Technological Aspects and Role of Big Data in Cinematics

Jaswinder Kaur[1] Dr. Vijay Bhardwaj[2]
[1]Research Scholar (Ph.D.), UCCA, Guru Kashi University, Talwandi Sabo, Punjab, India.
[2]Assistant Professor, UCCA, Guru Kashi University, Talwandi Sabo, Punjab, India.

Abstract
Big data is a new driver of the world's economic and societal changes. The world's data collection is reaching a tipping point for major technological changes that can bring new ways in decision making, managing our health, cities, finance, and education. While the data complexities are increasing including data volume, variety, velocity and veracity, the real impact hinges on our ability to uncover the `value' in the data through Big Data Analytics technologies. Big Data Analytics poses a grand challenge on the design of highly scalable algorithms and systems to integrate the data and uncover large hidden values from datasets that are diverse, complex, and of a massive scale. Potential breakthrough includes new algorithms, methodologies, systems and applications in Big Data Analytics that discover useful and hidden knowledge from the Big Data efficiently and effectively. The research paper discusses the challenges faced by Big Data. The research paper primarily focuses on the technological aspect responsible behind the working of the Apache Hadoop framework. The paper discusses about the different components of Apache Hadoop framework with the particular job assigned to them. The paper illustrates the working of MapReduce algorithm with suitable example to depict the working of three phases involved within it.

Keywords: Apache Hadoop framework, Big Data, Cinema, HDFS, MapReduce.

I. INTRODUCTION

In this jet-set age amount of data has been increasing with the passage of time at an unbridled pace. This happened because of sophisticated technology. The big data is the thriving technology that deals with different kinds of data. The term data does not only include numbers and text, but it also consists of picture audio and video. Data drives the modern organizations of the world and hence making sense of this data and unraveling the various patterns and revealing unseen connections within the vast sea of data becomes critical and a hugely rewarding endeavor indeed. There is a need to convert Big Data into Business Intelligence that enterprises can readily deploy [1]. Better data leads to better decision making and an improved way to strategize for organizations regardless of their size**,** geography, market share, customer segmentation , and such other categorizations. Big data is a blanket term for the non-traditional strategies and technologies needed to gather, organize, process, and gather insights from large datasets [2]. While the problem of working with data that exceeds the computing power or storage of a single computer is not new, the pervasiveness, scale, and value of this type of computing have greatly expanded in recent years. Big Data is a term used to describe a collection of data that is huge in size and yet growing exponentially with time [3, 4]. In short such data is so large and complex that none of the traditional data management tools are able to store it or process it efficiently. Big data handles the massive amount of data involving complex relationships and garnered from multiple heterogeneous platforms. The growth rate of data is increasing and would further increase in upcoming years. The term big data came into emergence when conventional data processing systems become less adequate and inefficient to extract and analysis the data due to exponential growth in the amount of data. The colossal amount of data hinders traditional technologies to work properly.

II. BIG DATA IN CINEMA

India has been the biggest movie producer worldwide for the last few years, releasing more than a thousand films each year. The value of the Film industry in India is INR 138.2 billion, which is remarkably huge. Data analytics holds a significant potential for the media and entertainment industry. It is becoming an important part of companies' growth strategy enabling them to gain insights into their customers. These insights can help companies to optimize their marketing efforts and deliver a better product. Needless to say, without analytics, companies are in the dark about their customers. Inevitably, companies with data have an advantage over those who run on pure intuition. Film Executives from major production companies have stated on multiple occasions that Hindi cinema is now in crisis mode. Our audience has become selective. There is an emotion attached to every region – about what kind of genres work & what don't. Geographically speaking, there are some movies that work in some regions while some don't.

The existence of the Indian diaspora urgently calls for big data analytics to assist the cinema executives in understanding the scenarios of how their films are being received over different regions of the country – be it the metropolitan cities or remote areas – and even globally [1, 5]. Experts have established that there seems to be no data available in the right form to help movie producers in making critical decisions to obtain maximum sales [6].

Data Science will help one out with the various and exact variables of a movie that will let one know which movie audiences will prefer over others. While analyzing the data, one will recognize the components which are not required for developing the movie. Below are the four reasons why Indian films require Data Science.

- **Blocks Failure**

Data Science is completely based on the facts and the numbers that include the data of the productions, actors, songs and almost every possible variable. The data is extracted through the actual information gathered from the people or social media, which is then calculated to find out if the movie will work or succeed and also this data helps to avoid undesired consequences with lessening the chance of failure.

- **Determination of Release Time and Day**

Data Science results in determining the release time and the day by using machine learning, natural language processing or through social media in order to earn the maximum audiences. This data gives away to the directors by providing the specific information required for the release of the movies of different genres on different days. For example, most of the animated movies get released during the time of children's vacations and most of the movies of Salman Khan's get releases on the day of Eid.

- **Movie can make more Money**

This is not more than a fact that Data Science can boost your movie to earn more money than any others. The analysis of data is conducted before the release of every movie, considering all the factors like production, songs, actors, directors, locations, story and whatnot. Data Science also helps you to know which part of the city, state or country will help you to gain more audiences which will help you to make more money. For example, the movies of the actor Dharmendra used to run really well in Punjab or other northern parts of India.

- **Story Based on Interest of the Audiences**

Data Science recognizes the top trends going all around the world through social media and other mediums to know what can get you the success. Nowadays maximum of the movie-makers in the industry started utilizing

the data received through the analysis to generate a story or plot in which the audiences will genuinely show their deep interest.

III.5V'S OF BIG DATA

The 5V's concerned with handling Big Data are shown in Fig. 1 and discussed as under [6, 7].

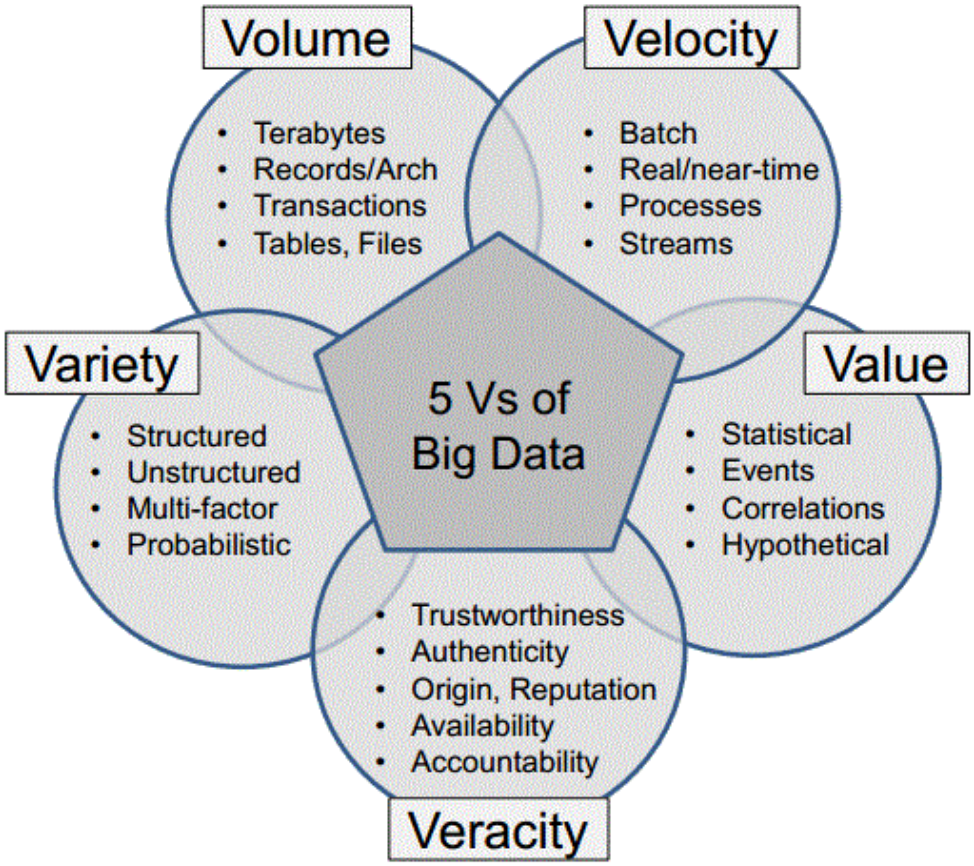

Fig. 1: Figure shows the 5 V's pertinent to Big Data

- **Volume**

The massive amount of data has been produced in different arena of the world. It is speculated that the database of the multinational companies gets twice every year. The network data has a major percentage in the global data contribution. The world has 800,000 petabyte data in 2000. It is been predicted that this amount would be anticipated to be 35 zettabytes in 2020. The data generation rate has been accelerating due to human preference for social media. Approximately 6 million persons are using digital media. Going by speculation it is predicted that around about 2.5 quintillion bytes data is produced every day.

- **Veracity**

Big data is produced from different data sources in a different format. Veracity refers to the preciseness of the data. In other words, it can be said that how much faith one can have on such data collected from multiple sources. To cope up with the problem veracity encompasses quantity, reliability, authenticity and trustworthy nature.

- **Variety**

Variety means different formats and different types of data. The variety has become a major hindrance in knowing ins and outs of data. Most of the time the generated data is unstructured having a slew of format. It is

predicted that organizations hold 90% unstructured data of the total data. It gets twice every three months. The storage and retrieval have become a major problem just due to the diversity of data. The diverse nature of data includes complexity, dimensions, features data types and formats. Only a high processing modern system is capable of understanding and fetching purposes. The main source of unstructured data is audio, video, images etc.

- **Value**

Not all data carries the same weight. Some amount of data may be more important to the organization and some may not be. Analysts use various tools and techniques to identify data relevant to an organization. Microsoft bought forecast which was a company of United States of America in 2008 having an air ticket prediction system and Microsoft implemented it into a bingo search engine this system led to saving the amount dollar 50 per ticket by 2012 with prediction accuracy 75%.

- **Velocity**

The data is generated at an alarming rate which cannot be imagined by a human. It must be analyzed to get value from that data. An unprecedented amount of data has been created by the social media mobile phone and retail sector only. Walmart has the capability to generate one million transactions per hour in the retail sector [2].

IV. CHALLENGES FACED BY BIG DATA

Data is produced at an enormous pace. Due to the high volume of data, various challenges are faced by big data experts. These challenges create problems during the time of analysis of data. All of these issues or challenges are explained below [8].

- **Growth Rate of Data**

The process of producing the data is rising without any stoppage. A major portion of data is covered by the unstructured data. It has become an uphill task for companies or organizations to store and analyze such an enormous amount of data. As per the revelation of the recent report the amount of such data is increased to 45% in 2016 from 30% in 2015. Whether the new technologies are being developed to deal with the gigantic store of data but the companies have to scale the hardware. Due to this the storage associated with the big data system hikes. In a single day, the 2.5 quintillion bytes of data are produced. It is said that 90% of the total is generated in the bygone two years. There are various sources of data such as search engines and online social internetworking sites [8, 9].

- **Problem of Filters and Hardware**

All the generated data is not useful. Valuable data need to be separate from invaluable data. Finding useful data is quite a time consuming task. The organization should have advance filters to segregate useful data. The requirement of filters has also become an issue for companies. Storing the massive amount of data that is useable for the companies and analysis of it requires complex hardware. The most used hardware only provides efficient productivity only for a certain period of time. The long period of use results in malfunctioning of hardware. To solve this problem the companies use the backup facility but the main challenge for them is to obtain the level of service when the crash occurs during the time when the file is being uploaded by the client [8, 10].

- **Privacy and Security**

It is always a moot issue of whether how the data should be used and shared without breaching the privacy of people. Data may contain confidential and private information like financial records that cannot be shared with anyone but companies cannot handle an enormous amount of data. They depend on third party people. By the

reason of this factor, the privacy of data and information is in peril. The valuable data is not received by the company because the data is not shared due to the loss of privacy and security. If all the data is made available to companies then a better decision can be taken but this does not happen usually.

- **Problem of Ownership**

Data is generated by multiple different sources. Most of the data sources are unknown and cannot be verified easily no one knows who the owner of data is. The accuracy of different data generation sources is going to become the biggest challenge for companies. This challenge makes its propagation towards the mining results. To overcome this problem the validation of data and provenance has become highly imperative in the overall knowledge discovery process. It is also seen as a challenge for big data users in the modern scenario.

- **Talent Shortage**

As the popularity and usage of big data will increase, the talent scarcity will also be seen. As per estimation, only the United States of America will witness the paucity of 140000 to 190000 experts of big data. Business organizations are not only suffering from the problem of deficiency of data talent but also for the successful implementation which wants an expert team of big data developers. Many companies are trying to solve this problem by providing education to big data developers.

- **The Database as Challenge**

In gone by times the RDBMS system was used for storing and analyzing the data. That can only deal with structured data. But it has no power to tackle unstructured and semi-structured data. Owing to an increase in the expansion of data the storage retrieval and management has become an issue for traditional database systems.

V. TECHNOLOGICAL ASPECT

The Apache Hadoop framework is composed of the following modules

- Hadoop Common: contains libraries and utilities needed by other Hadoop modules
- Hadoop Distributed File System (HDFS): a distributed file-system that stores data on the commodity machines, providing very high aggregate bandwidth across the cluster
- Hadoop YARN: a resource-management platform responsible for managing compute resources in clusters and using them for scheduling of users' applications
- Hadoop MapReduce: a programming model for large scale data processing

All the modules in Hadoop are designed with a fundamental assumption that hardware failures (of individual machines, or racks of machines) are common and thus should be automatically handled in software by the framework. Apache Hadoop's MapReduce and HDFS components originally derived respectively from Google's MapReduce and Google File System (GFS) papers [11].

This section details the Apache Hadoop framework and its constituting components along with the detailed working of the MapReduce algorithm.

Apache Hadoop Framework

Apache Hadoop is an open-source software framework used to develop data processing applications that are executed in a distributed computing environment. Applications built using HADOOP are run on large data sets distributed across clusters of commodity computers. Commodity computers are cheap and widely available. These are mainly useful for achieving greater computational power at a low cost. The data generated by multiple sources have not a specific format. Data may be structured, unstructured and semi-structured. The different kinds of data have emerged as a challenge to traditional data mining techniques. Apache Hadoop provides a framework to analyze and store all kind of data that was not in the approach of traditional

technologies. It is the most widely used and open-source platform and developed by 2007 with various tools. It is programmed in Java. It is inherited from MapReduce of Google and Google file system. It is developed by Dough Cutting and has driven its name from a toy elephant. The Hadoop is sponsored by Apache Software Foundation. It is used for processing and storing data. The storage and processing of large data sets are carried out in machines of commodity hardware. The data is analyzed and processed over multiple nodes by using MapReduce which has the ability to work with any kind of data [12, 13]. It has a feature of scalability, reliability and distributed computing. It offers the facility to transfer data among multiple nodes over HDFS. If any nodes fail during processing, another node can take over its work to provide maximum throughput. The Hadoop framework is divided into three layers: Storage, Resource management and processing management layer. Fig. 2 shows the composition of the Apache Hadoop framework. The different components of the Hadoop Framework are briefly explained below.

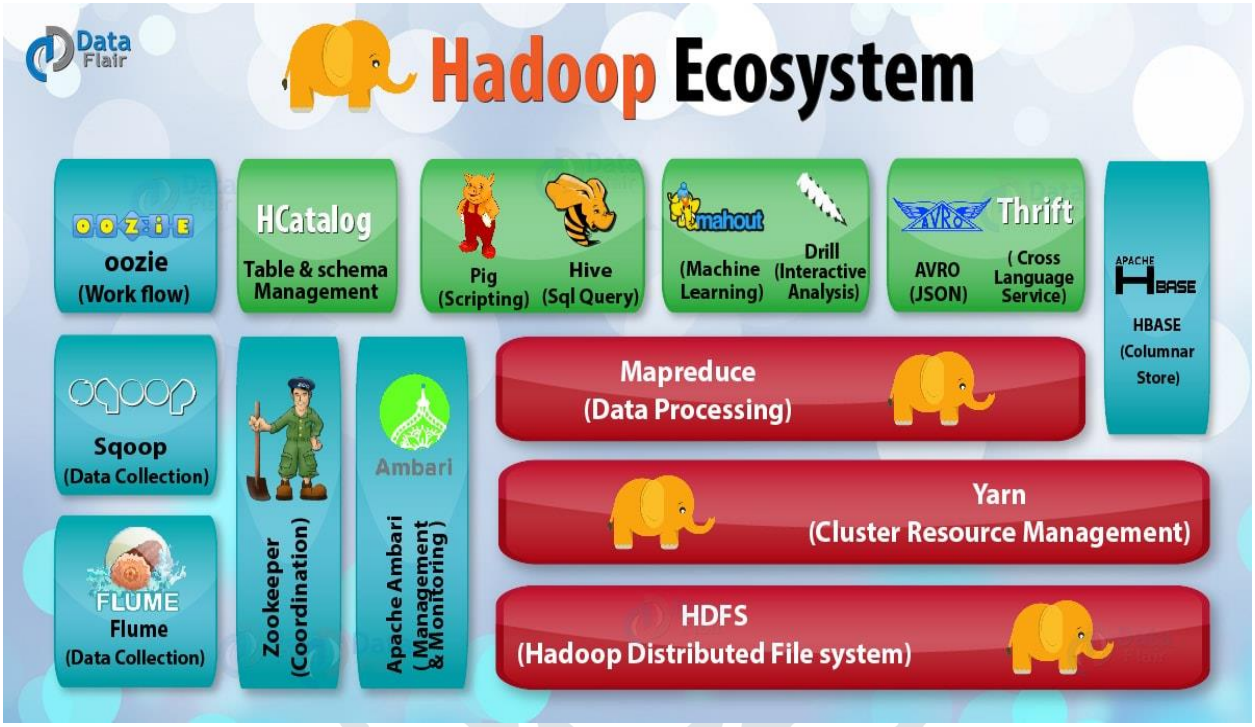

Fig. 2: Figure illustrates different components of Hadoop framework

- **HDFS**

It stands for Hadoop Distributed File System. It was developed to store datasets over commodity hardware with two nodes namely data node and name node. It follows master-slave architecture. The name node monitors all data nodes and stores all the Metadata. The data node actually stores and processes data and results back to the name node. In HDFS the file is divided into a sequence of blocks. It offers fault-tolerant features with the replication of data. By default, the replication factor is 3. The number of replicas that are generated of each block of a file is specified during file creation time.

- **YARN**

In Hadoop version 2.0 the YARN was not introduced. The tightly coupled hardware configuration is seen in Hadoop and MapReduce. MapReduce holds the responsibility of resource management but when the YARN was introduced this responsibility was held by it. With the segregation of programming model and MapReduce when any applications need to be executed the application manager process is launched by resource manager at the arrival of client request and then the application manager activates the container where the application is executed. It also distributes tasks across the cluster.

- **Pig**

The distributed analysis of data is done with the Pig. It uses Latin as a programming language. PIG tasks can not only be executed in MapReduce but can be done on another platform such as SPARK. It offers support for a user-defined function which is written in Java, Python etc.

- **Hive**

It uses an HQL which is similar to SQL and developed by Facebook. The System catalog of Hive stores Metadata of tables. It is also known as the Meta store.

- **Flume**

It is used for processing for pushing the semi-structured and unstructured data to HDFS.

- **Sqoop**

This tool is used mainly for transferring data between Hadoop and Relational databases. It is used to import and export data.

- **Zookeeper**

It maintains the coordination among different nodes over HDFS. It is also responsible for performing synchronization.

- **Apache Spark**

The processing of big data will be carried out using Apache Spark in the near future. It is developed at the University of California at Berkeley by researchers. The unprecedented feature of Apache Spark is that it has the ability to perform in-memory computations which separates it from Hadoop. It removes the I/O disk drawbacks of Hadoop by caching data in memory. It is faster as compared to Hadoop and supports Python, Scala, and Java. It can work with Hadoop and YARN and can also be used in standalone mode.

- **Apache Storm**

Apache Storm was developed for real-time computation with a view to removing all the drawbacks in the analysis and collection of data of the social media sector. It was launched by back type social media company and included in the popular list of the project of Apache in 2014. The increase of real-time computations led to the popularity of Apache Storm. The Storm architecture has two main components: spouts and bolts. Spouts work as an input stream. Bolts have a computation engine and process data in tabular form which is generated from spouts or bolts itself. The spout and bolt network is termed as topologies and presented by a directed graph. It offers ample benefits in terms of fault tolerance. Storm works with two nodes: master node and worker node. The master node is also called nimbus daemon which is responsible for monitoring the heartbeat of the worker node. The assignment of jobs is also carried out by nimbus.

- **Apache Flink**

The Flink has seen its development at the technical university situated in Berlin with the name Stratosphere. It was involved in Apache top-level project in January 2015. In this platform, the processing can be done in-stream as well as in batch mode. It provides scalability and in-memory computation. It can also run independently or can be integrated with Hadoop along with many functions, for instance, Map, Reduce Group, Join function etc. It has the ability to automatically pick a better execution model for each task. Its compatibility can be seen with MapReduce. It utilizes more resources and has the ability to execute every job in a short span of time.

- **H_2O**

It is an open-source platform that is still undergoing a thorough research phase. It is said that it would provide more speed but nothing has been published about this feature as yet. It includes analysis and machine learning kits along with tools for data processing and evaluation. It goes well with analysts who have not strong programming background due to its web-based interface. It offers support for Python, Scala, and Java. The data is fully processed in memory by data execution engine along with various execution methods.

- **MapReduce Algorithm**

MapReduce has a programming environment that has the capability to process a gigantic amount of data. Hadoop runs the MapReduce program which is written in many languages such as Java, C etc. This platform was developed by Dean and Ghemawat at Google used for processing data in Hadoop. It consists of two functions: the mapper function and the reducer function. Mapper takes the HDFS data as input and after processing sends results back to the reducer. The reducer aggregates all intermediate results to obtain the output. The parallel programming nature of data allows a large amount of data to be processed on multiple nodes in a short time. The overall execution process is handed over to two main components: job tracker and multiple task trackers. The job tracker maintains its position on name node and multiple task trackers are on the data node. The job tracker works as a master and the multiple task trackers work as a slave. The job tracker maintains the synchronization between different jobs which are executed on different nodes and monitors the result of each task [14]. The task tracker sends a heartbeat signal to the job tracker to tell about the current state of the system. Fig. 3 shows the working of MapReduce and phases which are responsible for the execution of the task.

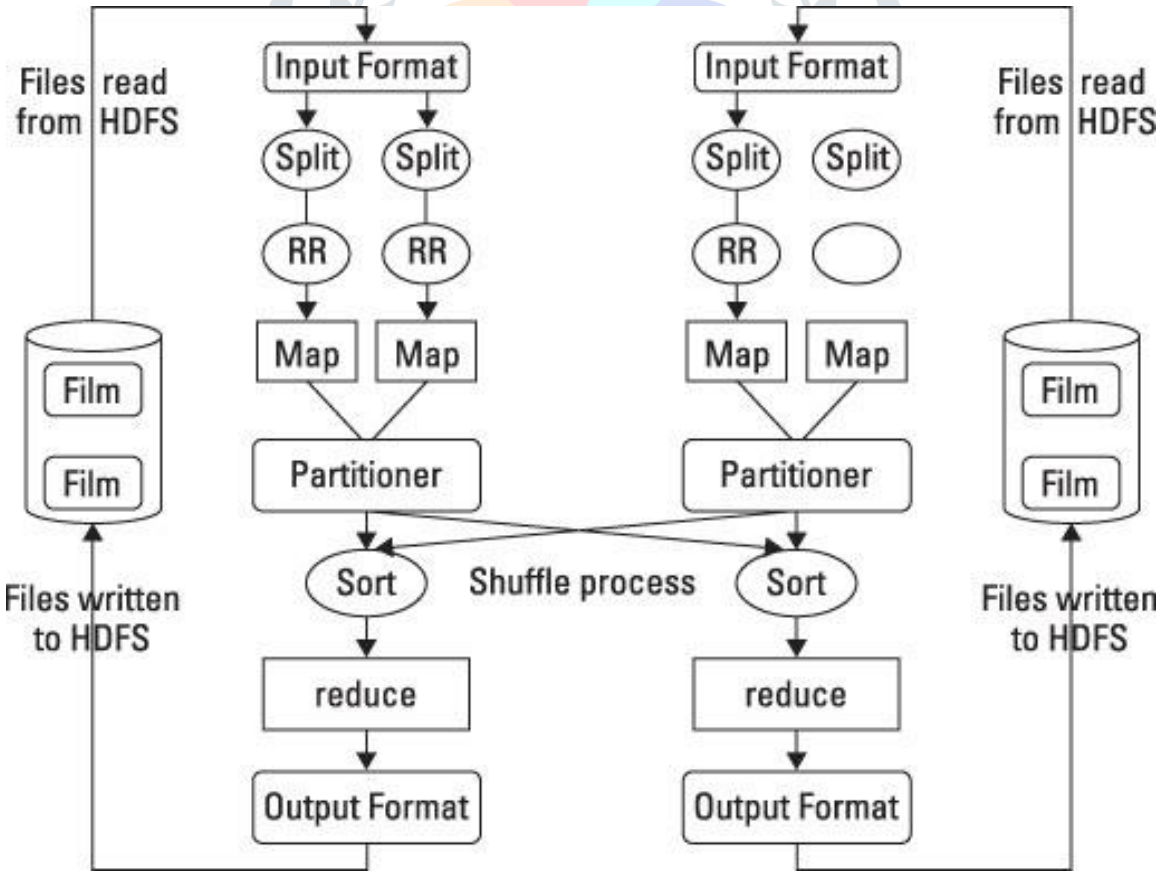

Fig. 3: Working of MapReduce algorithm

The working of MapReduce is divided into four steps which are explained below.

- The input passed to MapReduce is segregated into the different fixed-sized module. These fixed-sized modules are called splits. The divided input splits are passed to different mappers. Each mapper holds only one split.
- It is the initial step in the processing of data. Mapper functions perform processing over different splits. As a result, the key-value pairs are generated by the mapper function. These key-value pairs are also called intermediate results or records.
- This step takes the results of the previous phase as input and shuffles them to over multiple nodes.
- The reducer collects the results of the shuffling phase and aggregates them to produce single value results.

Example showing the working of the MapReduce algorithm.

Consider 5 people having Facebook accounts as A, B, C, D, and E. The friend list of each one is given on the right side against its name.

A -> B C D

B -> A C D E

C -> A B D E

D -> A B C E

E -> B C D

The MapReduce has been carried out as shown in Fig. 4 on the assumed example and the result has been extracted in relevance with the "Who knows Whom" concept as carried out by Facebook.

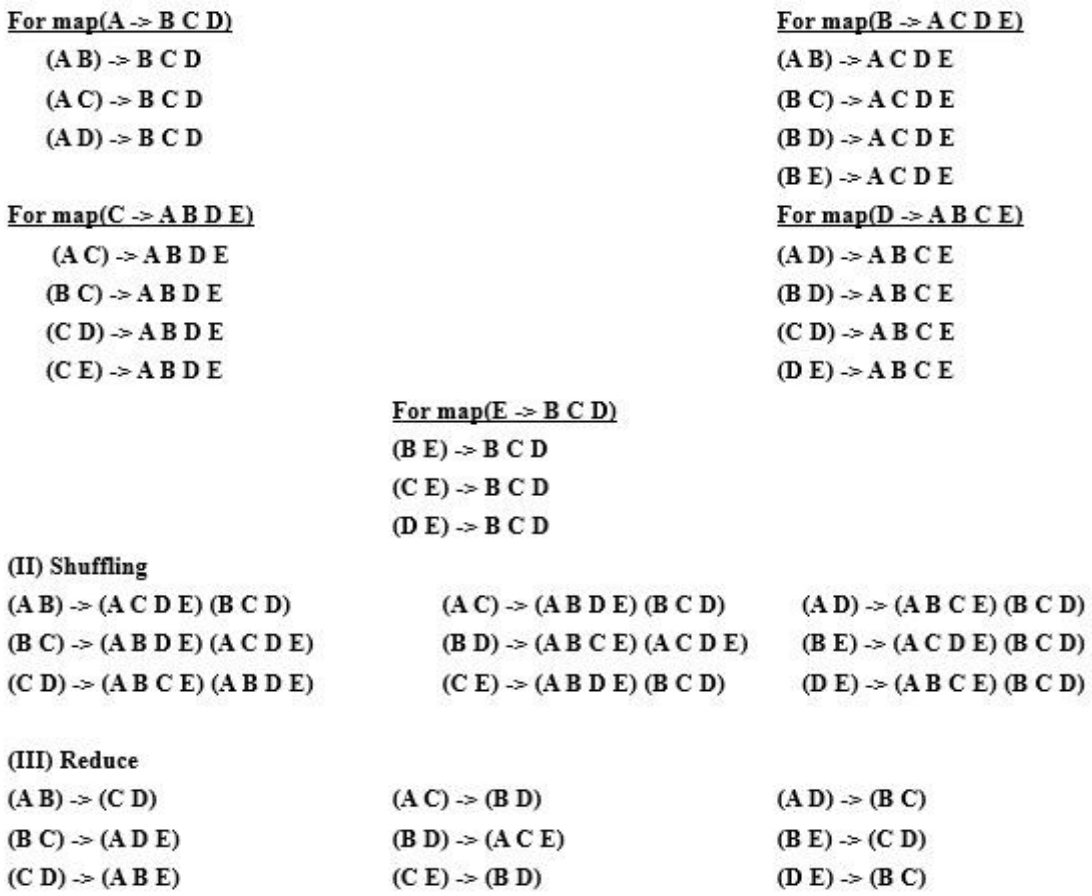

Fig. 4: Figure shows the working example of the MapReduce algorithm

VI. CONCLUSION

The research paper discussed the concept of Big Data, 5 V's, and challenges related to Big Data. The paper also elaborated on the challenges faced by Big Data. The paper also discussed the role played by Big Data in cinema. In the later sections, the paper focused on the technological aspect of Big Data and provided details on the Apache Hadoop framework and MapReduce algorithm.

REFERENCES

[1] Atta Badii, Ivo Keller, Mathieu Einig, Tobias Senst, Thomas Sikora, Volker Eiselein 2013 'Prediction of movies box office performance using social media'.
[2] Honghai Liu, Shengyong Chen,and Naoyuki Kubota 2013 Data cleaning for data mining ASurvey, in IEEE Computing for Sustainable Global Developmenton.9,3, pp. 1222.
[3] S. He and D. Feng. "Design of an object-based storage device based on I/O processor. SIGOPS Opera.Syst. Rev., 42(6):30–35.
[4] Jagdev, G. et al., "Excavating Big Data associated to Indian Elections Scenario via Apache Hadoop," International Journal of Advanced Research in Computer Science (IJARCS), ISSN 0976 – 5697.
[5] Jagdev, G. et al., "Big Data Proposes an Innovative concept for contesting elections in Indian Subcontinent," International Journal of Scientific and Technical Advancements (IJSTA), ISSN-2454-1532.
[6] Jeffery, D., Ghemawat, S., "MapReduce: Simplified Data Processing on Large Clusters." Google, 2004.
[7] Kaisler, S., Armour, F., Espinosa, J. A., Money, W., "Big Data: Issues and Challenges Moving Forward," International Conference on System Sciences (pp. 995-1004). Hawaii:IEEE Computer Soceity.
[8] Abdul Manan koli, Muqeem Ahmed, "Election Prediction Using Big Data Analytics-A Survey", International Journal of Engineering & Technology, 7 (4.5) (2018), pp. 366-369.
[9] Mario Callegaro, Yongwei Yang; "The Role of Surveys in the Era of "Big Data"; The Palgrave Handbook of Survey Research, 2017, pp. 175 – 192.
[10] Gagandeep Jagdev et al.; "Implementation of Big Data Concerned with Elections using Map-Reduce as Novel Mining Algorithm"; pp. 129 – 135; Proceedings of International Conference on Communication, Computing and Networking – 2017.
[11] Gagandeep Jagdev et al.; "Excavating Big Data associated to Indian Elections Scenario via Apache Hadoop"; International Journal of Advanced Research in Computer Science; Volume 7, No. 6(Special Issue), November 2016.
[12] Jagdev, G. et al., "Scrutinizing Elections Strategies by Political Parties via Mining Big Data for Ensuring Big Win in Indian Subcontinent," WECON-2015.
[13] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Ding, " Data Mining with Big Data", January 2014.
[14] Jeffrey, D., Ghemawat S, "The Google File System", 2003.