Available online @ https://jjem.jnnce.ac.in
https:www.doi.org/10.37314/JJEM.2022.060106
Indexed in International Scientific Indexing (ISI)
Impact factor: 1.395 for 2021-22
Published on: 31January 2022

Literature Survey on Big-Data Analytics and Tools

Aruna Kumar P[1]*, Pavan Kumar M P[2], Suchethana H C[3]

Department of Information Science and Engineering,
JNN College of Engineering, Shivamogga

arunkumarp@jnnce.ac.in, pavankumarmp@jnnce.ac.in, suchethanahc@jnnce.ac.in

Abstract

In the present technological world, a huge quantity of heterogeneous and complex data have been generated from various sources such as social media applications, sensors, IoT devices, Smartphone's, demographic data, business research and scientific data, worksheets, blogs, forums, E-commerce websites, search engines etc., These drastic data are more useful in one another way in our business or social activities to make some decisions. Thus, to get insight into data to find desired pattern it is necessary to make empirical analysis on the data. In this article, we have made a comprehensive literature that briefed the characteristics of big data, tools and techniques which are avail for processing heterogeneous-data and HDFS file systems to store the big-data.

Keywords: *Hadoop ecosystem, Big-data Analytics, MapReduce, HDFS.*

1. Introduction

Big-data refers very large, heterogeneous and schema less data. Big-data [8] is in unstructured or semi-structured form, so it is too complex to process using traditional data processing application software. Nowadays, data have been collecting from various sources, collected data need huge scalable storage. To fetch useful information from stored big-data, advanced analytical tools are required. In this paper, an attempt is made to explore the concepts of big-data and many analytical tools.

Nowadays mobile and internet users exchange or share information through social media like Twitter, Instagram, Research Gate, Telegram WhatsApp, Facebook, LinkedIN etc., which results big quantity of data have been generated in every second. Along with sensors and automated devices located in various places, various applications are produces a large quantity of data in the form of semi-structured or unstructured, called big-data. It is hard to store, process and analyse big-data by using traditional RDBMS applications [8, 12]. Thus, advanced data storage and processing tools are required in order to process the big-data.

In big-data domain, the structured data is well organized and such data defined with proper length and format. The data in the form of rows and columns like excel sheet are example for structured data. All traditional applications are store and process structured data. Structured data are schema based data. Tabular data which contains attributes as column names, all attributes defined with data types [1, 14] constraints and size etc. In case of semi-structured data [14], data are little bit organized, refers data is not conform formal structure. Log files and emails are example for semi-structured data. However, unstructured data [14] are not well organized, refers data is not able store in traditional relational database using row and column format. For example Textual data (Texts), multimedia data (Photos, Audio and Video).

1.1 Characteristics of Big-Data

The big-data characteristics [1-2, 6, 9-10, 12, 14, 16] are refined over a period of time due to its dynamic nature. Here, we have listed and described 8 characteristics popularly mark as 8V's as shown in Figure 1, which are Volume, Velocity, Variety, Veracity, Values, Validity, Volatility and Visualization [5,16].

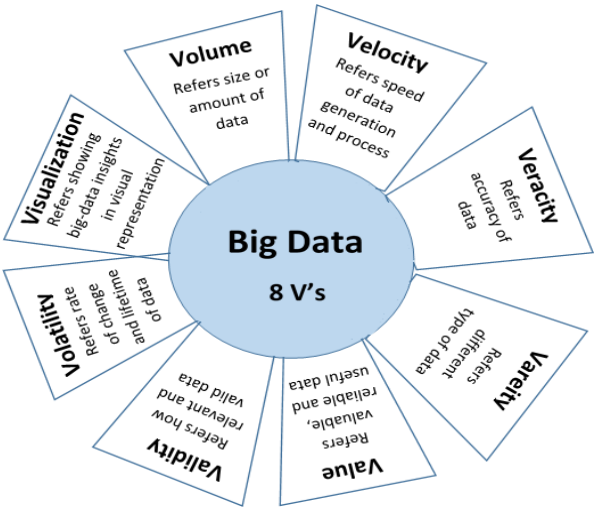

Figure.1: Characteristics of Big-Data.

Volume: The volume refers size of the data. Big-data is very huge in size and grows continuously. Every second the devices like mobile, computers, sensors etc., and software applications like Instagram, Telegram, Facebook etc., are generating data huge amount of data.

Velocity: Speed of data generation is another important characteristic, refers rate of data generation or how fast data grows. Devices and applications are producing enamours amount of data every fraction of second.

Veracity: How much accurate data is produced by devices or applications? Are there unreliable, inconsistent and uncertain data? Some time, devices produce data with noise, low quality which results inaccurate data. Also applications may generate invalid, unwanted data.

Variety: Indicates, different types of data. Data been generated from various devices which are heterogeneous in nature. Which will produce data may be in unstructured, semi-structured and structured format.

Value: Specify, the desired values or insight into the data to fetch desired pattern, which is useful to take some decision. The entire data which is collected may not be useful, need to dig to fetch required and useful information.

Validity: It determines, at what extent data is valuable and relevant?

Volatility: Find out how long the collected data is valid?

Visualization: Final insights are required to represent visually to the end users through graphics like bar charts, graphs etc.[16].

1.2 Big-Data Analytics

It is the way [2, 14] of finding the important and desired patterns in collected big-data by using advanced techniques and modern tools [14]. The process of collecting data from various sources, organize in suitable manner and making analysis on them to discover useful information and extract the desired patterns is called big-data analytics. Which will help to make better decision, risk management, better product development and improve customer experience in various sector. This process termed as knowledge discovery in databases (KDD) [4, 14] means discovering the useful knowledge from the database for decision making. The following figure 2 gives steps are taken in the process of knowledge discovery.

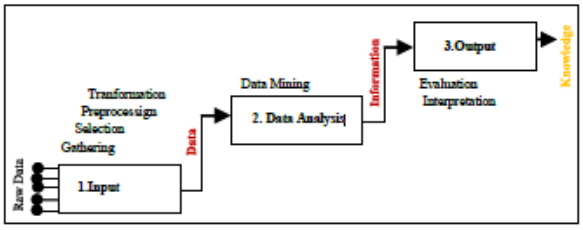

Figure 2: Phases of Knowledge Discovery in Database. (Courtesy: Chun-Wei Tsai et. al., [14]).

Life cycle of big-data analytics involves various stages as stated in the following section. Analysis of big-data from conventional data

analysis is not so easy, due to its complex nature and characteristics. So, its required well defined methodologies to analyse the big-data, following steps are required to sort out the actions and responsibilities involved with acquire, process, analyse the data.

There are nine stages are involved in big-data analytics lifecycle as shown in figure 3 and detailed explanations are clearly stated.

1. Business Case Evaluation
2. Data Identification
3. Data Acquisition and Filtering
4. Data Extraction
5. Data Validation and Cleansing
6. Data Aggregation and Representation
7. Data Analysis
8. Data Visualization
9. Utilization of Analysis Results

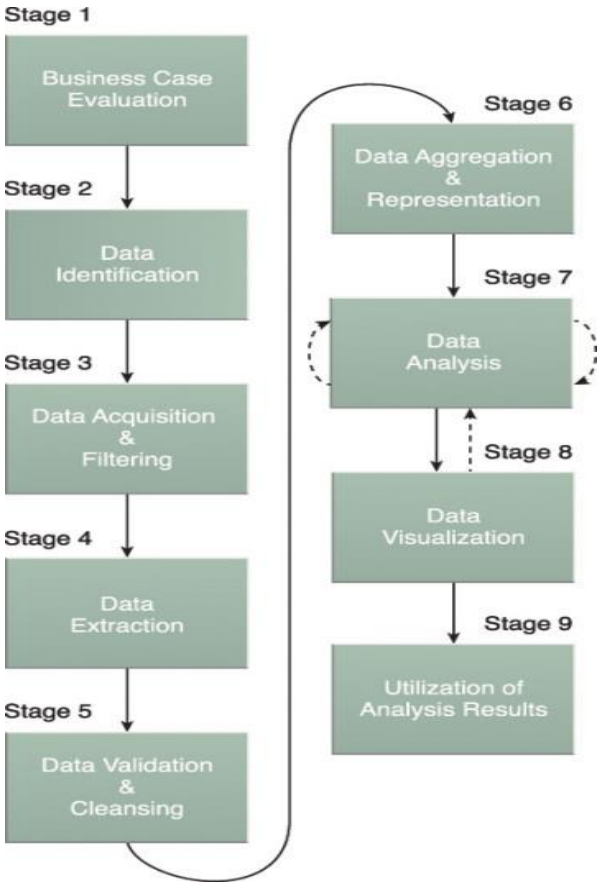

Figure 3: Phases in Big-data analytics lifecycle (Courtesy: Internet source as in below URL https://www.informit.com/articles/article.aspx?p=2473128&seqNum=11).

Business Case Evaluation is the first stage of big-data analytics lifecycle, must start with proper case study or problem statement that introduce a clear perceptive of the problem, motivation and goals of carrying out the data analysis. The detailed assessment of the problem statement helpful for decision-makers to understand the business resources and requirements.

The Data Identification is the step, which enforce to identify the data required for the analysis. Data identification helps to finding the desired pattern and correlations from available data.

Data Acquisition and Filtering phase will collect the data from various sources, which may contain noisy, corrupted and unwanted values. So, it's essential to filter the data. Hence, filtering plays vital role in removing corrupted and invalid values and will fill the missing values.

Data Extraction Data collected from various sources may not be in proper format, it is essential to transform it into proper format which is required to make analysis on the data. The data extraction phase will transform data into required format.

Data Validation and Cleansing is the process of identifying the invalid data that leads to wrong result and data cleansing phase responsible to validate the data by applying rules and removing invalid data.

Data Aggregation and Representation data may be scattered across many datasets, it is required to combine datasets together. Hence, this phase integrate datasets together which are scattered in multiple dimensions.

Data Analysis this phase is examining the acquired data and analyse the aggregated data. This phase is an iterative till to obtain the required pattern is found and its illustration are clearly demonstrated in figure 3.

Data Visualization Process this phase states

that, analysed data or results have been expressed graphically like bar charts, pie charts, graphs etc.,. Which will helps users to understand easily the data insights. Thus, big-data requires advanced data visualization tools.

Utilization of Analysis Results Phase, finally analysed results or values will be used by end user for making decisions in their business activities.

In this paper, at section-1, we have clearly stated about the brief introduction of big-data and its characteristics, importance of big-data analytics and lifecycle. In section-2, the detailed literature survey is stated, which includes 15 research articles, who made significant contribution into the Big-data and analytics field during 2010 to 2021. In section 3, detailed description of big-data tools and technologies are figured out. Finally, discussion of topic is concluded.

1. Related Works

Ahmed Elragal et. al., [1] presented the big-data analytics: A literature review analysis, in this paper, authors focussed on detailed discussion on characteristics of big-data, process of text mining, technical algorithms, processing, cloud computing, opportunities and challenges, reveal the research opportunities' in the data analytics research domain, and the work also given the hint to future research scope, article also mentioned the cutting edge research gaps that could motivate the vendors, researchers, and practitioners to broaden the research work on big data analytics tasks.

N. Elgendy et. al., [2] presented the article entitled on big-data analytics: A literature review paper, in this paper authors briefed about different analytics methods and tools, storage and management of big-data. The article also states the research scope provided by the application of data analytics in various decisions making domains are explained.

Sukhpreet Singh et. al., [3], presented an article on A review paper on big-data and hadoop, in this literature, authors figured out about basic characteristics of big-data, layered architecture of big-data framework, various problems with big-data processing, Hadoop and map-reduce architectures are explained.

D. P. Acharjya et. al., [4] published an article entitled A survey on big-data analytics: challenges, open research issues and tools, in this literature, they explored big-data future scope, and diversified tools associated with it. Thus, they also stated the researcher's constraints in big-data.

Ritu Ratra et. al., [5] introduced the big-data tools and techniques: A roadmap for predictive analytics, in this article, characteristics of big-data is demonstrated, and different mining techniques are clearly presented. The article also covered the usage of big-data analytical tools and hadoop architecture outcomes.

Umang Kumar et. al., [6] presented an article entitled big-data analytics: A literature review paper, in this paper, authors presented the dis-similar analytics methods and tools for data analytics. They also review the map-reduce architecture and HDFS file system techniques. In this article, they list out the applications of big-data analytics in different decision making domain.

A comprehensive review of tools and techniques for big-data analytics is presented by Amita Dhankhar et. al., [7].The article, brief the big-data concepts and highlighted the big-data systematic structure. The author explore the significant data assessment approaches and presented the acquisition tools advantages and disadvantages, and NoSQL storage systems working principle has been highlighted.

Dr. K. Venkatachalapathy et. al., [8] presented the research article entitled data mining for big-data: Issue and challenges. This article has been brief about data mining process, classification of data mining and cluster based mining methods in detail.

M.Ann Bency [9] published an article in the name of analytics methods and tools on big-data. This paper focused on advanced data analytical tools and techniques. The article has been figured out map-reduce and HDFS architecture and functionalities.

Pranav B et. al., [10] presented a review paper on big-data analytics. The advanced research work on big-data and its characteristics, diversified tools to analyze and store the result of big-data. They also assess the architectures of HDFS and map-reduce techniques to store and process obtained result.

Cheng Fan et. al.,[11] have presented an article regarding an application of data science techniques in construction of the efficient modern buildings, so that performance of the new buildings are enhanced based on analysis made on data collected during the design and construction. The data driven and IoT based approach gives the modern techniques to construct smart and efficient buildings.

Shubham Tripathi [12] presented a review paper on big-data analytics. This article reviewed about characteristics of big-data, analytical methods, decision making methods and analytical issues, challenges in big-data analytics process.

Rafi. SK, et. al., [13] presented big-data analytic processing, in this research article discussed regarding development of big-data strategy, big-data analytical processing and storage methods using map-reduce and HDFS respectively and big-data stream processing platforms.

Chun-Wei Tsai et. al.,[14] In this research article authors are precisely stated that expected trend of marketing of big-data between 2012-2018, data mining algorithms and efficient data analytics methods for data mining and frameworks for big-data analysis.

Nawsher Khan et. al., [15] In this survey, attributes of the big-data (characteristics, nature, growth rate, management, analysis and security) are clearly classified. The study gives detailed description of data life cycle and of big-data terminologies. Future research scope in big-data domain. Aforementioned research issues and challenges motivate the research scholars to develop the optimal techniques to address challenging issues in the domain of big-data.

Zhwan M. Khalid et. al., [16] In this article authors primarily provided reviews on need of data visualization and available tools and techniques. Later discussed about heterogeneous distributed storage description and their challenges using different methods. Kumar et al., [17] presented the importance of processing diversified data in High performance computing environment (HPC) [18-20] and their constraints and challenges are clearly stated.

Varsha M et al., [21] give a more detailed description of rice blast disease data. This work considers the historical big data of paddy disease covering period of ten years.

3. Analytical Tools and Techniques

Big-data analysis includes various phases of activities range from data acquisition to data visualization. Main components of big-data process are Data Ingestion, Data Storage, Data Analysis, Data Consumption, which are involved in the process of collecting, storing, analysing and consuming data respectively, detailed discussion is mentioned below.

i) Data Ingestion: is the process of congregation and preparing the data. This phases involves ETL process. The ETL (extract, transform, and load) steps are used to prepare the data. In this step, it's required to identify possible data sources and organize it through cleansing and validating. This phase empirically organize the data and optimizing the transformation process.

ii) Data Storage: the gathered data need to store for further process. It is essential and helpful to store the gathered data in data warehouse

or data lake efficiently. The storage of big-data must offer reliable, robust and available storage.

iii) Data Analysis: most important phase in big-data processing is analysis phase and it will analyse the data for fetching desired patterns or knowledge or values of interest. Analysing the data to generate valuable insights from the big-data to make better decisions. There are four kinds of big data analytics: prescriptive, predictive, descriptive, and diagnostic. In this paper, we have tried to explore tools and techniques available for analysis of big-data.

vi) Data Consumption: once the analysis of data is completed, results are shared with stake holders to make better decisions in their business activities. In order to furnish the better data visualization and storytelling to non-technical stake holders and project managers, advanced data visualization tools are used.

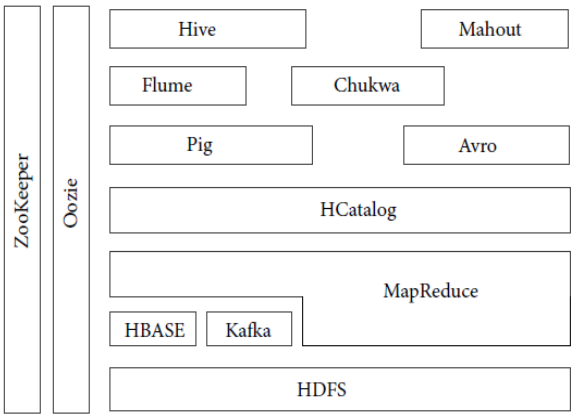

Figure 4: Hadoop ecosystem. (Courtesy [Nawsher Khan et. al.,15]).

Big-data management system includes many tools and techniques as shown in Figure 4 Hadoop ecosystem, each and every tools are responsible for handling big-data management like fetching, cleaning, transforming, storage, visualization etc,. The first and foremost tool is Hadoop Distributed File System (HDFS) [2, 6, 9] which run on commodity hardware and it is extremely fault tolerant and deliberate using low-priced hardware.

HDFS able to store huge amount of data and affords easy accessing. To store large amount of data, the files have been stored across multiple machines in the redundant fashion to make data available in case of failure and supports processing applications in parallel manner.

Some of the significant features of HDFS [10, 13] are given as below

- HDFS is an open source framework suitable for the distributed storage file system, which enable with features of distributed clusters computing model with data locality.
- Hadoop is a Java and Linux based system, which offers a command interface to interrelate with HDFS.
- HDFS provides file permissions and authentication, fault efficient, scalable and flexible storage system.

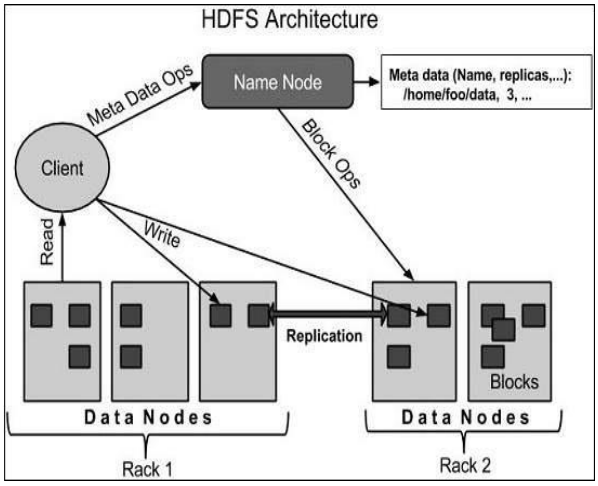

Figure 5: HDFS architecture

(Courtesy:https://www.tutorialspoint.com/hadoop/hadoop_hdfs_overview.htm).

Figure 5 shows the HDFS architecture / system, which go behind the master-slave architecture and it has the following elements to do particular task.

Name node: The HDFS system having the Name node and it acts as the master server and it provide the following jobs, Name node stores all information related to file system like

i) File section is stored in which part of the cluster
ii) Last access time for files
iii) User permissions like which user has right to use the files.

Name node executes HDFS file system operations such as opening directories and files, renaming, and closing.

Data Node: can manage and store the system data.

i) Data nodes carry out read and write operations on the file systems as required by the client. It compute the following operations such as deletion, replication and block creation, as per the Name node instructions received.

Table 1: Hadoop Components

Hadoop Component	**Functionalities**
HDFS	Distributed File System to store data with replication.
MapReduce	Distributed processing and fault tolerance
HBase	A distributed, scalable, big data store with Fast read/write access.
HCatalog	It is a table storage management tool for Hadoop that represent the tabular data of Hive metastore to other Hadoop applications.
Pig	It is a high-level platform for creating programs that run on Hadoop. The language for this platform is called Pig Latin.
Hive	It is the software of a data warehouse and set up on top of Hadoop for furnishing data query and analysis.
Oozie	It is a server based workflow scheduling system to handle Hadoop jobs.
ZooKeeper	It is a coordinates and provides active services for a Hadoop cluster and furnish distributed configuration service, synchronization service and naming registry for distributed systems.
Kafka	It is a software framework for stream-processing, messaging and data integration.
Mahout	It is Apache software framework to design a distributed or scalable machine learning techniques.

Blocks: users data in the big-data analytics are stored in HDFS files. The file in a file system will be alienated into one or more segments and/or stored in individual data nodes. These HDFS file segments are described as blocks. In other words, the least amount of data that HDFS can read or write is also called a block. The default block size of HDFS is 64MB, but it can be amplified as per the need of HDFS configuration.

Along with HDFS, Hadoop contains many essential components as required by big-data management system are shown in Table 1 [15].
Map-reduce [2, 6, 9, 10, 13] is the central hub or processing pillar of Hadoop, and it is a framework or programming paradigm, which enables mass scalability across numerous server in hadoop cluster [3, 15]. Map-reduce programming paradigm enables the specification of an operation to be applied on dataset, divide the problem and dataset, and execute in parallel [3].

Mainly it provides two important functionality, as shown in figure 6.

i) Distribution of a job based on client application task or users query to various nodes within cluster.

ii) Organizing and reducing the results from each node into cohesive response to the application or answer to query. Following diagram shows Map-Reduce architecture.

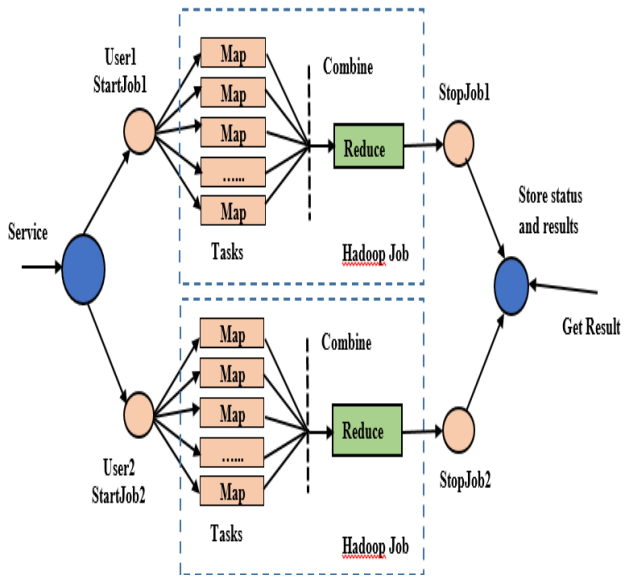

Figure 6: Architecture of Map-Reduce framework (Courtesy: Sukhpreet Singh [3]).

Mapper is the task, used to process all input records from a file and produces an intermediary set of key/value pairs. Mapper implements map function takes data as input and alters it into intermediate data, in which each data essential elements are split down into key/value pairs.

Reducer is another job, it will take the results of the mapper (intermediate key-value pair) as the input and generate the suitable results. The results of the reducer have been stored in HDFS. It has reduced the mapped data by using aggregation, query or user-defined function. Reducer functionality carried out in three phase are as follows.

i) Shuffle Phase: this phase facilitates to transmit data from the mapper to the required reducer. With the help of HTTP, the framework calls for related partition of the output in all mappers.

ii) Sort Phase: The results obtained from mapper that is certainly key-value pairs have been sorted on the origin of its key value.

iii) Once shuffling and sorting process has done the reducer unites the attained results and carries out the computation process as per the requisite.

However, map-reduce has an important framework in big-data processing, which share following features.

1. Provides automatic parallelization and distribution of computation.
2. Processes data stored on distributed clusters of data nodes and racks.
3. Allows processing large amount of data in parallel.
4. Provides scalability for usages of large number of servers.
5. Provides map-reduce batch-oriented programming model in Hadoop-v1.
6. Provides additional processing modes in Hadoop-v2 YARN based system and enables required parallel processing.

In Figure 7, we have shown an example of map-reduce task of heterogeneous data from different sources.

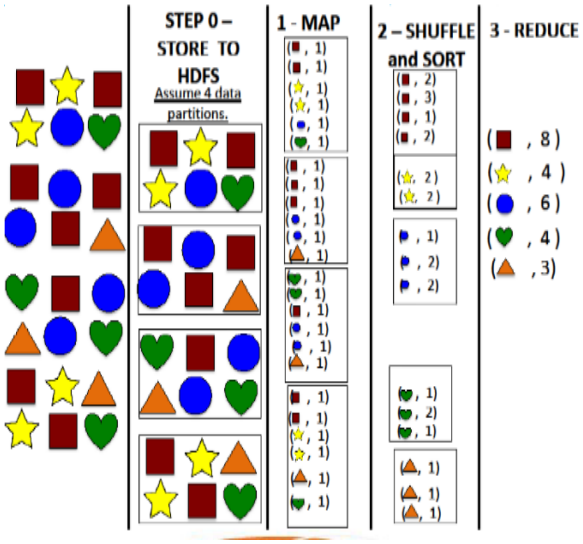

Figure 7: Example Map-Reduce Process

In the figure 7, different shapes represent dissimilar data, which processed in several steps. In step 0, fetched input data is partitioned as separate blocks, then mapper maps the data and produce key-value pairs (shapes considered as keys and frequency as value) then reducer takes output of mapper as

input and apply shuffle and sort functions, finally reducer combines the data based on values.

Consider the example, word count process to understand Map reduce process as shown in figure 8.

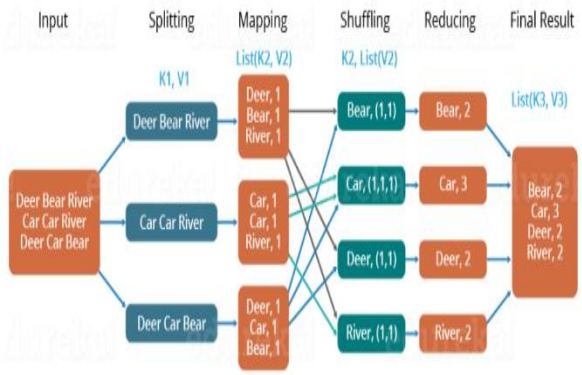

Figure 8: Map-Reduce Process for Word Count (Courtesy: Internet source https://www.edureka.co/blog/mapreduce-tutorial)

Figure 8 demonstrated the file which consists words of animal names given as input to mapper. In step 0, data will be partitioned, then mapper maps the partitioned data into key-value pairs. Output of mapper given as input to reducer, reducer will apply shuffle and sort functions, then combines data based on values.

HBase: It is an open-source, non-relational, versioned and distributed-database. It is enforced over the top of HDFS and it furnished the facility such as Google's Bigtable for Hadoop. It stores and process data by column based. HBase supports the distributed-data storage given by the fundamental distributed file systems spread across commodity servers. HBase has some important features include the following capabilities.

- It supports linear and modular scalability.
- It go behind strictly consistent reads and writes.
- sustains automatic and configurable sharding of tables.
- Supports the automatic failover and provides good communication between region servers
- Flexible Java API for client access

HCatalog: It is a tool used to manage HDFS and to store metadata. It is purely depends on Hive meta-store and integrated it with other kind of services (map-reduce and Pig) using a common data model.

Pig: it is the high level language and it allows writing the complex map-reduce conversion by the use of easy scripting language. Defines aggregate, join and sort transformations on data sets. Pig is used to extraction, transform and loading (ETL) data pipelines, quick research on raw data and iterative data-processing.

Hive: An Apache Hive provides a data-warehouse infrastructure and setup on top of Hadoop for providing a i) Ad hoc queries ii) Data summarization iii) Analysis of large data-sets using a SQL-like language called HiveQL. An Apache Hive offers following features

- The tools to facilitate data extraction, transformation and loading (ETL)
- Mechanism to enforce structure on variety of data formats
- Access the files, which are stored either directly in HDFS or in other data storage systems such as HBase
- Query execution via map-reduce and Tez.

Oozie: is a workflow director system and designed to run and manage multiple related Hadoop jobs. Oozie workflow jobs are represented as DAGs (Directed Acyclic Graphs) of actions as shown in figure 9.

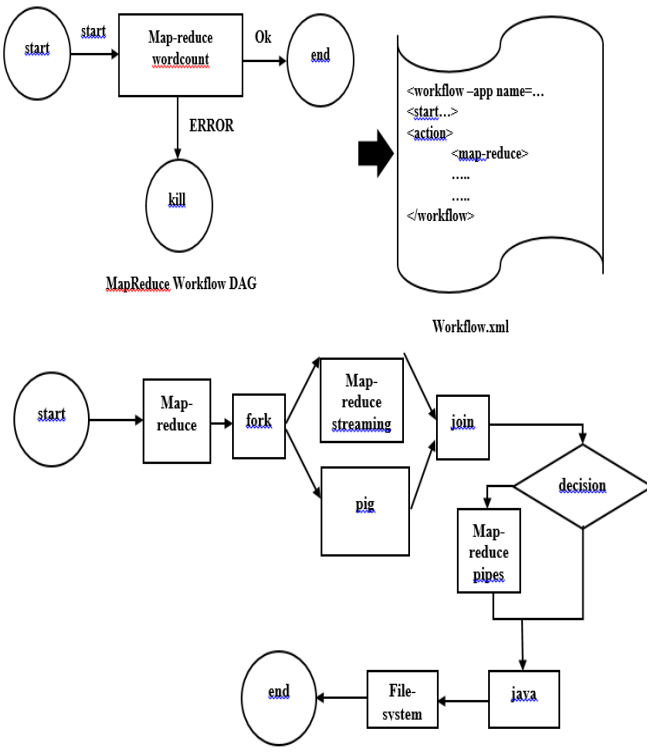

Figure 9: Oozie workflow graph
(Courtesy: Text Book: Hadoop 2 Quick-Start Guide: Learn the Essentials of Big Data Computing in the Apache Hadoop 2 Ecosystem, by Douglas Eadline, ISBN-13:978-9332570351)

Oozie performs three types of jobs as given below

1. Workflow- it is particular progression of Hadoop jobs with outcome-based assessment points and control dependency. Improvement from one act to another cannot ensure until the first act is completed.
2. Coordinator- is scheduled workflow job that has run at different time intervals or when data become available.
3. Bundle-is a higher-level Oozie abstraction that can batch a set of coordinator jobs. Oozie is incorporated with the rest of the Hadoop stack, supporting some types of Hadoop jobs out of the box (e.g., Java Map-reduce, streaming map-reduce, Pig, Hive, and Sqoop), as well as system-specific jobs (e.g., Java programs and shell scripts).

ZooKeepr: It is a coordination service, enables synchronization in cluster distributed applications and manages jobs in cluster. Zookeeper's main coordination services are:

1. Name Service: maps name to information associated with that name Ex: DNS maps domain name into IP
2. Concurrency Control: accesses shared resource in distributed system and controls concurrency
3. Configuration Management: new joining node can pick up up-to-date centralized configuration from Zookeeper when node joins system
4. Failure: automatic recovering strategy by selecting some alternate node

Kafka: It is an open-source tool and support for distributed event streaming platform used by thousands of companies for data integration, critical applications, streaming analytics, and high-quality data pipelines. However, it allows for distributed data store optimization for ingesting and data processing stream in real-time.

Kafka facilitates three main functions to its users which are as follows:

- Print and subscribe to the stream of data-records
- Successfully stores the streams of data records in the order in which data records were generated
- Process the real-time record streams.

Mahout: It is a library for linear algebra, machine-learning, data mining applications and it facilitates statisticians, mathematicians, and scientist to implement their own novel algorithms. By nature functionalities of Mahout are categorized into four groups:

i) Collective filtering
ii) Categorization
iii) Clustering
iv) Mining of parallel frequent patterns.

Important part of Mahout library is, it supports for distributed mode of execution and map-reduce execution process.

Table 2: Data Acquisition Tools

Name	Description
Flume	Basically flume is a distributed system. Hence, it collects the log data from dissimilar sources, combined data, and transfers the data to HDFS [7].
Sqoop	Import and export data between structured data stores and Hadoop [7].
Chukwa	Data compilation system for monitoring big distributed systems [7].
Semantria	Used to collect assorted information from different clients and it is impressively combining the various text analytics result [5].
OpenText	By nature it is a sentiment analysis tool used in process of classification to determine the various subjective patterns. Therefore, It is used to appraise the expressions of sentiment that is present in text form [5].
Trackur	This tool is designed to collect the information by using automatic sentiment analysis to gaze at the explicit keywords that the users are administer and after then decisions are carried out [5].
SAS Sentiment Analysis	Automatic extraction of sentiments in a real time scenario and it combine the various Natural language statistical modelling approaches and rules for assessment of sentiment.
Opinion Crawl	Sentiment analysis of current affairs in online social media has been done using Opinion Crawl and it allows diversified visitors' to evaluate the web / current affair sentiment on standard subjective matters [5].

Now let us review, process of fetching data from external source and processing further. Data acquisition and data collection is the process of fetching the structured-data, semi-structured-data or unstructured-data from different sources. Different tools and techniques are available for data acquisition and data collection process as given in Table 2. Initially let us discuss about data acquisition, which can be achieved through three steps

i) Data collection
ii) Data Transmission
iii) Data pre-processing

Figure 10 shows the taxonomy of data acquisition phase. In which, data collection steps involved in the process of collecting data from various sources like sensor data, surveillance data, inventory management data, log files, data from web systems etc,.

Collected data is may not be in proper format, it is very essential to transform collected data into proper format, so which can be processed further. Finally, before applying analytical methods on data, pre-processing is taken place to validate, clean or filter the data like eliminating redundant or duplicate values, fill missing values, remove invalid data, etc, are mentioned.

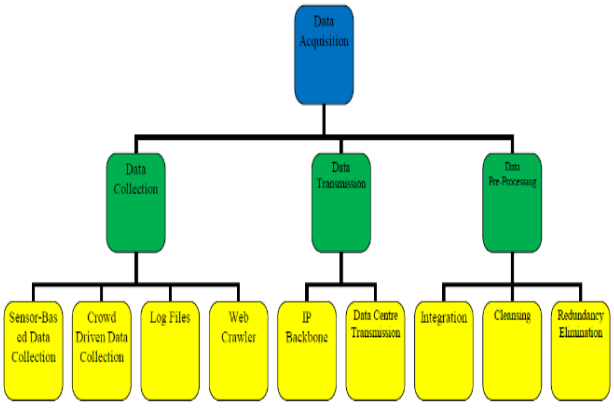

Figure 10: The taxonomy of data acquisition phase. (Courtesy: Amita Dhankhar et. al., [7]).

Table 3: Data Collection Tools

Name	Description
Scraper	Primarily it is an open source extraction tool it has supports and install as chrome extension with an option restricted information extraction. As a result, It is useful for analyzing the online commerce information and store the information to Google Spreadsheets [5].
Octoparse	This is an information extraction tool and used to flip the complete net into a structured format with this tool [5].
ParseHub	This is a web scraping tool and fetches the data from diversified site and fed into a spreadsheet or API. However, it facilitates to fetch the desired data from big data in the easiest manner [5].
Mozenda	The tool has captured the unstructured internet information and translates it into a structured format. Although, it uses point and click code tool to illustrate websites into structured information. Acknowledged by frequent names, like internet information gathering, internet scraping and internet information extraction [5].
Content Grabber	This is a cloud-based web scraping tool that assist businesses all sizes with data extraction [5].

Table 4: Data Cleaning and integration Tools

Name	Description
DataCleaner	This tool integrated with Hadoop and designed for data cleaning. It also support for data validation, assessment of quality data and data transformation and reporting [5].
MapReduce	This is a programming model of Hadoop framework. it has been used to retrieved the stored data in Hadoop File System (HDFS)
Rapidminer	This tool used to access, load and analyze any type of data. Thus, the tool has retrieve the valuable information from conventional structured data. Though, It can also transform the data from unstructured form to structured form [5].
OpenRefine	This is an open source powerful tool to manage cluttered data and unstructured data, make such data more precise. However, the tool has capable to edit, clean, reshape the data in an intelligent way. It also facilitated the additional features like reconciling, faceting, editing cells, clustering [5].
Talend	This is data integration tool for business insights. Similar to other tools, it facilitate the software services for cloud storage, enterprise application integration and data management

Once data have been collected, it is required to extract useful information by applying appropriate filters, to achieve this, need tools and methods, which are listed out in Table 3.

The extracted and filtered data need to be integrated and cleaned like removing of invalid values etc, for which it is, required some tools and technologies as briefed in Table 4.

5. Conclusion

In this paper, we have tried to explore the initiatives of big-data and its novel analytics approaches, diversified characteristics features, advanced tools which are available in the market to extract the data, store the data and analysis the data to render the best result. However, we enlighten the importance of

HDFS. In the present scenario HDFS file system is used to store the data in distributed manner and supports parallel processing. We empirically assess the map-reduce framework techniques. Thus, it has been used to transfer data from one format to another format and reduce the data by applying some aggregate functions. Our review article also identified and list out the various Big-Data analysis tools and such tools are used for supporting management of the big-data.

Acknowledgements
We would like to thank JNN College of Engineering and Department of ISE for their technical facilities to conduct literature survey on Big Data Analytics and Tools. Also thank to reviewers for their valuable comments and suggestions on the paper.

Declaration

Conflict of interest the authors certify that there are no conflict of interest with any organization for the present work.

References

1 Ahmed Elragal, Moutaz Haddara, Big Data Analytics: A Literature review analysis, Conference: NOKOBIT, Fredrikstad, Norway, November 2014.

2. Nada Elgendy, Ahmed Elragal, Big Data Analytics: A Literature Review Paper, Conference: Industrial Conference on Data Mining, August 2014, Lecture Notes in Computer Science 8557: 214-227,
DOI: 10.1007/978-3-319-08976-8_16.

3. Sukhpreet Singh, Ashwani Kumar, A Review Paper on Big-data and Hadoop, Proceedings of the International Conference on Recent Innovations in Science, Agriculture, Engineering and Management, 20th November 2017, pp. 732-739.

4. D. P. Acharjya, Kauser Ahmed P, A Survey on Big Data Analytics: Challenges, Open Research Issues and Tools, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 2, 2016.

5. Ritu Ratra, Preeti Gulia, Big Data Tools and Techniques: A Roadmap for Predictive Analytics, International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249-8958 (Online), Volume-9 Issue-2, December, 2019.

6. Nikhil Madaan, Umang Kumar, Suman Kr Jha, "Big Data Analytics: A Literature Review Paper", International Journal of Engineering Research & Technology (IJERT), ENCADEMS – 2020, Vol.8, Issue 10, 2020.

7. Amita Dhankhar, Kamna Solanki, A Comprehensive Review of Tools & Techniques for Big Data Analytics, International Journal of Emerging Trends in Engineering Research, Emerging Trends in Engineering Research, Vol. 7, No.11, Nov 2019, pp.556-562.
https://doi.org/10.30534/ijeter/2019/257112019
.

8. K. Venkatachalapathy, C.Krubakaran, Data Mining for Big Data: Issue and Challenges, International Journal of Research in Advent Technology (IJRAT), Special Issue, Available online at www.ijrat.org, International Conference INTELINC 18, 12-13 October 2018, Annamalai University, Chidambaram, Tamil Nadu.

9. M.Ann Bency, Analytics Methods and Tools on Big Data, International Journal of Scientific Research in Computer Science Applications and Management Studies, Vol. 7, Issue 4, July 2018.

10. Pranav B, Chethana Murthy, Review Paper on Big Data Analytics, International Research Journal of Engineering and Technology (IRJET), Vol. 07, Issue 07, July 2020.

11. Cheng Fan, Da Yan, Fu Xiao, Ao Li, Jingjing An, Xuyuan Kang, Advanced data analytics for enhancing building performances: From data-driven to big data-driven approaches, Building Simulation, Vol. 14, 2021, pp.3–24.

12. Shubham Tripathi, Review Paper on Big Data Analytics, Journal of Information and Computational Science, Vol.10 Issue 1, 2020, pp 372-378.

13. Rafi. Sk, Ramesh .B, Chenna Kesava Rao. M, Big Data Analytic Processing: processing platforms, South Asian Journal of Engineering and Technology, Vol.2, No.35, 2016, pp. 1–5.

14. Chun-Wei Tsai, Chin-Feng Lai, Han-Chieh Chao, Athanasios V. Vasilakos, Big data analytics: a survey, Journal of Big Data: a Springer Open journal, 2:21, 2015.
DOI 10.1186/s40537-015-0030-3

15.Nawsher Khan, Ibrar Yaqoob, Ibrahim Bakker Targio Hashem, Zakira Inayat, Waleed KamaleldinMahmoud Ali, Muhammad Alam, Muhammad Shiraz, Abdullah Gani, Big Data: Survey, Technologies, Opportunities, and Challenges, Scientific World Journal, Vol. 2014, https://doi.org/10.1155/2014/712826

16. Zhwan M. Khalid, Subhi R. M. Zeebaree, Big Data Analysis for Data Visualization: A Review, International Journal of Science and Business (IJSAB) , Vol. 5, Issue 2, 2021, pp. 64-75.
DOI:10.5281/ze nodo.4462042.

17.M.P Pavan Kumar, B. Poornima, H.S. Nagendraswamy, and C. Manjunath, Structure-preserving NPR framework for image abstraction and stylization, Journal of Supercomputing, Vol.77, No.8, 2021, pp. 8445–8513.
https://doi.org/10.1007/s11227-020-03547-w.

18. M.P. Pavan Kumar, B. Poornima , H.S. Nagendraswamy, C. Manjunath and B.E. Rangaswamy, Image-Abstraction Framework as a Preprocessing Technique for Extraction of Text From Underexposed Complex Background and Graphical Embossing Images, International Journal of Distributed Artificial Intelligence, Vol. 13, Issue 1, Jan 2021, pp 1–35. https://doi.org/10.4018/IJDAI.2021010101

19. M.P. Pavan Kumar, Poornima, B., Nagendraswamy, H.S. and B.E. Rangaswamy, HDR and Image Abstraction Framework for Dirt Free Line Drawing to Convey the Shapes from Blatant Range Images. Multidimensional Systems and Signal Processing, Springer. 2021. Multidimensional Systems and Signal Processing vol. 33, 2022, pp. 401–458.
DOI: 10.1007/s11045-021-00803-x.

20. M.P. Pavan Kumar, B. Poornima, H.S. Nagendraswamy, C. Manjunath, Structure Preserving Non-Photorealistic Rendering Framework for Image Abstraction and Stylization of Low-Illuminated and Underexposed Images. International Journal of Computer Vision and Image Processing, Vol. 11, Issue 2, 2021.
doi:10.4018/IJCVIP.2021040102.

21. Varsha M., Dr. Poornima B., Dr. Vinutha H. P., Pavan Kumar M. P, Predictive Model for Rice Blast Disease on Climate Data Using Long Short-Term Memory and Multi-Layer Perceptron: An Empirical Study on Davangere District, Annals of the Romanian Society for Cell Biology, Vol. 25, Issue 6, 2021, pp.4703–4722.