


DOI: <http://dx.doi.org/10.21123/bsj.2022.19.4.0887>

## New and Existing Approaches Reviewing of Big Data Analysis with Hadoop Tools

Watheq Ghanim Mutasher<sup>1\*</sup> 

Abbas Fadhil Aljuboori<sup>2</sup> 

<sup>1</sup>Information Institute for postgraduate studies, Iraq.

<sup>2</sup>University of Information Technology and Communications, Iraq.

\*Corresponding author: [ms201920531@iips.icci.edu.iq](mailto:ms201920531@iips.icci.edu.iq)

E-mail address: [abbas.aljuboori@uoitc.edu.iq](mailto:abbas.aljuboori@uoitc.edu.iq)

Received 1/3/2021, Accepted 24/6/2021, Published Online First 20/1/2022, Published 1/8/2022



This work is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/).

### Abstract:

Everybody is connected with social media like (Facebook, Twitter, LinkedIn, Instagram...etc.) that generate a large quantity of data and which traditional applications are inadequate to process. Social media are regarded as an important platform for sharing information, opinion, and knowledge of many subscribers. These basic media attribute Big data also to many issues, such as data collection, storage, moving, updating, reviewing, posting, scanning, visualization, Data protection, etc. To deal with all these problems, this is a need for an adequate system that not just prepares the details, but also provides meaningful analysis to take advantage of the difficult situations, relevant to business, proper decision, Health, social media, science, telecommunications, the environment, etc. Authors notice through reading of previous studies that there are different analyzes through HADOOP and its various tools such as the sentiment in real-time and others. However, dealing with this Big data is a challenging task. Therefore, such type of analysis is more efficiently possible only through the Hadoop Ecosystem. The purpose of this paper is to analyze literature related analysis of big data of social media using the Hadoop framework for knowing almost analysis tools existing in the world under the Hadoop umbrella and its orientations in addition to difficulties and modern methods of them to overcome challenges of big data in offline and real-time processing. Real-time Analytics accelerates decision-making along with providing access to business metrics and reporting. Comparison between Hadoop and spark has been also illustrated.

**Keywords:** Apache-Spark, Big Data, Hadoop, IOT, Social Media.

### Introduction:

There are more than one billion social media network users, many of whom are regularly involved around the world and can be Linked through their phones and tablets. Social networking, programs that allow users to Online networking, in recent years, has grown in popularity. The set that is commonly accessed is from those where consumers can use Post comments on social networking in near real-time<sup>1, 2</sup>.

The ongoing increase is in the output of large volumes of Blog info, tweets, Facebook, Twitter, social media, etc. Sites for networking, mechanical/electrical sensors, company processes open up a large range of possibilities for data collection, processing, and research organizations. Decision-makers use data to understand the needs

and expectations of customers that assist in formulating develop marketing techniques and achieve competitive advantage in the marketplace<sup>3</sup>.

To extract useful information, Hadoop is a common data set from these large datasets, an Open source platform for massive data sets to be processed on Hardware Distributed Product. But in today's tale, an environment for discussing this vast volume of knowledge is a world for a daunting job that requires costly hardware, Adaptation, dedicated storage, and complicated applications Prohibition of Big Data technologies for small companies. Cloud computing includes computing to prevent this bottleneck. Sources such as storage, servers, pay computing resources Rather than creating your costly peruse model, Infrastructure for hardware

and applications Cloud Use Computing allows businesses to concentrate on their business. Objectives and gains have been without thinking about problems like Resource supply, infrastructure, IT experts<sup>3</sup>.

### Characteristic of big data:

Colossal height, high dimensionality in nature and challenging. Data on a large scale does not mean enormous in size. Describes its meaning Fatten Description by 5V's3V'ss defined by Doug Laney Amount, pace and variety, they are:

1-Volume: Giant data in size which is high in volume Terabytes have been found in petabytes, too.

2- Velocity: This refers to the rate of data that can be transmitted into and out of the Sources, including a device, cell phones and so forth.

3-Variety: Identifies various formats of data, like Framed information, semi-framed information and disorderly data Sources and sources.

4- Veracity: Incompatible and discordant are identified Details that can make data messy at

times, consistency Accuracy and precision are hard to handle.

5- Value: It is below 5Vs when there is a massive amount of The company's invaluable data and all, unless you turn it into something useful<sup>4</sup>.

### Hadoop system:

Hadoop is free programs platform which ensures parallel processing for big datasets using programming models through commodity server clusters. It is designed to scale, and also a very greater rate of tolerance for the mistakes., as a single computer to a large number of computers. The Hadoop architecture consists of the distribution of data on the file system Hadoop (Hadoop Distributed File System), the Scripting model Map reduce as well as many other approaches. such as Hive, Pig, Sqoop, Hbase, Mahout, etc. named ecosystem tools for managing large datasets. Termed as components of the ecosystem<sup>5</sup> can be illustrated in (Fig. 1).

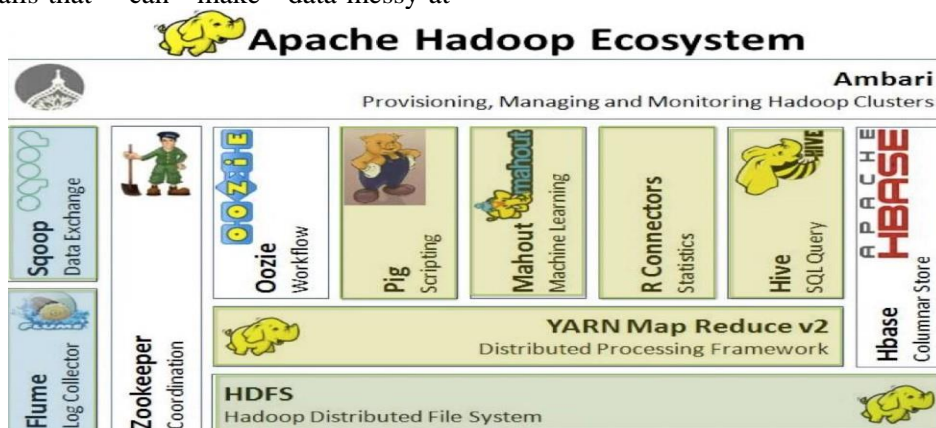


Figure 1. Hadoop system components<sup>5</sup>

Hadoop distributed file system and also Mapreduce are in the Hadoop ecosystem. Critical elements that are clarified in depth and accompanied by Other modules are optional<sup>5</sup>.

### Hadoop Distributed File System (HDFS):

Apache Hadoop is a distributed file system application Design of a device that stores a huge quantity of information and also provides methods simpler to reach more customer scattered around the board system. It is really tolerant of faults and is planned running on equipment at a low price (called

generic hardware). Files found in the files, the redundant HDFS, is stored across several machines Mode to restore the loss of data if it fails. Hadoop is applied as a file system block-structure when filenames are break through the fixed size block classed on Hadoop Platform<sup>5</sup>. The Hadoop distributed file system uses the Name and Name Master slave architecture Nodes with details. The Word Node functions as a master when there are several nodes of data served as slaves<sup>5</sup>. HDFS's architecture is displayed in (Fig. 2)<sup>5</sup>.

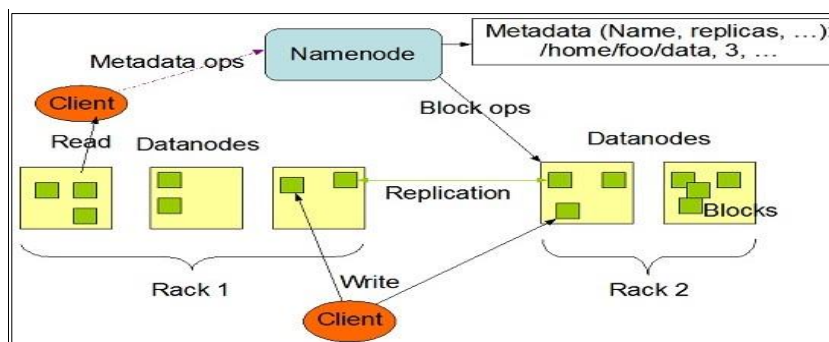


Figure 2. HDFS structure<sup>5</sup>

One name node is the sole authority for controlling the namespaces of the file system in the Hadoop architecture and controls customer access. It stores all file system metadata across the clusters. Several data nodes hold data to handle the storage connected toward the node which are running at. The Hadoop distributed file system client will do various operations I/O and file read or write operations stored on the HDFS data store<sup>5</sup>.

**Map- Reduce (Distributed Processing data):**

Decrease Map (Distributed Processing data), Map-Reduce is a method of programming that is used to process huge number of data of knowledge. The Map-Reduce operates On the process of splitting and conquering, where even a big complicated issue is a big problem as a smaller

group that are overcome and broken down to about the distributed clusters simultaneously and separately. Moreover, all the other alternatives were mixed with each other to achieve the goal, the initial problem response. The data processing primitives used are reduced in the map model and named the reducer and mapper. Any software that reduces maps must have at least one subroutine for mappers and reducers. The Mapper or the map method converts the input key and value pairs to any input key and value pair, number of different pairs of intermediate key values, whereas the reducer has a reducer that reduces the method of transforming pairs of intermediate key values whose value pairs are aggregated into any number of output keys<sup>5</sup>. The Map-Reduce programming is involved in multiple stages as shown below in (Fig. 3).

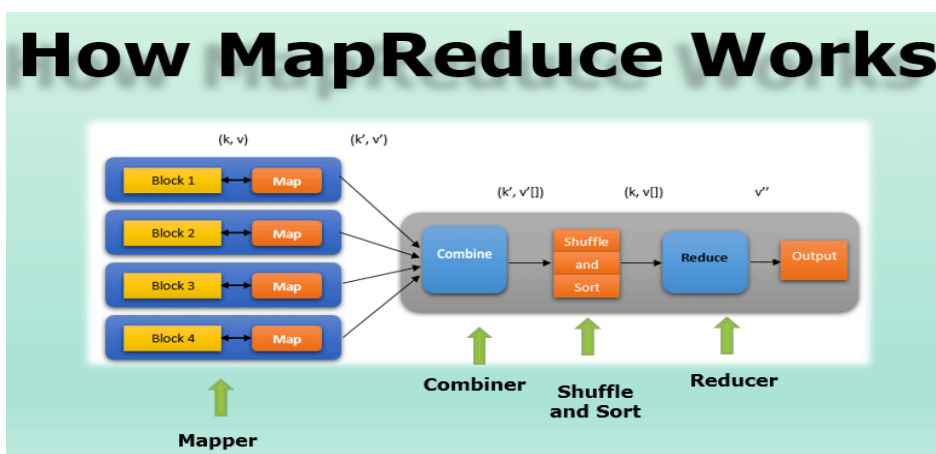


Figure 3. Map Reduce Architecture<sup>5</sup>

A broad dataset consists of <key, value> in the input process. The pair is given as a regular input for the program for mapping reduction. The data files used for the map reduction are stored on Hadoop Distributed File System stores that already have an Input Format Norm User-specified. Then the split is selected until the input file is selected. The input data is read and split into smaller phases Parts. Then the split segments are providing to mapper. The extract operations map and generate the related data Pairs of Middle Key and Value. It

displays split records from a data file and achieving medium outcome that converts and transforms input keys to output keys, value array info, list of values, which is then transferred to the combiner. Using the combiner is to minimize the data volume with both the mapper and the reducer move. Also it is denoted to as a semi-reduced that accepting the incoming values Output main value pair to reducer, from mapper and passes. The elements of the reducer are shuffle and form the shuffling is shuffling A partitioning method and the transfer of measured or

defined output towards the reducer, in which the reducer is assigned intermediate keys. Every partition is known as a sub-set. Every dataset entered the shuffling of the general process of the reducer in guarantees that partitioning goes on partition divides completed at suitable reducers where for http is used by the reducer Protocol to retrieve the mapper's own partition. The type, the sort is the duty of the stage to sort the intermediary keys on the one key node before even being transferred to the reducer automatically. Simultaneously, the sort and move of different stages appear where the mapped result is retrieved and combined. The shirker decreases a set of intermediary values that shared distinctive keys with the Value Set. Sorted input is used by the reducer to produce the Definitive performance. Using record authors, the final performance is written by the reducer is in the normal output format of the output file. Each map reduction program's final output is created with key results. Written value-pairs in the output file are written back to the output file Shop for HDFS <sup>5</sup>.

### Hadoop of Ecosystem:

To increase Hadoop's productivity and results, there are several inventing's that are founded on the head of the Hadoop. Also for those technology collections, Hadoop is The so-called Hadoop Environments. Basic system of Hadoop Eco It includes exploring new technologies <sup>6</sup>.

### Apache Pig:

Pigs is a scripting language that was originally created by Yahoo to build programs for Hadoop by Yahoo utilizing the procedural language classified as Latin Pig, which assists in the Production of the Hadoop cluster's large data set. The Pig is a Pig a Java alternative for the development of Map Reduce programs that Enables Programmers to spend less time writing code mappers & mapping Reduction systems rely mainly on datasets being evaluated. Such as real pigs that can eat almost everything; the management of every type of thing. Species of knowledge from during the programming languages of the Pig hence the name Pig. The general concept is that it takes Pig Scripts to write Compared to writing the program Map Reduce, 5 % of the time. The advantage you need just to write far lesser lines. Codes will reduce the total time to create and test <sup>6</sup>.

### Apache Flume:

The Apache Flume provides a distributed environment., secure and open product. A service to store, aggregate, and accumulate effectively and

transfer a huge volume of data over to HDFS for streaming. It a streaming-based, easy and versatile infrastructure easy and scalable infrastructure. Flow of data is stable and more forgiving of faults is also capable of being tune. Processes of reliability is for failure and repair. It utilizes easy, expandable data model of data which enables internet based data API for research <sup>7</sup> as illustrated in (Fig. 4).

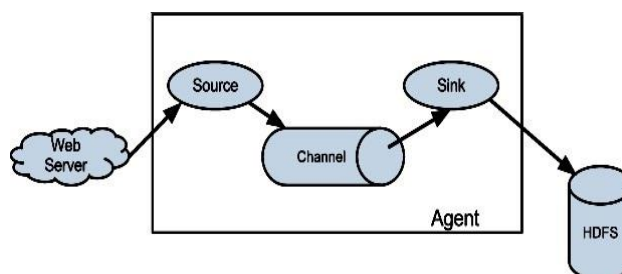


Figure 4. Flume Architecture <sup>7</sup>

A Flume occurrence is classified as a data flow unit having byte payload capacity, and the optional collection of attributes for strings. An Agent of Flume is a method for the Java Virtual Machine (JVM) Hosting the elements from which major events stream to the next destination from an external source .The source of Flume releases actions given to send it from an outside place, such as web application servers. The outside source is the external source which sends events in a format to flume that is recognized by Goal Source for Flume. When a source for Flume receives, it is stored in one or more channels for an event. The channel is a passive store that maintains the case until it is Consumed by a drain called Flume. The sink prevents the case, and moves it directly into the external source from the channel like Hadoop distributed file system or forward it refers to the next Flume data source, In the flow, flume agent working (next hop) Source and sink, respectively Run asynchronously with the agent inside the given agent. The channel staged events are conducted upon every agent and in a channel. The next available agent or terminal repository is then delivered into the stream (like HDFS) and deleted from one channel Just after storing them in the next agent's channel or site. They are in a repository for the terminal. That's how the single-hop operation is in Flume. Message delivery semantics have end to end delivery semantics. The stream reliability flume utilizes a transactional instrument method to ensure the efficient distribution of the accidents <sup>7</sup>.

### Apache Hive:

Pig is a language semantically similar to Pig Latin for scripting Hadoop and Hive are identical to regular SQL inquiries, such as queries



for Hadoop, which enables developers to write Hive Query Language Inside (HQL). For the developer who is recommended to Hive, who SQL is familiar with Initially Hive, which was created by it was later adopted by the Apache application on

Facebook, Base under the name and it is also open - source software Hive Apache. The Hive is suited for OLAP and is quick, scaling, Query Language and scalable<sup>6</sup> can be illustrated in (Fig. 5).

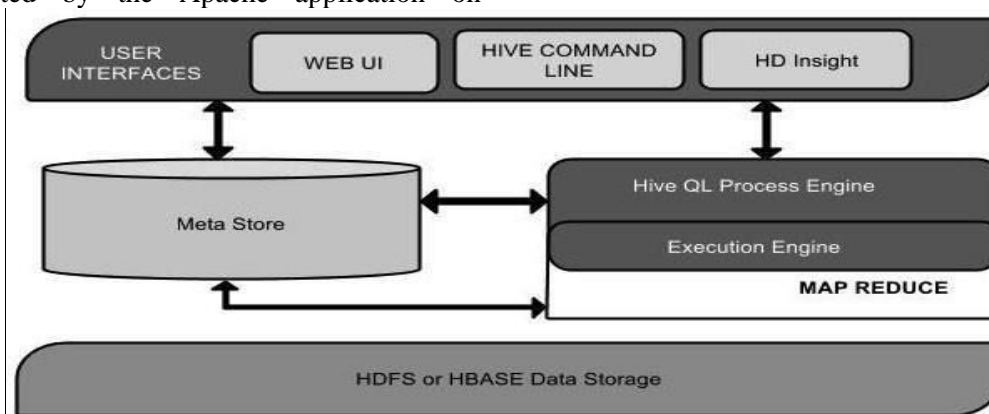


Figure 5. Hive Architecture<sup>6</sup>

**Hive work is explained by the following components<sup>6</sup>:**

1. User Interface: Three user interfaces are provided by Hive, like Web UI, Graphical User interface., and Perspective for Hive HD.
2. Store: For storage the tables' metadata or schema., Table columns, their data forms, and database HDFS map. at Hive, the server has been used.
3. HiveQL processing engine.: HiveQL is like SQL which is close to SQL, possible replacement Map Reduce's conventional method Uh Program.
4. Executing Engine: The engine is the way of processing the execution engine. Question and

produce outcomes comparable to Map Reduce Outcomes.

**Apache Sqoop:**

Sqoop is a method way for transmitting information Between Hadoop and servers for relational databases such as Oracle, MySQL. In short, sqoop is a way for importing relational information from Hadoop HDFS storage, and export information from the Hadoop File Relational Information Program<sup>6</sup> can be illustrated in (Fig. 6).

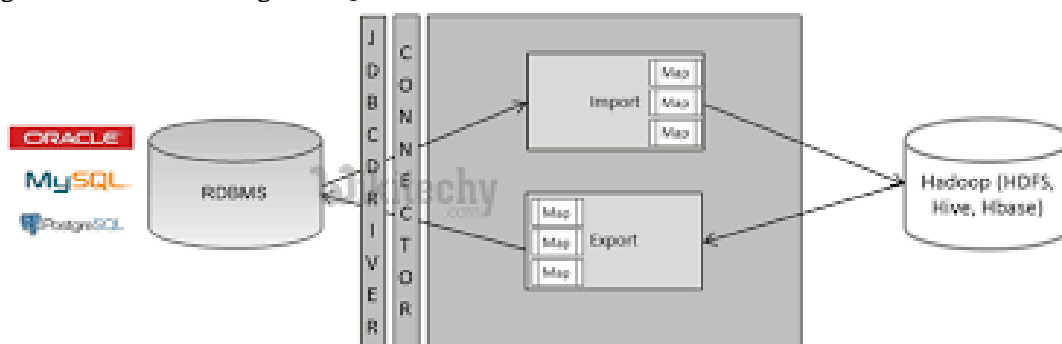


Figure 6. Sqoop Architecture<sup>6</sup>

**Apache spark:**

Spark is an effective open-source processor. It is designed across top speed, due to the ease use and advanced engine about analysis. In a number of ways, the spark engine works Environments, from Hadoop Clusters to cloud services. A range of famous growth is supported by Spark Scala, Python and Java languages, among others. Spark offers the appealing, elegantly designed Improvement API and makes it possible for data workers to quickly

through machine learning and data, iterate over data techniques in science which needs quick, in-memory information. Sorting 10-100 times quicker than Map Reduce, Spark is Providing speed time for perspective into more information, resulting in faster information and better management decisions and better outcomes. for consumers.

The cluster manager for handling clusters and distribution are needed by Spark System for Storage. Spark supports cluster management for

Hadoop YARN or Apache Mesos standalone. Spark is capable of the large range for multiple storage interfaces, Hadoop distributed file system, Cassandra, OpenStack Swift, and Cassandra S3 Amazon. Two main concepts are in Spark, Resilient Distributed Dataset and acyclic graph directed (DAG) Engine with execution.

RDD is a memory abstraction that is distributed. This enables computation in memory on large - scale distributed cluster centers with elevated fault-tolerance. There are two varieties of RDDs from Spark, Parallel collections dependent on current one's groups of programming like a set, a map, etc. and files. They're kept on HDFS. Two types of processes are carried out by RDD.

Transformations and behavior that are transformed produce new transformations Datasets from the current or input RDD (e.g. map or Map-Filter), and behavior to return a value after a filter is executed. Dataset calculations (e.g. decrease, collect, count, Save As TextFile, and so on).

Conversions are sloth's ones Activity that only determines the current RDD when performing actions that complete the real calculation and compute the outcome or save it to your external drive. Execution Engine Guided Acyclic Graph (DAG, whenever an action is executed on RDD by the user, a guided action Considering all the acyclic graphs, all of them are generated and dependencies for Change. This abolishes the Multi-stage execution model and conventional Map Reduce. It also increases performance <sup>7</sup>.

### Apache Hbase:

Hadoop has the restriction when that massive Data set quantities are stored, except for the many obvious tasks. One has to search for the whole dataset so only details can be accessed. The HBase databases that are used sequentially are the databases to store large quantities of data in a random way and access the data Style <sup>6</sup>.

### Mahout Offers:

Libraries implemented on top of Hadoop for scalable machine learning algorithms. It is used or implementations along with the Map Reduce system <sup>5</sup>.

### Zookeepers:

It is a centralized service that is used to preserve settings Hadoop cluster data along with distributed clusters Coordination and synchronization. It is also used to map the jobs with Hadoops <sup>5</sup>.

### Oozie:

This is a system for data processing, used as a process manager to handle the workers at Hadoop. It is possible to combine it with Other elements of Hadoop, such as Mapreduce, Hive, Pig, and Sqoop <sup>5</sup>.

### Ambari:

It offers a provisioning software platform, Managing and tracking clusters of Hadoops. It also offers a Web UI of Hadoop management to configure and monitor various different web UIs Hadoop service with the support of RESTful APIs <sup>5</sup>.

### Literature analysis:

Hadoop is an open source program that runs on distributed applications, storage of massive databases, and distributed computing. Doug The Doug in 2005, Cutting and Mike Cafarell produced Hadoop, published in Javen Specially designed to scale from a single server to a single server. Thousands of computers and algorithms function on Map Reduce Where data is processed on various CPUs simultaneously, Nodes.

Marouane Birjalía et al <sup>8</sup> Presents how to resolve the constraints of conventional methods using the Hadoop ecosystem to streamline the processing of large cluster data. Which are primarily collected from social networks by Apache Flume unstructured format and the JAQL script used to extract important data, converting them into a simpler structure to convert the delimited file by commas, and stored in Hadoop storage to perform the processing using Map Reduce. InfoSphere techniques are used to analyze big data in real time to feeling. InfoSphere's main include a significant range of (IBM) technologies enhancing Hadoop performance.

Aditya Bhardwaj et al <sup>3</sup> removed on-site operational difficulties hardware investment, IT support, and installation, configuration of Hadoop components such as HDFS and MapReduce, the Cloud-based Hadoop cluster environment was implemented for the processing of Big Data. Microsoft Azure cloud services have been used to incorporate multiple MapReduce jobs such as Pi, TeraSort, WordCount on cloud-based Hadoop deployment. MapReduce job performance has been measured with respect to CPU execution time with varying Hadoop cluster size. From the experimental outcome, as the amount of data nodes in the HDInsight cluster increases, it is found that CPU execution time to finish the jobs decreases and suggests the good response time with better efficiency as well as greater customer satisfaction. So it would be a safe choice for Big Data initiatives to install Hadoop on the cloud computing platform.

Binod Kumar Adhikari et al <sup>9</sup> Built a system for sensitive data detection from voluminous data varieties that meets the shortcomings and challenges of current methodologies. The data obtained from the network in real time is stored in cloud drives, other storage devices and then transferred to the distributed file system of Hadoop and processed using distributed computing concepts, machine learning algorithms and advanced statistical methods with MapReduce processing. Using distributed computing and Hadoop clusters, big data analytical techniques were introduced to extract confidential information from a wide scale of data. Using ANOVA, Parametric Levene's test, Pearson Correlation, and Kruskal-Wallis Test, the result was statistically analyzed and it demonstrated that big data analytical methods are more applicable than conventional statistical methods to obtain sensitive information.

Divya Sehgal and Dr. Ambuj Kumar Agarwal <sup>10</sup> To sentiment analysis in the twitter data, which is also known as big data, they used HADOOP. This project is also applied to the social media site and movie review such as blogs and comments and likes per day and study and improves the efficacy of data analysis to assess the importance of sentiment. The accuracy is entirely worth observing. For this project, using hash tags and emoticons is also very helpful and important for social media data. This strategy was mainly based on Speed analysis of results and even precision Complete output of the study of sentiment in the context of the Big data is data which also eliminates the common ones, only big data technology concerns are coming up. In this one, it is an approach that performs sentiment analysis on big data. Achieved by separating the various modules and the list of Production of the next steps in coordination with for mapping it on various computers, HADOOP. It is easy to open a portion of speech and tagged using opennlp, it is easy to open to solve the problem.

Mudassir Khan and Aadarsh Malviya <sup>11</sup> presented sentiment analysis technique by adapting a Hadoop framework and classifier for deep learning. The important characteristics of the extraction process, such as all-caps, emoticon, hashtag, elongated units, sentiment lexicon, negation, and punctuation, are extracted using Twitter data input. In the shuffle, where lists of specific features are picked, the obtained features are then fed in. The reducer fed the list of specific characteristics. The features are categorized using deep recurrent neural in the reducer step. The difficulty lies in evaluating the sentiment of the

tweets due to the specific characteristics of Twitter data, which classifies the characteristics into two groups: namely positive review and negative review. Performance analysis is conducted utilizing measures such as accuracy of classification, sensitivity and specificity. The proposed approach gave improved accuracy of classification of 0.9302, greater sensitivity in comparison to classical strategies. The suggested deep RNN(Recurrent neural networks) method based on Hadoop produced maximum precision, sensitivity and specificity of 0.9302, 0.9404 and 0.9157, respectively.

Bharat Tidke et al <sup>12</sup> proposed and introduced new architecture for user clustering using sentiment value and similarity measures based on stream data from the online social networking platform. For several applications, such as community identification, behavioral similarity detection and group recommendation, this approach can be used. The objective is to efficiently manage broad social data using cost-effective fetching tools and to query unstructured data and algorithms to analyze scalable, uninterrupted data streams with finite memory and resources. Using Apache Flume, the authors collect streaming tweets from the Twitter API to detect user clusters with similar feelings. The proposed solution uses a scalable and fault tolerant framework (i.e., Hadoop) that usually uses HDFS for data storage. Data processing paradigm and map-reduce. Apache Hive is used to operate on top of Hadoop and to use the AFINN(Finn Årup Nielsen) dictionary to evaluate Tweet Sentiment and to get similarity between Tweets implemented in the programming model Map reduce java software.

Devika Harikumar1 et al <sup>13</sup> A framework for filtering and evaluating posts on YouTube was proposed to delete sensitive content from comments before posting harsh comments. Using Hadoop, MapReduce and YouTube API, they used the NLP (natural language processing) parser technique to define the sensitive characteristics and proposed technique. In order to maximize time consumption, linear search methods are used. Authors believe that the video can be evaluated for sensitive information in future work by looking at the vocabulary used in the video. They also have a way to increase the speed of analysis even further.

Chetna Dabas et al <sup>14</sup> introduced a method for the thorough review and classification of YouTube video comments. In terms of execution time for the queries designed as a part of this work, the proposed framework has yielded promising results. The submitted application uses the multinomial Naïve Bayes classifier and analysis to

identify the comments on YouTube videos. Analytical language using Hive. The findings obtained are Via graphs and charts visualized. Additionally, to render the A more user-friendly Graphical User Interface framework (GUI) is designed using the Tkinter library in Python. As a result, analyzing this feedback plays a key role in optimizing video content to satisfy the need for time. In particular, comments will greatly help a new user by forming an educated and definitive opinion on a specific video.

ERRAIS Mouhssine and Choug dali Khalid<sup>2</sup> proposed a system for polling and uncovering online social media data for extremist content. Using this system, machines can learn how to extract group messages from public Facebook pages automatically, use API graph calls, filter out messages without opinion, classify their feelings about interest patterns (i.e. positive and negative) and the aim of this model is to create a big data application that gets a stream of public data from the social media network of Facebook, which can help law enforcement and cybercrime analysts evaluate and monitor social media in the search for digital monitoring of violence or extremism, which can be used in more digital forensic investigations. In this context, Hadoop with Apache Hive, Apache HBase.

Parmeet Kaur et al<sup>15</sup> intended to do a comparative study through MongoDB and Hive for the analysis of news results. With limited resources and large-scale unstructured data, MongoDB has resulted in less execution time than Hive for the execution of similar queries. The data is maintained in the Hadoop ecosystem in a NoSQL database, MongoDB and the Hive data warehouse. For both MongoDB and Hive, the analysis was performed on a single computer device. The data is shared (partitioned) in MongoDB as well as partitioned in Hive using the date field as the sharing key and the partitioning key in the respective databases for better data handling and fast query processing.

Farhan Amin et al<sup>16</sup> proposed a new term 'Socio-Cyber Physical Network' architecture for identifying human activities using big data analytics, proposing a device architecture that combines social network with the technological network. Innovative tools that address the challenges posed by big data volume, variety, and velocity are required for the growing gap between

users and big data analytics. The proposed framework uses a graph definition that helps the devices find a better fit and can forecast the future. Network architecture output fulfills the needs of the users associated with it, whether the input data is both real-time and offline when performing real-time actions.

Alireza Ashayer et al<sup>17</sup> concentrated on the effectiveness and efficiency with using a Hadoop cluster to evaluate user activity depending on their behaviors in systems with a huge number of user which are processed in an in-memory manner for quicker processing and querying than disk storage of data stored in computer memory. Due to its specific features and similarities in present issues in academia and manufacturing, sentiment analysis development as a base model in performance analysis, performance analysis in both related areas is included. The cluster's function in larger datasets and subsequent analysis. It has been shown that the number of static allocation executors does not affect output in our cluster. It has also been shown that using more than one thread for every input stream would lose the resources of the device and reduce output. The consequences of message propagation between nodes were analyzed and found that the tend to increase in the replication process factor could have an outsized impact, the cluster's network capacity. More precisely, an increase in the replication factor of 50 percent raises network load by 110 percent.

Pushpita Ganguly<sup>18</sup> mainly concentrated on the methods and comparative studies available for big data analysis. When running on multiple nodes in a shared cluster, these tools should follow proper scheduling algorithms. Fair approaches to synchronizing multi-node activities in distributed environments need to be provided. But this, obviously, Apache Spark, which improves the feat not only by processing powerful high-level functionalities in memory, must be laid out on a more sophisticated platform. Apache Spark is also able to solve the weak problems of MapReduce and is also more suitable for the multi-iteration platform needed for the data flow of the next generation, as seen in the following comparison Tab.1 View job scheduling algorithms that analyze their characteristics, strength and weakness in each system.



**Table 1. Comparison between MapReduce and Apache Spark**

MapReduce	Apache Spark
In multi-operational or iterative processing, MapReduce is not sufficient. It stores data on the hard disk and uses the same data set twice, making it slower compared to Apache Spark.  As memory size is not constrained, in batch processing, map reduce is faster than Spark.  In real time data processing, it is difficult to perform due to high inactivity map reduce. Replication is used by Map reduce to conduct fault tolerance  If the loop stops in the centre, it will still take up memory in Map reduce.  Users need to focus on other resources for stream processing, database querying, machine learning, since they mostly use a batch processing technique.  Map Reduce carries out a lot of I/O operations that reduce latency	The highly suitable Apache Spark for iterative processing. It saved the data for quick computation in memory, RAM-based processing so that it is 100 times faster than MapReduce It is not that good for larger data sets in the case of Apache Spark, because it is slower than map reduce in batch processing. Apache Spark is sufficiently effective for real-time implementation To accomplish fault tolerance, Apache Spark uses RDD (Resilient Distributed Datasets). Apache Spark has a special lazy assessment function that will not store data before the final process occurs. All in one approach, Apache Spark offers spark streaming for stream processing, Spark sql for database querying, MLlib for machine learning, which refers to this methodology's flexibility. Apache Spark contains extremely less I/O tasks

Ashwitha T A et al <sup>19</sup> presented the Using Hadoop platform; Hive tool is used with the Hadoop framework for the study of the movie dataset. The datasets are derived from the IMDb website. The dataset includes releasing year information, title, language, iamb ratings, FB likes, genre, etc. Four datasets for study have been considered by authors in this paper as Movie dataset. It is impossible and difficult to store such an enormous amount of information in conventional data warehouses. Hive queries have been performed and its response time has been contrasted with SQL response time. For smaller datasets, MySQL is well equipped. As the size of the dataset increases, it might take longer to process it and more memory might also be needed for processing. But in the case of Hive, for larger datasets, it is better suited. The experiment showed that the response time in MySQL is longer. Therefore, using Hive for large datasets is more effective. Compared to conventional systems, the processing time for analyzing datasets has been substantially increased. Hive is primarily used for enormous volumes of data processing. On a low-cost system, it's quicker than SQL.

Cameron Seay et al <sup>20</sup>, Big Data technology such as Hadoop are transforming analytics and processing, but there is a role for mainframe to play in enhancing big data analytics and exploring some possible mainframe benefits and its advanced virtualization capability will provide analytics with Hadoop to access the vast benefits seamlessly Big data tools such as Hadoop can be used by the

mainframe to offer innovative solutions in order to address the continuous influx of data that comes in various formats, particularly semi-structured or non-relational data.

Experts in big data analysis who are familiar with the use of standard methods. analysis, such as Hadoop, have been identified. This work would offer an improved access to knowledge recently covered from those in the industry. Using mainframe Hadoop will allow you to reduce analytical computing costs and significantly increase enforcement by not making unnecessary duplications Off the source (mainframe) network, the info. Hadoop also removes the bottleneck at the user's interface - their processor(s) - as well as why there is a network bottleneck because the data is not transmitted from multiple servers to the user's interface - as well how the bottleneck. is eliminated through the use of Hadoop and does some pre-processing of the data needed before it is sent to the analyst's workstation.

João Cunha <sup>21</sup> proposed a generic functional architecture for managing, storage and analyzing broad data for use in various variety of fields with the Apache Hadoop system and Mahout. It concludes that the forth coming Twitter data must be carefully analyzed MAHOUT has many variety of classification algorithms such as word count and hash count. However, some may be added to improve its classification efficiency due to certain semantics or sentences. Even Mahout provides a library of algorithms designed to solve various problems, and Hadoop is capable of parallel

processing. Even the punctuation of sentences such as exclamation or citation-like terms or phrases in sentences change the classification of the sentence, interfering with the efficiency of the classification algorithms used by Mahout, but other clustering algorithms are also successful. But Mahout does not really have any mechanism to store the top terms Healthcare agencies, hospital networks and many others will be able to examine the broad tweet datasets. To achieve effective enhancements. Potential advantages include the diagnosis of illnesses at earlier stages, allowing prevention-based care to be treated more efficiently and effectively. Any trends or effects can be forecast or estimated on the basis of vast quantities of historical data and Focused on people's tendencies. Such analytics may help to minimize waste and inefficiency in various areas of health.

WenTai Wu et al <sup>22</sup> presented insights and described potential research directions, including energy efficient cluster partitioning, data oriented resource classification and provisioning, optimal utilization based resource provisioning, EE and locality conscious task scheduling, optimization of job profiling with machine learning, elastic containerized power saving Hadoop and efficient big data analytics Review the studies on enhancing the energy performance of Hadoop clusters and summarize them in five sections, including the management of energy aware cluster nodes, energy aware data management, allocation of energy aware resources, scheduling of energy aware tasks and other energy saving schemes. Hadoop defects are also revealed in several areas, including data management, resource management, scheduling, scheduling. By optimizing several parameters, such as the number of active) servers, data movement, server usage, etc. the energy efficiency of a Hadoop system can be achieved. Nevertheless, it is also commonly adopted to estimate or forecast device power or cost as a decision predictor. A predefined model with several parameters is required for the second form.

Awais Ahmad et al <sup>23</sup>, the proposed system architecture focuses on the environment analysis generated by Smart Cities, wearable devices (e.g. body area network) and Big Data to assess human activities and dynamics of humans. This makes the world smarter and provides an intelligent space to sense our behaviors or behavior, and the ecosystem's evolution. The developed framework carefully tackles the comprehension challenge by providing users with input that provides them the opportunity to enhance their actions using the warning message taxonomy, i.e., white zone, gray zone, and red zone. The aim is to explain human

actions in real-time in the social sphere. Through the amount of knowledge generated by smartphones, social networks, and smart cities, these aims are beginning to be feasible. The similarity between different data attributes is exploited by Big Data in SIoT and promotes the interpretation and description of human behavior. The idea of 'defining human behavior using Big Data' was raised through careful investigation of citizens' inclusion and engagement with the evolving Smart Cities and demonstrated its applicability using the Hadoop ecosystem.

### Conclusion and Future work:

Hadoop is commonly used with map reduce technique for the processing of big data. But, therefore, in this paper, some particular areas where map reduce cannot be considered as an acceptable option unfit for real-time processing and unfit for processing graphs are also unfit for OLTP (online transaction processing), also The RDBMS was not able to handle unstructured data and total time taken for executing query, challenges of on-site hardware investment, IT support, and installing, configuring of Hadoop components such as HDFS and Map Reduce. also the semantics of sentences may change the meaning of words.

Therefore, to overcome this challenges must focus on used data mining algorithm on Hadoop framework and the definition of big data with used addition some tools to work with Hadoop such as apache spark, Flume, Kafka...etc. Which help in enhancement performance and computation time and overcome on most challenges and difficult implementation and execution algorithm to extend specific Hadoop development tools.

At this point, in the future, the interest should be on investigating the same path, during a universal integrated framework for dealing with all issue of big data can be applied to various application and Hadoop deployment on cloud computing platform will be good choice for the big data project.

### Authors' declaration:

- Conflicts of Interest: None.
- We hereby confirm that all the Figures and Tables in the manuscript are mine ours. Besides, the Figures and images, which are not mine ours, have been given the permission for re-publication attached with the manuscript.
- Ethical Clearance: The project was approved by the local ethical committee in Information Institute for postgraduate studies.

### Authors' contributions:

The idea of the search was by Prof. Dr. Abbas Fadhil Aljuboori and supervise it. The executed this idea by researcher wathaq ghanim mutasher with help and support Dr. Abbas.

Our goal to explain and prove from during related works that ecosystem Hadoop have more tools to analysis big data. Some tools deal with real-time data and map reduce can't deal with it efficiently. unlike other tools which can deal with data in both offline and real-time with high efficiency such as apache spark also advice researcher to used cloud to overcame on running problem and infrastructure as hardware and reduce processing time because environment Hadoop work in parallel processing.

He also provided us with analytical results based on the description we obtained from a review of the references, so that the researcher might save time and effort in his tasks by using big data technologies.

### References:

1. Gole S, Tidke B. A survey of big data in social media using data mining techniques. ICACCS - Proc 2nd Int Conf Adv Comput Commun Syst. 2015;5-10.
2. Mouhssine E, Khalid C. Social Big Data Mining Framework for Extremist Content Detection in Social Networks. Int Symp Adv Electr Commun Technol ISAECT 2018 - Proc. 2019;1-5.
3. Bhardwaj A, Singh VK, Vanraj, Narayan Y. Analyzing BigData with Hadoop cluster in HDInsight azure Cloud. 12th IEEE Int Conf Electron Energy, Environ Commun Comput Control (E3-C3), INDICON 2015. 2016;
4. Monika, Bhat A. An analysis of Crime data under Apache Pig on Big Data. Proc 3rd Int Conf I-SMAC IoT Soc Mobile, Anal Cloud, I-SMAC 2019. 2019;330-5.
5. Jadhav B, Patankar AB, Jadhav SB. A Practical approach for integrating Big data Analytics into E-governance using hadoop. Proc Int Conf Inven Commun Comput Technol ICICCT 2018. 2018;(Icicct):1952-8.
6. Bhardwaj A, Vanraj, Kumar A, Narayan Y, Kumar P. Big data emerging technologies: A CaseStudy with analyzing twitter data using apache hive. 2015 2nd Int Conf Recent Adv Eng Comput Sci RAECS 2015. 2016;(December).
7. Farhan MN, Habib MA, Ali MA. A study and Performance Comparison of MapReduce and Apache Spark on Twitter Data on Hadoop Cluster. Int J Inf Technol Comput Sci. 2018;10(7):61-70.
8. Birjali M, Beni-Hssane A, Erritali M. Analyzing Social Media through Big Data using InfoSphere BigInsights and Apache Flume. Procedia Comput Sci [Internet]. 2017;113:280-5. Available from: <http://dx.doi.org/10.1016/j.procs.2017.08.299>
9. Adhikari BK, Zuo W, Maharjan R, Han X, Amatya PB, Ali W. Statistical analysis for detection of sensitive data using hadoop clusters. Proc - 21st IEEE Int Conf High Perform Comput Commun 17th IEEE Int Conf Smart City 5th IEEE Int Conf Data Sci Syst HPCC/SmartCity/DSS 2019. 2019;2373-8.
10. Sehgal D, Agarwal AK. Sentiment analysis of big data applications using Twitter Data with the help of HADOOP framework. Proc 5th Int Conf Syst Model Adv Res Trends, SMART 2016. 2017;V:251-5.
11. Khan M, Malviya A. Big data approach for sentiment analysis of twitter data using Hadoop framework and deep learning. Int Conf Emerg Trends Inf Technol Eng ic-ETITE 2020. 2020;1-5.
12. Tidke B, Mehta R, Rana D, Jangir H. Topic sensitive user clustering using sentiment score and similarity measures: Big data and social network. Int J Web-Based Learn Teach Technol. 2020;15(2):34-45.
13. Harikumar D, Kapoor D. Youtube Data Sensitivity and Analysis Using Hadoop Framework. Int Res J Eng Technol. 2019; 06 (04): 3133-3139
14. Dabas C, Kaur P, Gulati N, Tilak M. Analysis of Comments on Youtube Videos using Hadoop. Proc IEEE Int Conf Image Inf Process. 2019;2019-Novem:353-8.
15. Kaur P, Dabas C, Singhal V, Nangru S, Sehgal A. News Data Analysis from Facebook Through MongoDB and Hive. Proc IEEE Int Conf Image Inf Process. 2019;2019-Novem:454-8.
16. Amin F, Ahmad A, Choi GS. To Study and Analyse Human Behaviours on Social Networks. Proc 4th Annu Int Conf Netw Inf Syst Comput ICNISC. 2018;233-6
17. Ashayer A, Yasrobi S, Thomas S, Tabrizi N. Performance Analysis of Hadoop Cluster for User Behavior Analysis. Proc - 20th Int Conf High Perform Comput Commun 16th Int Conf Smart City 4th Int Conf Data Sci Syst HPCC/SmartCity/DSS 2018. 2019;805-9.
18. Ganguly P. Big Data Analytics : Using Hadoop Inspired MapReduce and Apache Spark. Int. J. Adv. Sci. Technol. 2020;7(2):72-82.
19. Ashwitha TA, Rodrigues AP, Chiplunkar NN. Movie Dataset Analysis Using Hadoop-Hive. 2nd Int Conf Comput Syst Inf Technol Sustain Solut CSITSS 2017. 2018;1-5.
20. Seay C, Agrawal R, Kadadi A, Barel Y. Using hadoop on the mainframe: A big solution for the challenges of big data. Proc - 12th Int Conf Inf Technol New Gener ITNG 2015. 2015;765-9.
21. Cunha J, Silva C, Antunes M. Health Twitter Big Bata Management with Hadoop Framework. Procedia Comput Sci [Internet]. 2015;64:425-31. Available from: <http://dx.doi.org/10.1016/j.procs.2015.08.536>
22. Wu WT, Lin WW, Hsu CH, He LG. Energy-efficient hadoop for big data analytics and computing: A systematic review and research insights. Futur Gener Comput Syst [Internet]. 2018;86:1351-67. Available from: <https://doi.org/10.1016/j.future.2017.11.010>
23. Ahmad A, Rathore MM, Paul A, Rho S. Defining human behaviors using big data analytics in social internet of things. Proc - Int Conf Adv Inf Netw Appl AINA. 2016;2016-May:1101-7.

## الطرق الحديثة والحالية لتحليل البيانات الضخمة باستخدام أدوات هادوب

عباس فاضل الجبوري<sup>2</sup>

واثق غانم مطشر<sup>1\*</sup>

<sup>1</sup> معهد المعلوماتية للدراسات العليا، الهيئة العراقية للحاسبات، العراق.

<sup>2</sup> جامعة تكنولوجيا المعلومات والاتصالات، العراق.

### الخلاصة:

الجميع متصل بوسائل التواصل الاجتماعي مثل (الفيس بوك وتويتر ولينكدان والانستغرام... الخ) ، التي تتولد من خلالها كميات هائلة من البيانات لا تستطيع التطبيقات التقليدية من معالجتها ، حيث تعتبر وسائل التواصل الاجتماعي منصة مهمة لتبادل المعلومات والآراء والمعرفة التي يجريها العديد من المشاركين ، على الرغم من هذه السمات الأساسية ، تساهم البيانات الضخمة أيضاً في العديد من المشكلات ، مثل جمع البيانات ، والتخزين ، والنقل ، والتحديث ، والمراجعة ، والنشر ، والمسح الضوئي ، والتصوير ، وحماية البيانات ... الخ. للتعامل مع كل هذه المشاكل، ظهرت الحاجة إلى نظام مناسب لا يقوم فقط بإعداد التفاصيل، بل يوفر أيضاً تحليلاً ذا مغزى للاستفادة من المواقف الصعبة، سواء ذات الصلة بالأعمال التجارية، أو القرار المناسب، أو الصحة، أو وسائل التواصل الاجتماعي، أو العلوم، الاتصالات، البيئة... الخ. يلاحظ المؤلفون من خلال قراءة الدراسات السابقة أن هناك تحليلات مختلفة من خلال Hadoop وأدواته المختلفة مثل المشاعر في الوقت الفعلي وغيرها. ومع ذلك، فإن التعامل مع هذه البيانات الضخمة يعد مهمة صعبة. لذلك فإن هذا النوع من التحليل يكون بكفاءة أكثر كفاءة فقط من خلال نظام Hadoop البيئي، الغرض من هذه الورقة هو تحليل الأدبيات المتعلقة بتحليل البيانات الضخمة لوسائل التواصل الاجتماعي باستخدام إطار Hadoop لمعرفة أدوات التحليل تقريباً الموجودة في العالم تحت مظلة Hadoop وتوجهاتها بالإضافة إلى الصعوبات والأساليب الحديثة لها للتغلب على تحديات البيانات الضخمة في المعالجة غير المتصلة وفي الوقت الفعلي. تعمل التحليلات في الوقت الفعلي على تسريع عملية اتخاذ القرار إلى جانب توفير الوصول إلى مقاييس الأعمال وإعداد التقارير. كما تم توضيح المقارنة بين Hadoop و spark.

**الكلمات المفتاحية:** اباتشي سبارك، البيانات الضخمة ، انترنيت الاشياء ، هادوب، وسائل التواصل الاجتماعي .