warehouses** and **Data Marts** differ in their area of coverage, with the Data Marts being pertinent to only particular business areas in an enterprise, whereas the Data Warehouse being an enterprise-wide storage entity. More recently,**Data Lakes** have emerged, which provide a 360^{o} view of the data being captured. Business Intelligence (**BI**) practices like Extract-Transform-Load(**ETL**), are used to exploit and analyse concerned data, by **aligning the BI capabilities with the enterprise business norms**, to improvise **process, bottom-line and beneficials**leading to **greater efficacy** to organizational goals and vision. The architecture and platforms associated, as depicted below, like Hadoop, facilitates the same.[8]

Big Data, is breaking dormancy in jet-speed, hence circumscribes the organisational daily work management through **Data Analytics** (descriptive, diagnostic, predictive and prescriptive), **Enterprise Resource Planning** (ERP), **Customer Relationship Management** (CRM), **Contract Management** (POS Documentation and Negotiation), **Developmental progression** (New Product Development, New Business Development, Sustainable Propagation Concepts); and also extending to the world of **Academics** (Research and Learning), particularly in **Data Science** and its augmented implementations.


Big Data / Data Sciences Analytics Architecture
Information Access Applications
Traditional BI Reports, Charts
Information Delivery & Decision-Making
Static Reports
Custom Reports
Multi-Dimens. Analysis
Information Integration
Big Data Solution
OLAP Cubes
Data Warehouse
Data Marts
Data Extract, Transformation, Load
Unstructured Data Transformation
ETL Systems
Enterprise Systems & Content Stores
Content Mgt Systems
Databases
Unstructured Data
File Systems
ERP
CRM
SCM
SOA, ESB, Web Service

Among the several **advantages** of Big Data, the following are noteworthy:

- **Data Locality**: It provides a coherent, cost-effective platform for huge data storage and processing, withthe special capability to *"move code to data"* instead of the traditional *"move data to code"* approach, whereby higher level of convenience is achieved by simply rendering the code to be applied on the data by moving it to the datalocation.
- **ReliabilityandFlexibility**:Itprovidesareliableandflexibletechnology-enabled data analytics environment.
- **Robust Support**: It supports high-speed data acquisition, processing and result generation.
- **Adaptive**: It leverages the potentiality of other systems, in terms of handlinghigh-volume, high-velocity and high-variety data.
- **Cross-Functionality**: It accentuates a tight handshaking mechanism and arrives at a comprehensible trade-off between Information Technology (IT), Business and Data Science.
- **BusinessIntelligenceManagement**:InBusinessEnterprises,itlaysthefoundations for easily collecting, organizing, summarizing, analysing, synthesising data and therefore carves the path for better decision making and profit-generation.

While such virtues bejewel the Big Data World, it is **yet not impeccable**, due to the following Roadblocks:

- **Security and Privacy Dilution:** pilferages amplified all the more due to Cloud Computing and Virtualization concepts.
- **Data-Retention Tenure:** Crux ofmatters, the voluminous and versatile data posing challenge towards retention duration and debatable data utility with time being on the forward.

- **Interpretability:** absentia of rigid schema or structure and even the heterogeneity in the data structure, makes processing it, a tedious job for the Big Dataapplications.
- **Efficacy:** Accommodation of scalability, continuous availability, partialtolerance, data consistency and avoidance of data redundancy, can be sometimes arduous in certain Big Data applications.
- **Visibility:** Data visualisation and error-handling techniques sometimes may tend tobe cumbersome in Big Data applications.
- **Resource availability:** Lastly, since the tools and technologies are in the emerging phase, and since this is a relatively new hype, obtaining skilled Big Data scientists and professionals are a bit tedious, as ofnow.

This Paper is authored, thereby, with the core objective of **inspecting the methodologies, technologies, properties, pros and cons, and associated jargons, relevant to Big Data, and its most popular implication, the Hadoop Ecosystem.** Besides, the paper also explores the possibilities of **combining and/or interchanging the various components and/or techniques intrinsic in Big Data**, eliciting the cons of Big Data. Further ahead, this work delves on the author's proposition, to build a system that inculcates, **Online Aggregation of Map Reduce Sub-System in Hadoop**, [13][10] overcoming the faults in of real-time online parallelization and pipelining of the Map Reduce functions. This perspective of redesigning the traditional Map Reduce sub-system, by incorporating distributed and parallel computation, **prolifers and elevates the abilities of the Hadoop Ecosystem**, pertaining to **scalability, fault tolerance, and online real-time processing speed**, interim to the Map Phase and Reduce Phase. Platformed on Hadoop Map Reduce, the Map Reduce supports Online Aggregation and stream processing, simultaneously **enhances utility and shrinks response time**. The work going ahead, focusses on Map Reduce, a **prime action enabler for satiating Big Data demands** pivoted on parallel processing exploiting numerousness of commodity nodes in the Hadoop Ecosystem for scalingvolumization.

TECHNOLOGIES AND METHODS

Big Data is associated with some technological and methodological jargons, which had helped in paving the way towards the advent of the Hadoop Ecosystem, and the methods and techniques used therein. Some such terminologies are stated below:

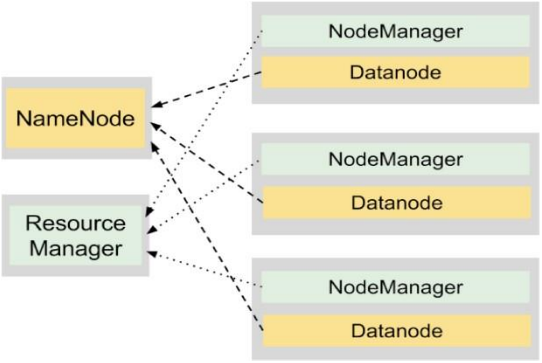
- ***In-Memory Analytics*** – storing all processing-relevant data in the Random Access Memory (RAM), before-handed, for faster access, and rapidprocessing.
- ***In-Database Analytics*** – It combines the data warehousing sub-systems withthe analytical sub-systems, through the process of Extract-Transform-Load (ETL), thereby eliminating the necessity to export and import datarepetitively.
- ***Symmetric Multi-Processor System (SMP)*** – A tightly-coupledmulti-processing system, with shared main memory between processors, with each processor having full access to all I/O devices, with their own high-speed memory, cache memory, and controlled by a single operating system instance.

- ***Massively Parallel Processing System (MPP)*** – A well-coordinated system ofparallel processors, each with dedicated memory and own operating system, and each processor working on a different part of the process or program, and communicating via message-passing protocols.
- ***Distributed Computing System*** – Unlike parallel systems which aretightly-coupled, distributed systems are loosely-coupled systems, composed in turn of individual sub-systems, each running their individual application and having the data distributed across several sub-systems. It follows a ***Shared-Nothing Architecture***, and as per the ***Brewer's (CAP) Theorem***, in a distributed environment, one can achieve only any two of **Consistency (C), Availability (A) and Partition Tolerance(P)**.

Hadoop Ecosystem, is a popular open-source project work of the **Apache Software Foundation**, written originally in Java, and based on **Google's Map Reduce and Google's File System (GFS)**. Doug Cutting, the developer of Hadoop, named it after his son's toy elephant, and henceforth the emblem of Hadoop is a yellow elephant. It was originally developed to support the text search engine, "Nutch".[9]

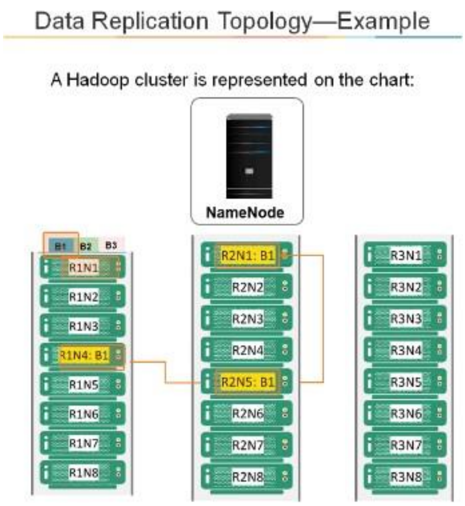
Hadoop acts as an **open-source distributed computing platform** for offering exactly **optimized solution for massive & disparate** Big Data utilizing **easily available and comparatively cheap hardware** components, **replicated *Nodally***, so as to render the key advantages of **fault tolerance and isolation, scalability, reliability, robustness, flexibility, high processing speed and availability.** There are majorly two version releases of Hadoop – Hadoop 1.0 and Hadoop 2.x.

Hadoop has two main sub-parts (detailed in following pages):

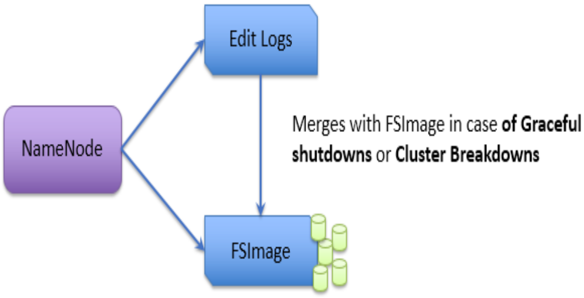


- ***Data StorageFramework:*** The *Hadoop Distributed File System (**HDFS**)* acts as the storage unit for the Hadoop Ecosystem, rendering itself to be a general-purpose file system to store the data in its native form.

  The HDFS in turn consists of **nodesor *daemons* working in *master-slave architecture***, the master node being the ***NameNode***and the slave nodes being the ***DataNodes***. The NameNode is responsible for managing the **file system namespace and metadata information**, and also **store and maintain the data block related information along with the block indexes**. DataNodes are the **work-horses** of the HDFS, acting as the slaves, and responsible for **storing the data in blocks**, and for performing the **read-write low-level operations**. The DataNodes sends **heartbeat signals** as well as block information reports to the NameNode for making the latter aware of their existence and their constituents.



The data in HDFS is replicated across different DataNodes, in blocks therein, with a default replication factor of 3. HDFS has a **well-defined Network Topology**, incorporated with a **Rack Awareness Algorithm** for optimal placement of data blocks and their replications, to facilitate data locality, reduce latency and provide fault resilience.


Edit Logs
NameNode
Merges with FSImage in case of Graceful shutdowns or Cluster Breakdowns
FSImage

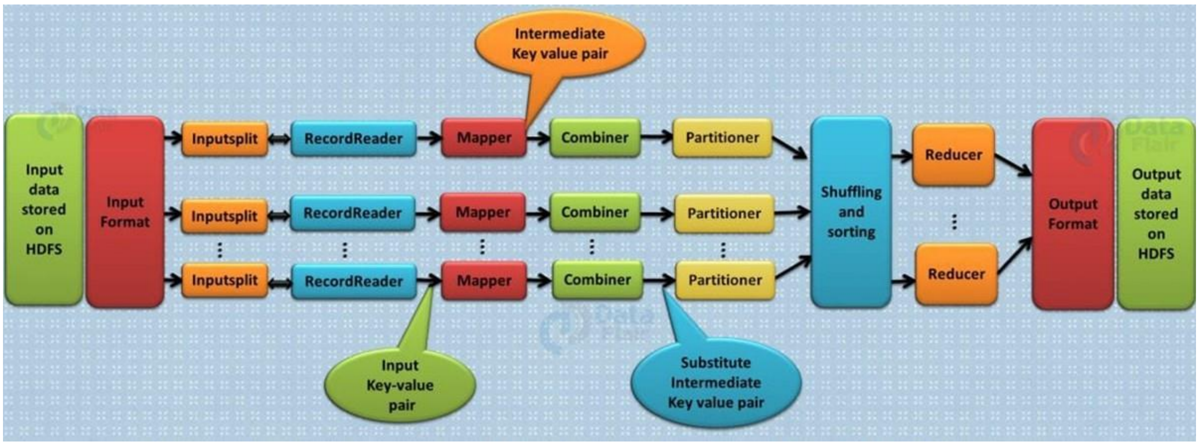
In HDFS NameNode, the metadata information is stored in ***editLogs***(containing all transactional data) and in ***FSImage***(containing the filesystem image, which is refreshed periodically via ***check-pointing***, and is used for NameNode**recovery** in case of cluster breakdown, a cluster being a collection ofnodes).

In Hadoop Version 1.0, another daemon existed, the ***Secondary NameNode***, which acted as a **pseudo-backup node for the NameNode**. However, it is not a true backup and henceforth, the NameNode acts as a **Single Point Of Failure (SPOF)**. Also, in an **enterprise-wide Big Data HDFS Solution**, it may so happen that a single NameNode**may not be able to engulf all the available data**. Both of these above stated issues have been taken care of in Hadoop Version 2.x, respectively, by means of *HDFS 2.x High Availability* (with a true backup available for the NameNode, working in Active- Standby Mode), and by *HDFS 2.x Federation*, by designing separate NameNodes pertinent to different subject areas.

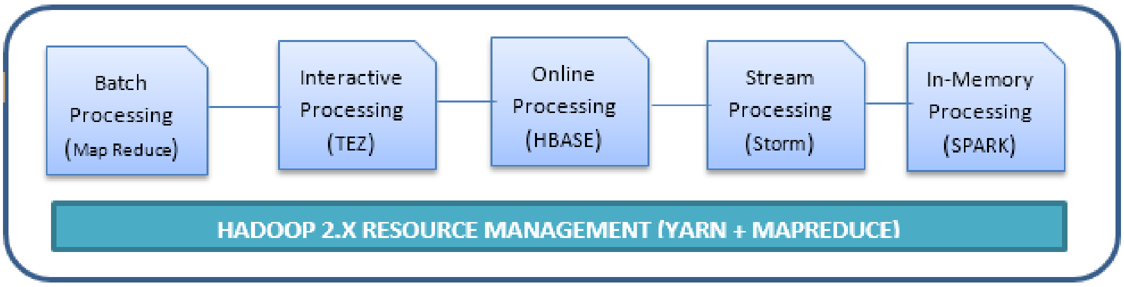
- ***Data Processing Framework*** : The *Map Reduce* and *Yet Another Resource Negotiator(**YARN**)*actsastheprocessingplatformforHadoop,allowingformassive data processing in parallel. The data to be processed in broken into smaller units, for achieving distributed computations, and faster processing speed. [4]

The processing tasks are in turn split into sets of ***Map* tasks and *Reduce* tasks**. Each task processes the small subset of data assigned to it. The Mapper tasks comprise– **loading, parsing, transforming and filtering** the data. The Reducer component takestheintermediate**Mapperoutputs**astheirrespectiveinputs,afterproper

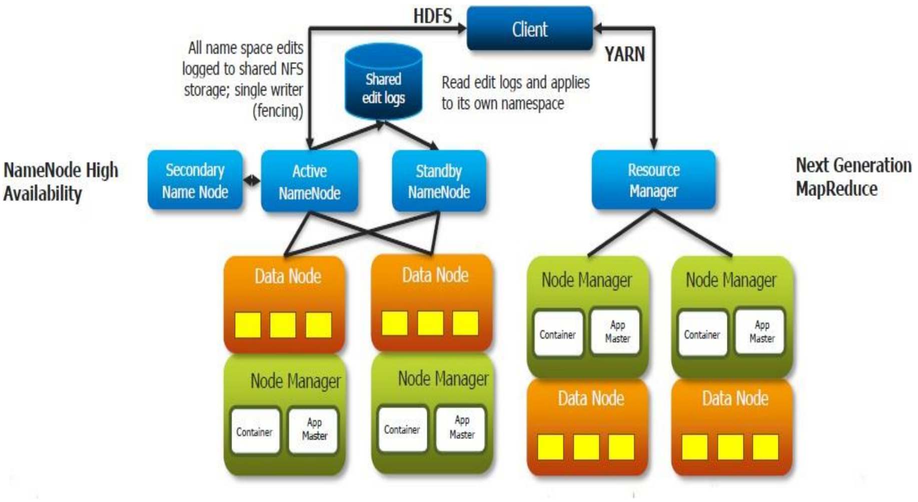
shuffling and sorting, and combines the former to generate reduced final output. The Mapper and Reducer components deal with key-value pairs, analogical to hash maps or Python dictionaries.[12][13]



In Hadoop 1.0, only MapReduce is present, working in a master-slave architecture, as in HDFS; with the ***JobTracker*****being the master daemon and the *****TaskTracker*****being the slave daemon**. However, in Hadoop 1.0, there were certain limitations like, availability of only **Batch mode of processing, rigidity in terms of programming language with the only available option being Java, and also the problem of an over-burdened JobTracker, acting as a Single Point Of Failure (SPOF).** Hadoop 2.x outgrows these limitations, by bringing into picture, another **sub-system**,the **YARN**, along with the MapReduce sub-system. YARN brought about **flexibility**in terms of **real-time processing**, **easy interfacing** with other applications and programming languages apart from Java. Also, YARN achieved **modularity**, by employing a master ***ResourceManager***daemon, consisting of further sub-components to deal with the different functionalities of the version 1.0 JobTracker (Scheduler, Application Manager), a slave *NodeManager*daemon taking up the responsibilities of version 1.0 TaskTracker, and a per-application-specific ***ApplicationMaster***, for application and program-level flexibility and interfacing. In Hadoop 2.x, due to presence of YARN and some additional applications and components, apart from the Batch Processing of MapReduce, one can achieve the below-mentioned different modes of processing :[3]

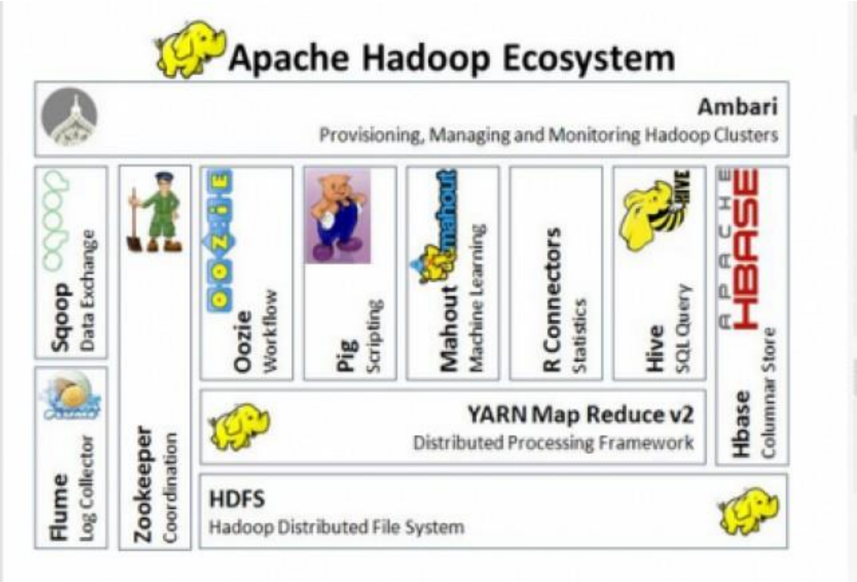

Batch Processing (Map Reduce)
Interactive Processing (TEZ)
Online Processing (HBASE)
Stream Processing (Storm)
In-Memory Processing (SPARK)
HADOOP 2.X RESOURCE MANAGEMENT (YARN + MAPREDUCE)

Also, it should be noted that, MapReduce and YARN sub-system and HDFS sub-system run on the same sets of nodes, henceforth, such configuration facilitating effective assignment of tasks/jobs on generally the nodes where the data is present, a feature known as Data Locality, as depicted aside.



Apart from the above main two sub-systems of the Hadoop Ecosystem: the HDFS and the MapReduce-YARN, there is are **additional suite of supporting components**, which help Hadoop enhance its functionalities, by adding further dimensions to it, as described in a jiffy below :[1]

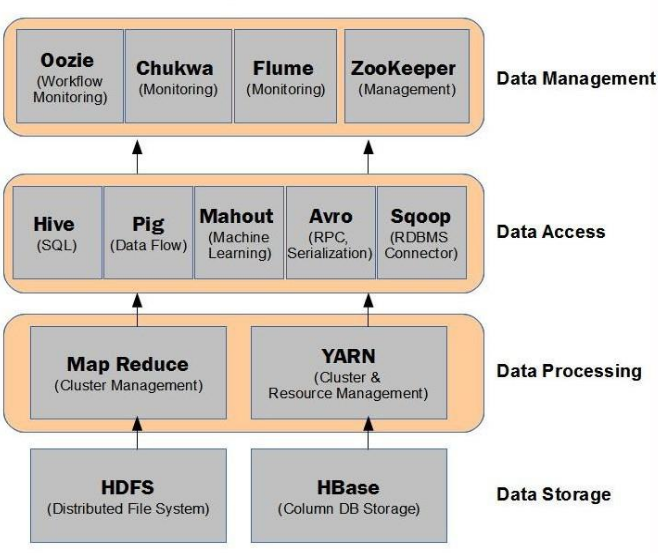
- ***HBase***– The official non-relational, open-source distributed database for Hadoop, which stores structured data in column-oriented manner, in terms of key-valuepairs.
- ***Pig*** – A high-level programming environment, which uses Pig Latin as thescripting language, and is used for depiction of high-level data flow. Pig Scripts are automatically converted to equivalent Map Reduce programs.
- ***Hive***–AdatawarehouseframeworkforHadoop,whichinturnusesSQL-likeQuery Language, Hive Query Language (HQL), and is used mainly for analytical purposes.
- ***Sqoop***– It is used for bi-directional data transfer between HDFS and/or HBaseand structured relational database systems, acting as Source and Sink Systems.
- ***FlumeandFlumeNG***–Itisusedfortransferringstreamingdata,suchaseventdata, sensor data, log file data from sources to a central location like HDFS/HBase. Flume New Generation (NG) is an improvised version, acting as a real-time loader for streaming data.
- ***Oozie***–Itactsastheworkflowschedulersystem,withprovisioningforbuilt-indecision control and branching in the workflows.
- ***Fuse/Samba***–BothserveasHDFSmountingsystems,wherebyHDFScanbehaveasa regu;ar file system, with the ability of using regular Linux/Unix commands for file processing.
- ***Chukwa***–Itisadatacollectionsystem,formanaginglargedistributedsystems,and monitoring purposes.
- ***Mahout***–Itisascalablemachinelearninganddatamininglibrarycomponent,usedfor predictive analytics and other advanced analytics practices.
- ***Whirr*** – It is the Clod Computing provision or add-on for the HadoopEcosystem.
- ***Storm*** – It is a distributed, fault-tolerant, high performance real-time and/or streamingdata processing system.



- ***SPARK*** – It is an open-source,parallel data processing system, having provisions for both in-memory and on-disk processing, and combining the aspects of batch, streaming and inter-active analaytics.
- ***Avro*** – ARemote Procedure Call (RPC) and Data Serialization-Deserialization framework, forHadoop.
- ***NoSQL***–NoSQLorNotOnlySQLisalight-weight,non-relational,open-source database, that does not conform to the normal ANSI-SQL constructs.
- ***Cassandra*** – It is a NoSQL database, using CRUD (Create, Read, Update, Delete) operations, in terms of cqlshscripting.
- ***CouchDB***– It is a document-oriented, open-source NoSQL database. It is implementedin the concurrency-oriented language, Erlang, and uses JSON to store data, and JavaScript as its formal query language.
- ***BigSQL***– It is a highly distributed, SQL-on-Hadoopengine.
- ***JaQL***– A high-level query language designed for use with datain JSON (Java Script Object Notation) format, and similar semi-structured data; resulting in low-level Map Reduce jobs.
- ***MongoDB*** – It is also a NoSQL database platform, that uses BSON (Binary JSON),acting as an open-standard for complex semi-structured data. It is well-equipped with a rich query language base, along with a fast in-place updating technique in place.
- ***Lucene/Solr***–High-performing,full-featuredtextsearchenginelibrary,providingfast indexing and ranked data search.
- ***Hue***–HueprovidesabrowserbasedGraphicalUserInterface(GUI)forHDFSandother Hadoop components like Pig, Hive etc.
- ***Ambari***–Itisaweb-basedtool,forprovisioning,managingandmonitoringHadoop clusters.
- ***ZooKeeper***– It acts as a centralized coordination service, formaintaining configuration information, providing distributed synchronization and group services.
- ***Apache DRILL*** – A distributed system for interactive analytics, oflarge-scale datasets, inspired by Google’s Dremel.
- ***UIMA***–UnstructuredInformationManagementArchitecture(UIMA)isanopen-source platform, from IBM, used for real-time content analysis, text and unstructured data processing to unearth the latent relationship buried therein.
- ***TEZ***–Itisaneffortfordevelopingagenericapplicationframework,forprocessing complex datasets relatively faster, and also for providing a set of re-usable data processing primitives that can be used by other Projects.

- ***THRIFT*** – Cross-language data serializationframework.

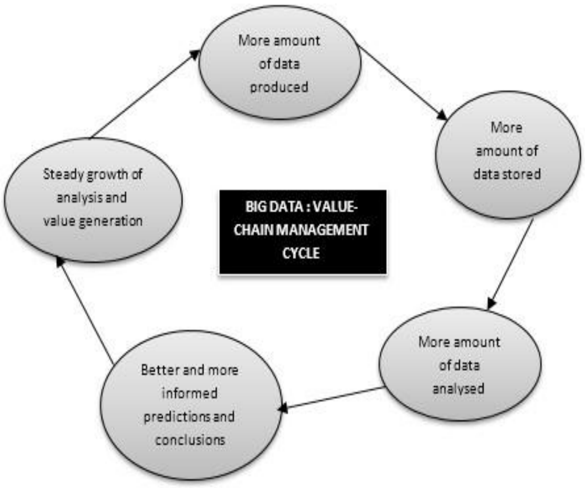
For **index web searches, spam email detection, recommender systems, financial services and strategies forecast, biological genomic manipulation and for analysis of unstructured data such as log, text and clickstream**, Hadoop is being leveraged extensively, although many are implementable in a relational database(RDBMS) as well. We may note that the core of the Hadoop framework is operationally varied from RDBMS. For e.g. **Complex information processing** (Modification of Unstructured data into Structured data); **Complex but parallelizable algorithms needed in place of heavily recursive query processing**; **Cost Complexity Mitigation** by resolving **Spacio-Temporal issues of processedData**; **Job Scheduling-Significant Custom Coding-Fault Tolerance** armed optimal resource.[7]



The Hadoop Ecosystem along with one or more of the additional components, have been consolidated in **convenient distribution software packages**, by different companies like ***Cloudera, IBM InfoSphere, HortonWorks, Apache, MapR***and several others. Applications of Hadoop are spread across leading world-wide companies like **Google, Twitter, Facebook, Yahoo,** among many others. The supported operating systems for Hadoop Ecosystem, are **Windows, Linux, BSD and OS X**.

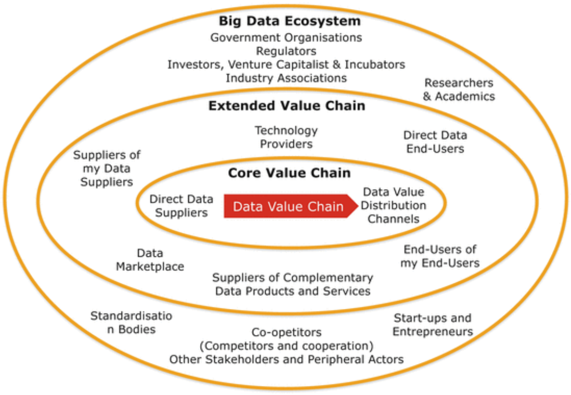
EXPERIMENTAL ANALYSIS

Big Data Tools and Techniques unwraps to us an entirely new perspective of handling gigantic volumes of data, and thereby probe further into data processing, to conclude valuable insights, directly and indirectly from data under inspection. This entire process of converting the enormous quantity of data, to valuable information, can be broadly classified into the following sub-phases, the entire process being referred to as ***"Big Data Value-Chain Management"***:[2]


More amount of data produced
More amount of data stored
More amount of data analysed
Better and more informed predictions and conclusions
Steady growth of analysis and value generation
BIG DATA : VALUE-CHAIN MANAGEMENT CYCLE

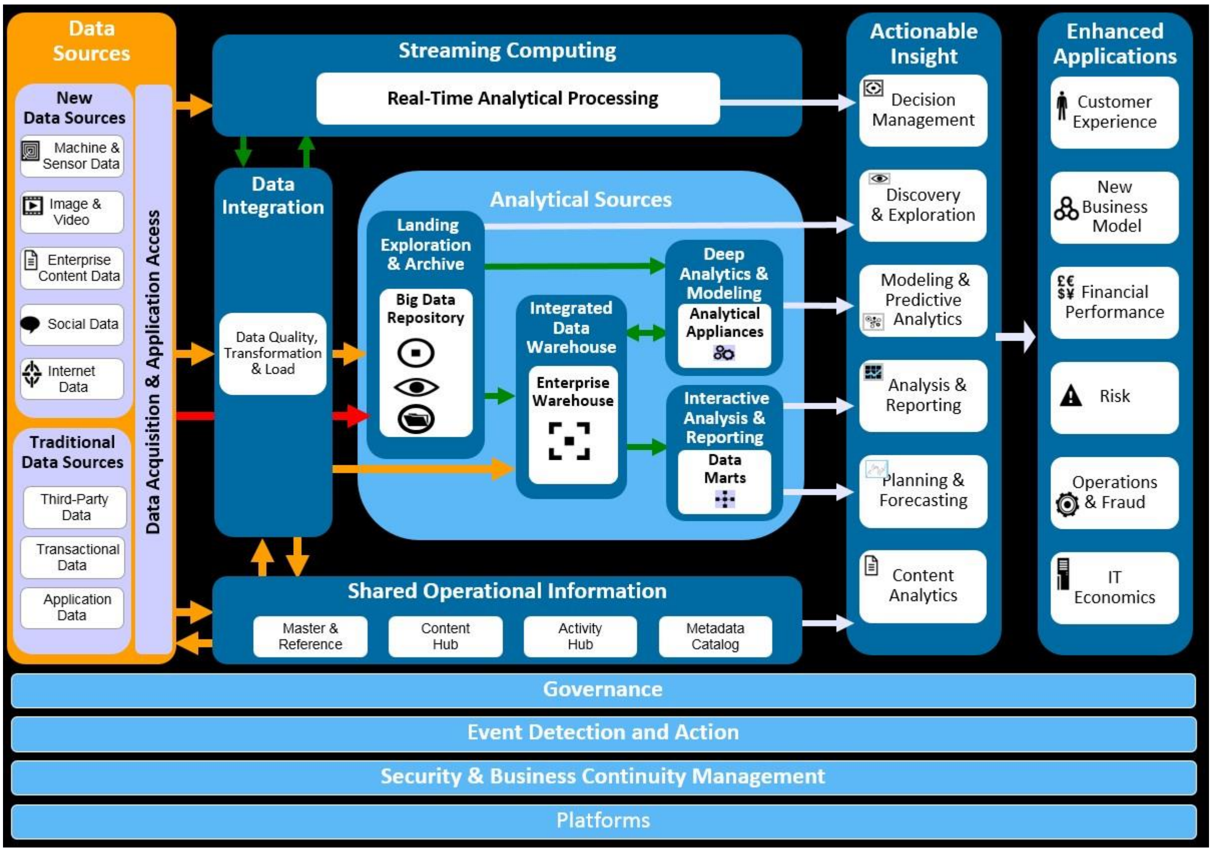
1. ***Data Generation*** – This is the first step, whereby colossal amounts of data are generated from profusely diverse sources like databases, data warehouses,

HTML/XML/JSON, Web Pages, Audit and Log Files, Multimedia Data like image, audio, video, sensor data, documents, spreadsheets, social media, email, text messages and chats and several other such sources. [9]


Big Data Ecosystem
Government Organisations
Regulators
Investors, Venture Capitalist & Incubators
Industry Associations
Researchers & Academics
Extended Value Chain
Technology Providers
Direct Data End-Users
Suppliers of my Data Suppliers
Core Value Chain
Direct Data Suppliers
Data Value Chain
Data Value Distribution Channels
Data Marketplace
End-Users of my End-Users
Suppliers of Complementary Data Products and Services
Standardisatio n Bodies
Start-ups and Entrepreneurs
Co-opetitors (Competitors and cooperation)
Other Stakeholders and Peripheral Actors

2. ***Data Acquisition/Collection*** – Encapturing of generated data for subsequentprocessing.
3. ***Data Pre-Processingand Organization*** – Laundering and pre-processing transformation of data to be inter-operable and compatible with the target systems or functions that need to be applied on the data. Thisprocessisalso,sometimes, known as Data Curation, and it involves the very basic checks for Data Validity and Data Quality.
4. ***Data Storage*** – It deals with the proper, persistent and distributed storage of thelarge-scale data, so acquired, and can be stored in relational as well as non-relational databases, in-memory databases, cloud-based virtualised storage media; adhering to the basic principles of data security, consistency, accuracy, scalability, partition and fault tolerance, continuous availability and data standardization.
5. ***DataIntegrationandSummarisation***–Sometimes,thedatatobestoredforfurther analysis, may need to be aggregated, consolidated and/or integrated, to deliver only meaningful summarised information, thus avoiding highly detailed, low-level data, in certain cases.
6. ***Data Processing and Analysis*** – Computation & Algorithmic implementationand model building of pre-treated data, typically by "moving code to data" as per the Big Data paradigm. Data processing aims to unveil latent patterns, trends, associations and or correlations and dependencies in the data, which helps in supplemental analysis. Analysis on data can range from simple descriptive analytics, to diagnostic analytics, and furthermore advanced levels of predictive analytics and prescriptive analytics, and can involve data mining practices like regression, classification, association mining, collaborative filter-based recommendations, text analysis, survival analysis, social media analysis, sentimentanalysis.
7. ***Deduction of Insights*** – The patterns so deduced from the data analysis results, aidin summarizing conclusions about any process/system/facet or the datum itself, so as to further utilize the knowledge gained in adorning the organisational or research value chain, for analytical reporting, data visualisation, strategizing, decision making and or other knowledge representation purposes.
8. ***Conclusive Decision Making*** –Decision making, the key of paramount significanceto any organisation for harvesting the best, is enhanced by and based on aforementioned cascade, particularly, when the acumen is toiled to beget the better and leaping forward the best alternative out of the ones available at every juncture. It enables empowering the best course of action and also risk –taking, as and when required. Better governance to any business is thus ensure en-route this channel. [5]

Hence, the Big Data Value Chain Management categorises the various activities, necessary to enable effortless knowledge discovery from the disparate, ginormous Big Data at hand,which otherwise would fail to yield any value.[11]



CONCLUSIONS

In this paper we have scrutinized various technologies to tackle the big data and their architectures. We have also discussed the challenges of Big data (volume, variety, velocity, value, veracity) and various advantages and a disadvantage of these technologies alongside the architecture and ecosystem for distributed data processing over a cluster of commodity servers. The main goal of our paper was to engage in a survey of various big data handling techniques which cradle humongous amount of data every moment, commensurate to the framework of any entity at its root.

REFERENCES

1. V. K. Jain (Khanna Publishing), ***Big Data & Hadoop***, 1st Edition, Pages 3-11,2017.

2. Curry E.,***The Big Data Value Chain: Definitions, Concepts, and TheoreticalApproaches***, In: Cavanillas J., Curry E., Wahlster W. (eds) New Horizons for a Data-Driven Economy.Springer, Cham, 2016.Lakshmi et al., International Journal of Advanced Research in Computer Science and Software Engineering 6(8), p. 368-381, August- 2016.

3. Tom White (O'Reily), ***Hadoop The Definitive Guide***, 4th Edition, Pages – 69-74, 80,2015.

4. AmoghPramod Kulkarni, Mahesh Khandewal, ***Survey on Hadoop and Introduction to YARN***, International Journal of Emerging Technology and Advanced Engineering, ISSN 2250-2459, Volume 4, Issue 5, May2014.

5. Ms. VibhavariChavan, Prof. Rajesh. N. Phursule, ***Survey Paper On Big Data***, International Journal of Computer Science and Information Technologies, Vol. 5 (6),2014.

6. MrigankMridul, AkashdeepKhajuria, Snehasish Dutta, Kumar N, ***Analysis of Big Data using Apache Hadoop and Map Reduce***, Volume 4, Issue 5, May2014.

7. Suman Arora, Dr. MadhuGoel, ***Survey Paper on Scheduling in Hadoop***, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 5, May2014.

8. Yuri Demchenko, ***The Big Data Architecture Framework (BDAF)***, Outcome of the Brainstorming Session at the University of Amsterdam, 17 July2013.

9. TekinerF. and Keane J.A., ***Systems, Man and Cybernetics (SMC), Big Data Framework***, IEEE International Conference, 1494–1499,2013.

10. Sagiroglu, S. Sinanc, D, ***Big Data: A Review***, 20-24,2013.

11. Dong, X.L.; Srivastava, D. Data Engineering (ICDE), ***Big data integration***, IEEE International Conference on, 1245–1248,29(2013).

12. Kyuseok Shim, ***MapReduce Algorithms for Big Data Analysis***, DNIS 2013, LNCS 7813, pp. 44–48, 2013.

13. Jimmy Lin, ***Map Reduce Is Good Enough? The control project***, IEEE Computer 32,2013.

14. Aditya B. Patel, Manashvi Birla and Ushma Nair, ***Addressing Big Data Problem Using Hadoop and Map Reduce***, in Proc. Nirma University International Conference On Engineering, 2012.

15. Margaret Rouse, ***Unstructured data***, April2010