REVIEW: BIG DATA V'S MODELS, CHALLENGES, HADOOP ECOSYSTEM, ISSUES, USES, BENEFITS AND APPLICATIONS

M. A. Raza[1] , H. U. R. Kayani[2], Muhammad Gulraiz Zahid Awan[3], M. Gul[4], A. Suleman[5], M. T. Aslam,[6]

[12456] The University of Lahore, Sargodha Campus, Pakistan
Department of Computer Sciences, The University of Lahore, Sargodha Campus Pakistan
University of Sargodha, Pakistan, The University of Lahore, Sargodha Campus, Pakistan
The University of Lahore, Sargodha Campus, Pakistan
[3] Electrical Engineering Department, National University of Computer and Emerging Sciences, Peshawar, Pakistan
Corresponding Author Email: mindshacker007@gmail.com

ABSTRACT: Big Data, encompassing colossal volumes of archival and consequential information, stands as the greatest invaluable treasure for any establishment, holding the potential to fortify business decisions grounded on factual evidence rather than mere perceptions. The most sophisticated method of defining big data is 56V's, provided in the paper. In this work, we have defined extensive data using the V technique, beginning with the 3V's and moving on to the latest definitions with the 5V's, 7V's, 10V's, and 14V's. Undoubtedly , forthcoming technological and corporate efficiency contests will amalgamate into extensive information exploration. The Hadoop cloud computing platform, which is open source and developed by the Apache Foundation , is also included in this study . It features a distributed system known as HDFS and MapReduce software programming platform . The leading big data technologies include Hadoop, Map Reduce, YARN, Hive, Flume, Apache Spark, and No SQL. The handling of massive data can significantly benefit from these technologies . This document also summarizes these tools' capabilities , advantages , and disadvantages . The concluding section of this research paper also delves into utilizing large-scale data in diverse sectors such as banking , finance , education, healthcare, and agriculture.

Keywords: 5V's, 7V's, 10V's, 14V's 56V's, BDA, Hadoop Framework, MapReduce, Ecosystem, Apache Spark, Healthcare, Education, Agriculture, Smart Cities.

INTRODUCTION

The latest and most advanced concept of Big Data is currently in utilization; everyone likes to talk about it, and it frequently appears in the media, and businesses work to take advantage of the increased amount of information at their disposal. Big data denotes expansive and intricate data assemblages that pose challenges when scrutinizing them using conventional data processing methodologies. The world of massive information is a critical topic in the IT business. Nurturing the big data giant needs unique techniques, as traditional security and confidentiality protections must be improved in the face of complex distributed computing across varied data kinds. Each new type of data has its own set of challenges and mysteries. Countless scholarly endeavors have begun to untangle the complexity of big data, beginning with Doug Laney's seminal manuscript and over the last two decades. At its heart is the enormous issue of managing massive data volumes while assuring their safe passage through the vast expanse of the internet and arriving at their destination uninjured. The "3Vs," "5Vs," "7Vs," "14Vs", and "56Vs" are commonly used abbreviations for the features of extensive data and are as follows.

The first 3V interpretation of big data refers to these three fundamental elements depicted in Figure 2
Volume: The data is so large that it cannot be stored or processed on a single computer.
Velocity: The data is generated and collected at a high rate in real-time.
Variety: The data gathered are structured, semi-structured, and unstructured.

Another aspect of Big data definition is to cover it with five Vs. According to a widely accepted definition, big data is distinguished by five features, 1: Large volumes of data which are being produced at a fast rate 2: a wide variety of data that cannot be stored in conventional relational databases 3: Data is processed at very high velocity 4: cost-effective value mining requires experienced data mining solutions; 5: data veracity might affect analysis accuracy, as shown in

Volu
Velocity
5V
Veracity

Figure 1: 5Vs of Big Data

Volu
3V
Velocity

Figure 2: 3Vs of Big Data

Viscosity
Volu
Velocity
10V
Validity
Value
Vis
zati

Figure 3: 10Vs of Big Data

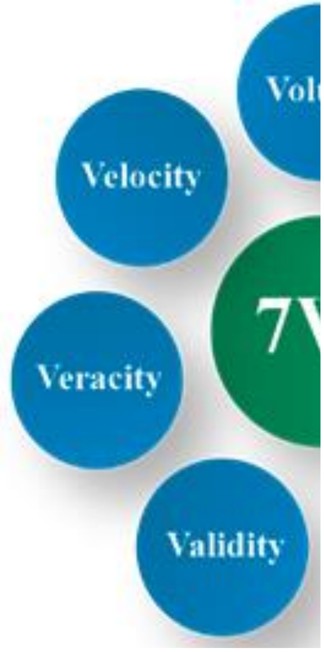

Figure 4: 7Vs of Big Data

(ZHANG Yaoxue1, 2017)

With the rising growth of data, the scientist proposed the “7V” definition of big data.

IDC developed the four V's model in 2011 as big data technology advanced. Further developments have allowed scientists to access 10V’s model of Big Data, is shown in

Table **1** shows the 14Vs and an explanation of each V. (Gayatri Kapil).

Figure 1: 5Vs of Big Data

Figure 2: 3Vs of Big Data

Figure 3: 10Vs of Big Data

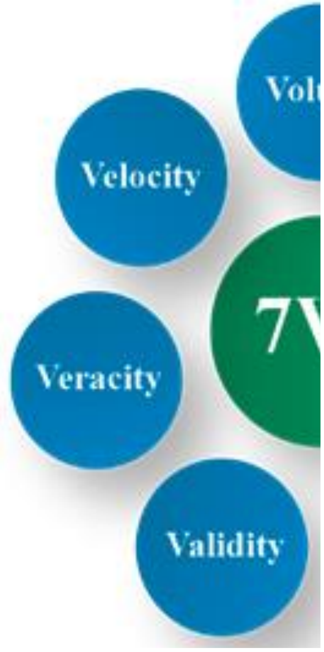

Figure 4: 7Vs of Big Data

. (Chaha, 2020)

The fundamental research on big data revolves around a set of 14V's, aiming to govern and leverage the vast amounts of data available effectively. Below,

Table 1: Big Data 14 Characteristics.

Sr. No	Big Data Characteristics	Explanation	Description	
1		Volume	Related to data size	The cornucopia of amassed and preserved information. The magnitude of data sprawls across terabytes and petabytes, an ocean of knowledge.
2		Velocity	Deals with data speed	The data transmission speed between the source and destination.
3		Value	Data significance	It succinctly signifies the potential economic worth that can be extracted from the vast volumes of big data.
4		Variety	Deals with data type	Various forms of data, such as images, videos, audio, and more, are received at the destination.
5		Veracity	Explains about quality	Authentic analysis of captured data is virtually useless if it's not accurate
6		Validity	Deals with Authenticity	The exactness and fidelity of the data utilised in the derivation of outcomes in the shape of valuable and significant information.
7		Volatility	Duration of Usefulness	It refers to the persistence of stored data and its duration of relevance and usefulness to the end user.
8		Visualisation	Data act	It entails the procedure of depicting abstract concepts or ideas.
9		Virality	Spread Speed	It is characterised as the ratio by which data is disseminated or propagated via user and subsequently received by multiple users for their purposes.
10		Viscosity	Lag of Event	It represents the temporal disparity between the occurrence of an event and the description of said event.
11		Variability	Differentiation of data	Data continuously streams in from diverse sources, and its effective discrimination between little noise and significant information is of paramount importance.
12		Venue	Platforms of various types	Many data types are received from distinct sources through various platforms, including personnel systems, private and public clouds, and others.
13		Vocabulary	Data Terminology	Data-related terminologies such as data models, data structures, and related concepts are commonly encountered in the field.
14		Vagueness	The indistinctness of existence in a Data	Vagueness pertains to ambiguity or lack of precision in information, indicating insufficient consideration regarding the intended meaning conveyed by each element.

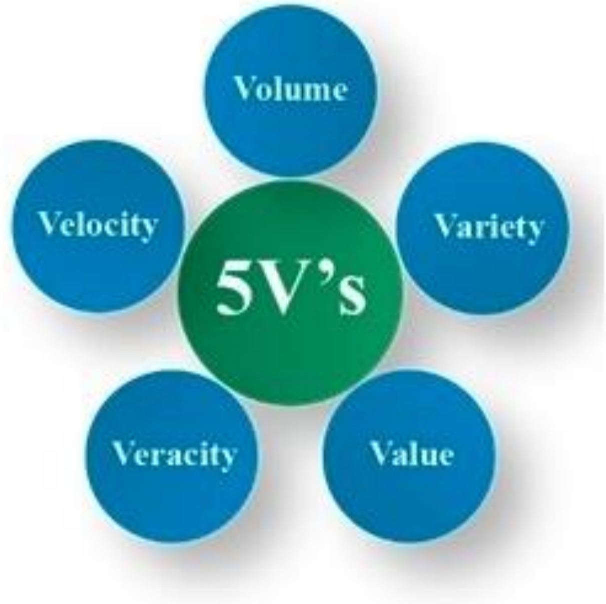

Figure 1: 5Vs of Big Data

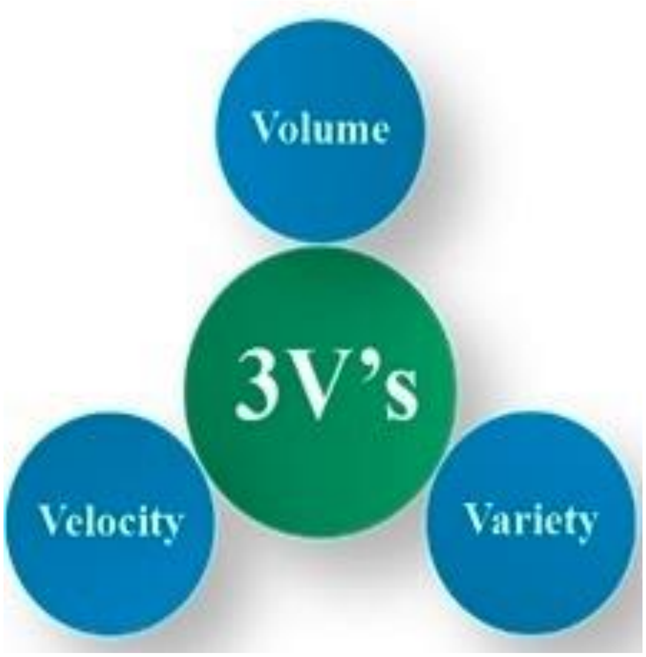

Figure 2: 3Vs of Big Data

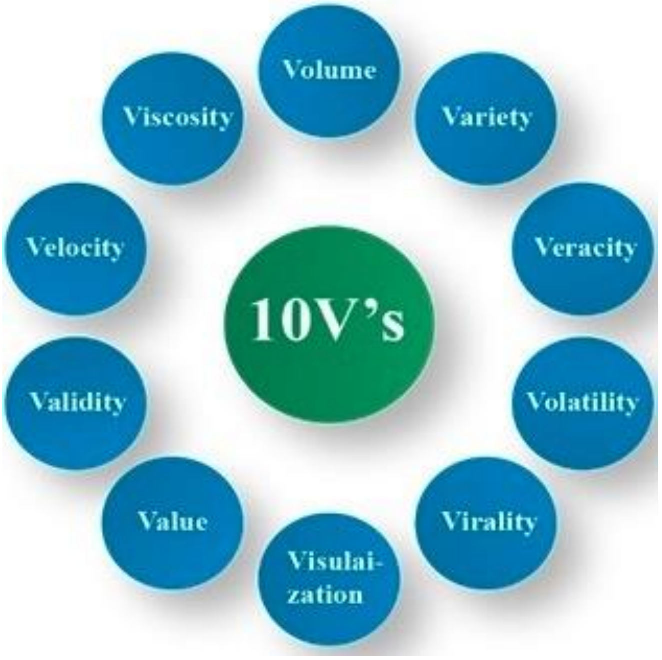

Figure 3: 10Vs of Big Data

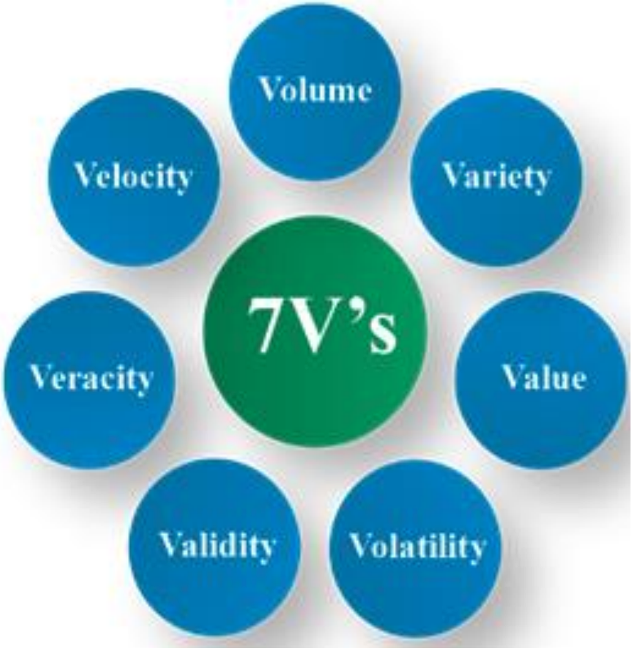

Figure 4: 7Vs of Big Data

METHODOLOGY

This academic research will use an investigation framework with a pair of methods. The first technique will highlight critical concepts relevant to the fifty-six Big Data V's defining attributes. Meanwhile, the second way will involve thoroughly studying scholarly publications using a widely accepted systematic literature review methodology. This endeavour aims to identify the famous tools with their comparison and concrete benefits provided by the enormous world of big data in the magnitude with which it provides apparent benefits.

56-V's attributes: The 3 Vs (Variety, Velocity, and Volume) are the first features of big data that many researchers pay close attention to, which prompts them to add a few additional Vs to the definition of big data. Below, Table ***2*** describes the 56V characteristics of big data.(Hussein, 2020)

Table 2: Big Data 56 Characteristics

Sr. No	Characteristics	Description
1	Volume	In the shape of data or information that may originate from all conceivable sensors, records, numerous watch times of YouTube videos, and tons of PB from worldwide mobile traffic, numerous companies already possess a significant volume of conserved 'Sea of data. (Hussien, 2020)
2	Variety	Large-scale data consists of diverse formats and data types, encompassing organised, partially organised, multi-structured, and predominantly non-structured data. Due to its heterogeneity in terms of both magnitude and nature, it cannot be integrated into a relational database.(Hussien, 2020)
3	Velocity	It may be referred to as the speed at which data flows or travels from one place to another, as well as the speed at which it is processed when data and information are received at a rapid rate. (Hussein, 2020)
4	Veracity	Does the data come from a valid source? Is it a crucial issue that requires a clear, definitive response? (Hussein, 2020)
5	Validity	The data must be accurate, consistent, precise, reasonable, and true for its desired purposes. (Hussein, 2020)
6	Value	Big data may be meaningless and unusable if a large amount of data is not converted into value. (Hussein, 2020)
7	Variability	The environment translates the variability in the data and requires deviation and aberration detection techniques before any relevant analytics can be performed. (Firican, 2017; Hussein, 2020)
8	Venue	Massive amounts of data can be differentiated by its heterogeneous distributed data from other platforms, from multiple owners' systems, with varying formatting and access requirements, whether confidential or public. (K. Borne, 2014; Hussein, 2020)
9	Vocabulary	Metadata encompasses data models, schemas, meanings, taxonomies, and other elements that define the arrangement, syntax, substance, and source of the data. (K Borne, 2014)
10	Vagueness	Regardless of the abundance of available data and the extent of its unambiguity, the interpretation of the data is often profoundly uncertain. (K. Borne, 2014)
11	Vulnerability	Since no system is flawless, some flaws will likely exist in the hardware or software. This implies that any linked data is vulnerable to theft or manipulation. (Firican, 2017)
12	Volatility	What time does information remain relevant and need to be stored? (Firican, 2017)
13	Visualisation	Refers to more modern data visualisation tools, which show real-time changes and more visualised graphics, moving beyond the use of pie, bar, and other traditional charts. (Firican, 2017)
14	Viscosity	It may also be used to represent data loss, latency, or delays with respect to the phenomenon being described. (Vorhies, 2014)
15	Virality	Determines how quickly data can spread over a network. (Vorhies, 2014)
16	Virtual	Big data virtualization allows businesses and other organizations to use all of the data assets they collect to achieve various goals and objectives. (Hussein, 2020)
17	Valences	It is a metric that shows how packed the data is. (Hussein, 2020)
18	Viability	Viability is the review assessment of data properties that are most likely to predict important outcomes to enterprises. (Hussein, 2020)
19	Virility	Big Data is self-creating. Big Data grows more powerful the diverse applications. (Darrin, 2016; Dhamodharavadhani, 2018; Gartner, 2013)
20	Vendible	Big Data clients' very existence is vital evidence that they value it; this is seen from the interaction of some established methods of data trading with subscribers. (Dhamodharavadhani, 2018; Gewirtz, 2016; GoodStratTweet., 2015)
21	Vanity	Data that is vain, content with the impact it has on other people. (Gewirtz, 2016; GoodStratTweet., 2015)

22	Voracity	Big Data has the potential to be so greedy that it might influence, control, and even have the capacity to consume itself. (Gewirtz, 2016; GoodStratTweet., 2015)
23	Visible	Relevant information must exist and be made apparent to the targeted person at the appropriate time. (Laney, 2012)
24	Visual	We currently live in a world where people view, exchange, and share photos and videos online, whether they are of themselves, their products, or the weather. (Laney, 2012)
25	Vitality	An essential perspective that is included in the concept of worth is the vitality of the data. (Laney, 2012)
26	Vincularity	Its precise meaning suggests connectedness or linkage. In the internet-connected world of today, this concept is extremely relevant. (Cartledge, 2016)
27	Verification	The process of checking the validity and verification of data. (GoodStratTweet., 2015)
28	Valor	The precise data potentially produce value and provide direction for doing so. (Borne, 2014)
29	Verbosity	Processing effectiveness depends on rapidly distinguishing between the meaning you remember and its repetition. (Laney, 2012)
30	Versatility	The data's versatility reveals how beneficial it is in various situations. (Laney, 2012)
31	Veritable	Data that is true, authentic, and not made up or fabricated. (Hussein, 2020)
32	Violable	Data is likely to be or potentially violated. (Hussein, 2020)
33	Varnish	End consumers need to interact with our work and polish matters. (Hussein, 2020)
34	Vogue	Does Artificial intelligence becomes? (Sivarajah, 2017)
35	Vault	Importance of data related to security. (Sivarajah, 2017)
36	Voodoo	Deliver outcomes that have an actual impact. (Sivarajah, 2017)
37	Veil	Investigate latent variables from a hidden perspective. (Sivarajah, 2017)
38	Vulpine	Data refer to new technology. (Sivarajah, 2017)
39	Verdict	People impacted by the model's choice. (Sivarajah, 2017)
40	Vet	Putting the assumptions to the test with facts. (Sivarajah, 2017)
41	Vane	Unclear decision-making process. (Hussein, 2020; Sivarajah, 2017)
42	Vanilla	If used carefully, simple techniques can be beneficial. (Sivarajah, 2017)
43	Victual	Big Data is the data science's fuel. (Sivarajah, 2017)
44	Vintage	Advantaged perspective on complicated systems. (Sivarajah, 2017)
45	Varmint	Software flaws increase in size as data does. (Sivarajah, 2017)
46	Vivify	Data science's capacity to deal with every facet of daily existence. (Sivarajah, 2017)
47	Vastness	It refers to the bigness of data. (Sivarajah, 2017)
48	Voice	The capacity to speak intelligently. (Sivarajah, 2017)
49	Vaticination	Ability to predict. (Sivarajah, 2017)
50	Veer	Adapt a course according to the needs of customers. (Hussein, 2020; Sivarajah, 2017)
51	Voyage	Improving our knowledge. (Sivarajah, 2017)
52	Varifocal	It concerns forests and trees. (Sivarajah, 2017)
53	Version Control	Are you utilizing it properly? (Sivarajah, 2017)
54	Vexed	Data science's potential to solve challenging issues. (Sivarajah, 2017)
55	Vibrant	The insight provided by data science. (Sivarajah, 2017)
56	Vogue	What will artificial intelligence become? (Sivarajah, 2017)

Challenges: Every opportunity brings its fair share of challenges. Big Data offers a multitude of enticing prospects, but it also presents numerous hurdles related to the gathering, storage, sharing, searchability, analysis, and visualization of such vast datasets. Unless we can surmount these obstacles, Big Data remains an untapped resource, akin to unmined gold. The current limitation lies in our need for more tools to explore the ever-increasing magnitude of information effectively. An enduring problem in computer architecture has been the imbalance between CPU-intensive tasks and input/output performance. This disparity hampers the progress of discovering insights from big data.(C.L. Philip Chen, 2014)

Following Moore's Law, CPU performance experiences a twofold increase approximately every 18 months, and the same holds for disk drive performance. However, in the past decade, the rotational speed of the disks has seen minimal growth. As a result, an imbalance exists where random input/output (I/O) speeds have seen slight enhancements while sequential I/O speeds have gradually improved alongside increased data density. Additionally, while the amount of information is growing exponentially, the pace at which information processing techniques develop could be faster. Modern methods and tools, especially those for real-time analysis, often need more optimal solutions in many significant Big Data applications. So, we have just now had the suitable instruments to utilize the gold ores best. (C.L. Philip Chen, 2014)

Challenges encountered in BDA consist of data inconsistency, incompleteness, timeliness, data security and scalability. Data construction in a suitable manner is a prerequisite for practical data analysis. Each sub-process of data-driven applications brings forth distinct challenges. In the subsequent subsections, we will provide a concise overview of the difficulties encountered in each phase.

a) Data Capture.

Big data can be sourced from various outlets, such as transactions, metadata, social media, sensors, and experiments. Collecting and consolidating data from diverse origins presents challenges regarding scalability and the sheer volume involved. In the future, organizations that can not only amass larger and higher-quality datasets but also leverage them efficiently on a large scale may gain a competitive edge. The handling of data preprocessing, automated metadata creation, and data transfer are additional concerns linked to this subject matter.(Lisbeth Rodríguez-Mazahua, 2015)

b) Data Storage.

Handling Big Data necessitates ample storage capacity and innovative data administration approaches on extensive distributed systems, as traditional database systems need help to cope with obstacles posed by Big Data. MapReduce facilitates automatic parallelization and scalable data distribution across multiple machines. Apache Hadoop, an open-source software, stands out as the preferred implementation. (Gantz J, 2012; Schadt E, 2010; V, 2013)

c) Data Search

Given the requirement for data to be timely, dependable, and comprehensive to support informed decision-making, this necessity becomes crucial. Query optimization is vital in effectively addressing a wide range of complex analytic SQL queries. Demand for higher-level query languages (such as HiveQL, Pig Latin, and SCOPE) remains strong, even for the latest frameworks based on MapReduce and its derivatives. As data movement between nodes in parallel platforms incurs high costs, optimizing queries and refining physical designs remain essential infrastructure elements. (Chaiken R, 2008; Olston C, 2008; C. S, 2012; Thusoo A, 2010)

d) Data Sharing

In the current landscape, data sharing holds equal importance to data generation. While data is now being produced to allow for integration and sharing across different parts of an organization, professionals still generate and utilize information tailored to their business requirements. Challenges arise in data curation and safeguarding privacy when managing these aspects effectively. (D, 2012; E, 2012; Hampton SE, 2013; Zhang X, 2014)

e) Data Analysis

In the present Era, the ability to perform timely and cost-effective analytics on large-scale datasets is crucial for the success of numerous business, scientific, technical, and governmental endeavors. One solution to address these challenges lies in the appeal of system scalability, as exemplified via cloud computing. (Agrawal D, 2011; Begoli E, 2012; Borkar V, 2012; Bu Y, 2012; Chen H, 2012; Manyika J, 2011; McAfee A, 2012; M. S, 2012; S, 2014; Walker DW, 1996; Wu X, 2014)

f) Data Visualization

The characteristics of Big Data pose challenges when it comes to visualizing the data. Visual interfaces are suitable for the given purposes: (1) Exploring data at different scales in conjunction with statistical analysis, as stated by Fisher et al., (2) Preserving context by representing data in the form of a smaller subset of a larger dataset, displaying correlated variables, and more. (3) Facilitating the ongoing discovery of patterns in data streams over the long term. In the emerging field of visual analytics, the aim is to present large datasets in visually appealing ways, enabling users to identify significant relationships. Creating multiple visualizations across diverse datasets is imperative for effective visual analytics. (Fisher D, 2012; J, 2014; Light RP, 2014; Shen Z, 2012)

1. History of Hadoop

Doug Cutting, the creator of Apache Lucene, a broadly used textual content exploration library, invented Hadoop. Hadoop's origins can be traced back to Apache Nutch, an open-source online exploration engine directly connected to the Lucene project.

The term 'Hadoop' is not an initialism but a fabricated designation. Doug Cutting, the project's initiator, elucidates the name's genesis: “It was bestowed upon a plush, yellow elephant by my child.

To conform my criteria for naming, it had to be succinct, relatively effortless to spell and articulate, devoid of intrinsic meaning, and unutilized elsewhere. Children excel at generating such appellations. 'Googol' is another term conceived by a child.” (White, 2012)

Hadoop evolved from Nutch, an open-source crawler-driven search engine on a distributed system. Google released the Google MapReduce and GFS papers in 2003-2004. MapReduce was implemented in the Nutch framework. Doug Cutting and Mike Cafarella established Hadoop. When Doug Cutting joined Yahoo, a new initiative based on the same concepts as Nutch was formed, which became known as Hadoop, while Nutch remained a distinct sub-project. Several versions occurred then, and additional sub-projects began integrating with Hadoop, building the Hadoop ecosystem. (Achari, 2015)

2. Big Data: Hadoop Ecosystem.

A Hadoop cluster has the potential to encompass a multitude of nodes, rendering it intricate and arduous to operate manually. Consequently, diverse elements facilitate the complete Hadoop System setup, upkeep, and administration.

For specific individuals, Hadoop serves as a data governance system that amalgamates extensive volumes of organized and unstructured data, permeating nearly all strata of the established data infrastructure within an enterprise strategically positioned to assume a central role within a data center. Alternatively, it is perceived as a highly parallel execution framework that democratizes the potency of supercomputing, poised to energize the execution of enterprise applications. Some regard Hadoop as an open-source community that fabricates tools and software to tackle the challenges Big Data poses. Due to the extensive range of capabilities offered by Hadoop, adaptable to address diverse problems, it is widely recognized as a fundamental framework. (Lublinsky, Smith, & Yakubovich, 2013)

Core Components of Hadoop Ecosystem: Hadoop's various components are configured using its configuration API. These components collaborate effectively to form a robust Hadoop ecosystem that can be applied to solve a wide range of real-world problems. The core components are shown in Figure 5.

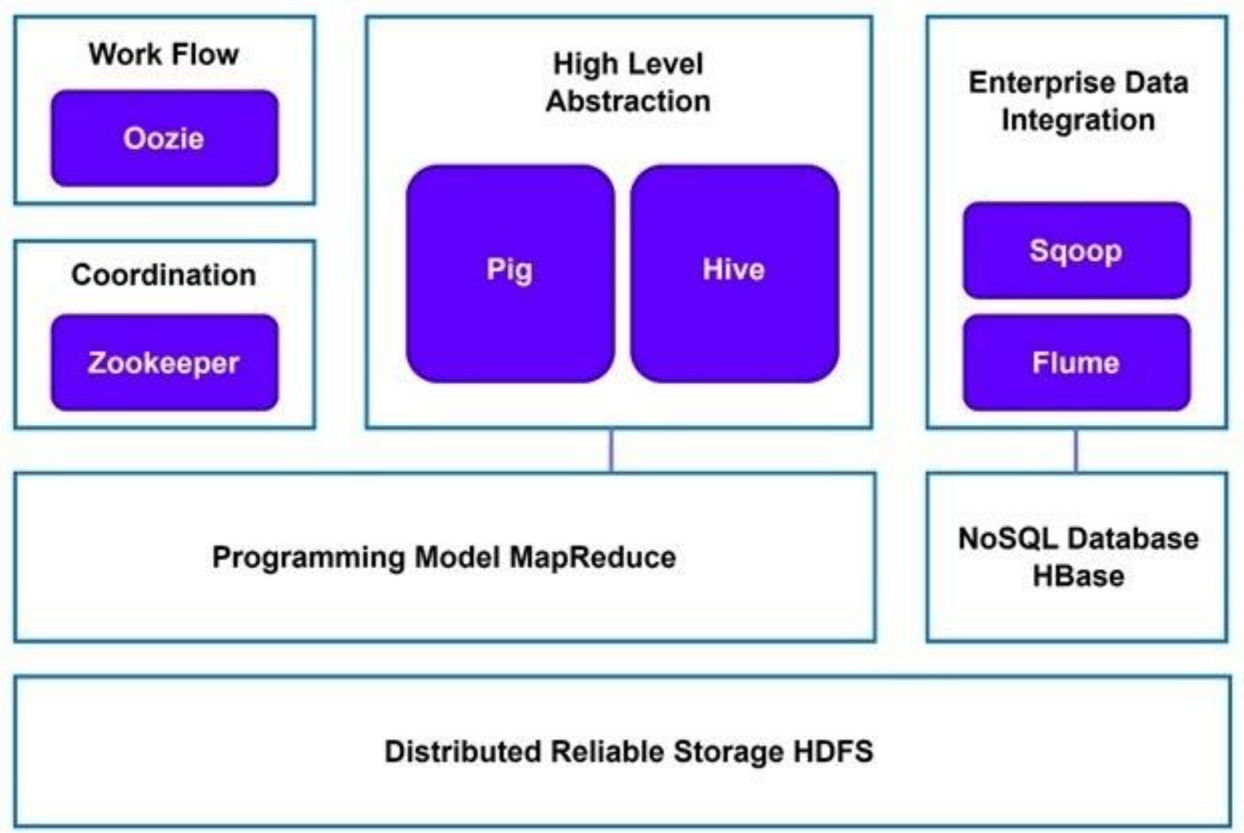

Figure 5: Core components of Hadoop

a) HDFS

Hadoop employs a distributed file system and disperses extensive files across numerous local storage-equipped Data Nodes within the cluster. While processing, the Name Node partitions the original file, turns into blocks, typically sized at 64 MB, and replicates the blocks into various Data Nodes based on a predefined protocol. The Name Node consistently updates the metadata associated with replication and allocation. (Gurjit Singh Bhathal *, 2019) (Zafar, 2021)

b) Map Reduce

Much like HDFS, MapReduce is based on a master-slave model and works on a parallel processing framework. It encompasses a solitary master (Job Tracker) daemon paired with three enslaved people (Task Tracker), with each slave operating a daemon within a cluster. Data is processed parallel by Map Reduce using several algorithms. The process entails mapping the task and then reducing it. In order to facilitate parallel processing, this Job Tracker separates the dataset into several units known as tasks (Map tasks) and distributes them by default to three Data Nodes (Task Tracker).

In case of disruption, the Job Tracker is responsible for monitoring and rescheduling any tasks. If a task fails to show progress within a designated timeframe or a Data Node experiences a complete failure, all jobs, including incomplete one, are restarted on an alternative server. Additionally, if a task runs unusually slow, the Job Tracker restarts it on a different server to ensure the timely completion of the work according to the schedule. (speculative execution).

c) Zookeeper

Zookeeper serves as the distributed coordination service for Hadoop. Engineered to operate across a cluster of machines, it functions as a remarkably dependable service employed for the administration of Hadoop operations, upon which numerous Hadoop components rely.(Lublinsky et al., 2013)

d) Hive

Facebook pioneered the introduction of a Hadoop interface akin to SQL. Hadoop's Hive interface resembles SQL, allowing SQL users to create MapReduce jobs without requiring familiarity with MapReduce, using familiar SQL commands and a relational table structure. Hive treats all data elements as if they were tables, enabling the creation of table definitions based on data files. Furthermore, it organizes metadata of unstructured data into tables and converts inputs into MapReduce jobs.

e) Oozie

A Hadoop cluster uses the workflow and coordination tool Oozie. It is implemented on a supercomputing platform. Thanks to this, jobs may run concurrently while awaiting input from other jobs. Oozie has several intriguing benefits, including a highly sophisticated scheduling tool. This enables the supercomputing platform to coordinate jobs that are awaiting other requirements. (" http://hive.apache.org/,")

f) HBase

This is a well-known Hadoop-based NoSQL columnar database. The Apache project HBase is based upon Google's Big Table data storage paradigm. It offers a column-oriented representation of the data and has no schema. ("Apache Oozie Workflow Scheduler for Hadoop," 2019)

g) Pig

The Pig platform is an elevated abstraction layer for the complexities inherent in MapReduce programming. It encompasses an execution environment and a scripting language called Pig Latin, designed to facilitate the analysis of data sets in the Hadoop ecosystem. Leveraging its compiler, Pig seamlessly translates Pig Latin scripts into sequences of MapReduce programs. ("Apache HBase," 2019)

In addition to the fundamental constituents illustrated in the aforementioned

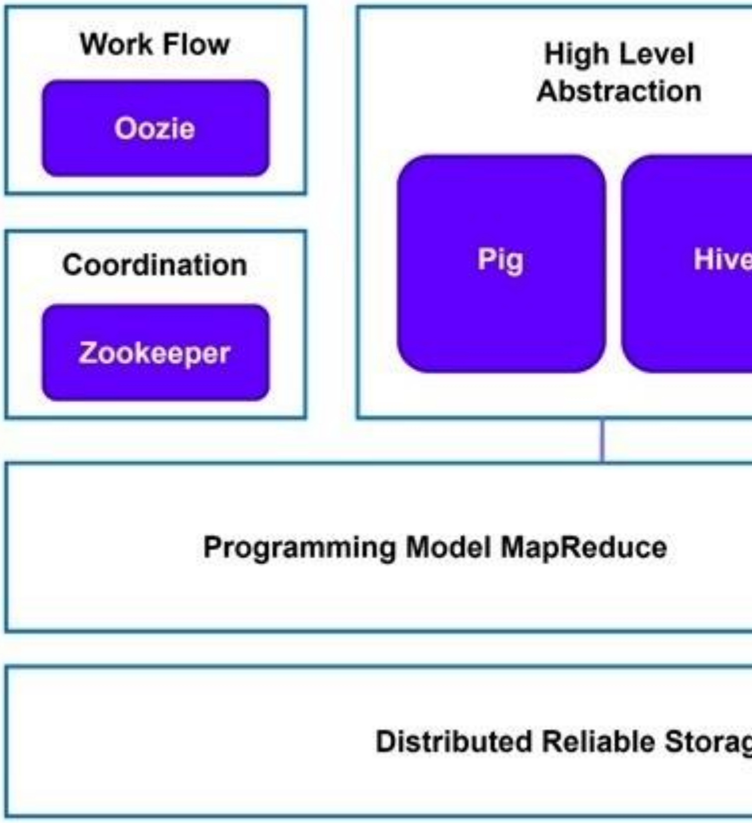

Figure 5: Core components of **Hadoop**, Hadoop's ecosystem is progressively expanding to offer advanced capabilities and additional components. Features, Strengths and Weaknesses of the core components of Hadoop and the newly introduced elements are described in Table 3. (Lublinsky et al., 2013)

Big Data: Issues, Benefits and Uses: Big data encompasses various issues and considerations that require careful attention and resolution. However, it is equally important to acknowledge the substantial uses, benefits, issues and widespread applications that big data offers across diverse domains and industries, which are presented in Table 4

Table 3: Features, Strengths and Weaknesses of the core components of Hadoop and the new elements

Sr. No	Tools	Features	Strengths	Weakness
1	(HDFS)	It helps in storing mass volumes of data. It is trusted and faults tolerant	Less expensive data storage will allow for one-time data reading and multiple-time data writing.	More efficient support for retrieving small data portions through random reads is needed. Furthermore, the administration of Hadoop clusters poses significant challenges.
2	MapReduce	Distributed programming platform for managing compute jobs, scheduling resources, and batch processing.	Cost-effective and immensely scalable due to the capacity to store enormous volumes of scattered data	Need help to process interactive, in-memory, or graph data. Additionally, map-reduce needs to be set up for small datasets.
3	YARN	Accountable for scheduling jobs and allocating resources in Hadoop. The OS of Hadoop 2.0 manages resources across several clusters, maintains information meta-data, and monitors user data.	The addition of YARN to Hadoop helps to guarantee practical resource usage and good data availability.	Accurate parameter setting is complex and requires an in-depth understanding of each parameter.
4	Hive	Hive is mainly utilized for big data tasks like querying, data analysis, and summarization utilizing an interface that resembles SQL.	Utilizing an indexing strategy makes creating and handling the vast dataset easier and more reliable.	Processing transactions online is not a task for Apache Hive. Additionally, it does not handle database operations like updates, deletes, and subqueries.
5	Flume	For importing and exporting data into and out of Hadoop, utilise Apache Flume.	Provide a user-friendly and flexible framework for swiftly aggregating and transferring vast data streams into HDFS.	Low scalability and a high failure point.
6	Apache Spark	Hadoop tools for computer vision and real-time processing.	Efficient for reading and writing operations, batch processing, joining streams, and the capacity to manage worker node failures.	Providing real-time processing takes much work. Additionally, they need help manually processing and optimizing for a particular dataset.
7	Oozie	A tool for arranging and managing Hadoop cluster jobs for parallel processing.	Enable fault-tolerant workflow for the execution of many jobs. Moreover, it offers a web service API for scheduled jobs.	Off-grid scheduling could work better with Oozie.
8	HBase	It is used for storing data and column-oriented data views. NoSQL column database is used.	A solution that allows enormous datasets to be stored on top of the Hadoop distributed file system.	Joins and cross-data operations take time to implement.
9	Flink	A perfect tool for handling batch and streaming operations. In Hadoop, it is effective for distributed stream processing and real-time analysis.	Flink provides a centralized run-time environment that may be used for batch and streaming data processing.	For processing large amounts of data, Flink is not frequently used and receives few community contributions.
10	Storm	It is used for real-time	Efficient for low latency, high	Lack of significant

	streaming and processing and data analytics for examining vast amounts of real-time data.	throughput, and straightforward streaming operations.	functionality for state management, data aggregation, and event time processing

Table 4: 6. Big Data: Issues, Benefits and Uses.

No	Uses	Benefits	Issues
1	Storage and Transport Issues	Medical: Diagnoses of Cancer.	Targeted advertising
2	Security and Privacy Issues	Fraud detection: Used to detect fraudulent behavior.	Healthcare analytics and precision medicine.
3	Management Issues	Efficient resource allocation: Optimize the allocation of resources	Climate prediction
4	Processing Issues	Decision-making: derive meaningful insights from large datasets.	Supply chain optimization

Applications: BDA technology has brought advantages to a range of corporate and business sectors. These sectors generate massive volumes of data, necessitating applying BDA methods for effective and streamlined decision-making. Some domains where these techniques are applied are farming, medical care, traffic management, educational institutions, and the banking and financial sectors.

a) Agriculture

In the agronomic field, big data states the use of all current technologies and data analysis as a building block for decision-making based on data. Massive Data has been utilized to advance several facets of agriculture, including remote sensing, decision-making by farmers, insurance and financing for farmers, knowledge of the change in climate, fields, studies of animals, crops and soil, and food availability. (H. B. U. Haq, 2020)

b) Smart City Traffic Control.

An essential element of intelligent urban centers is the efficient management of traffic flow, which aims to enhance the city's transportation systems, reduce commute times for residents, and optimize overall traffic patterns. As the population expands, challenges such as traffic congestion, environmental impact, and economic concerns arise. Consequently, smart cities implement advanced traffic signals and intelligent traffic management systems to address heavy traffic and congestion effectively. The ideal approach involves collecting data from all traffic signals across the city and leveraging this information to develop intelligent algorithms for making informed decisions, thereby providing the most optimal services for intelligent traffic management. (Aguilera G, 2013; Sepúlveda, 2021)

c) Healthcare

Enhanced well-being and profitable socio-economic growth rely on advancements in healthcare. The healthcare industry generates substantial data that can empower physicians and healthcare practitioners to make more informed choices. Furthermore, leveraging massive data in healthcare can facilitate the development of real-time disease assessment, ultimately benefiting public health. To establish the correlation between data and environmental risks in public healthcare, comprehensive tracking, monitoring, storage, and analysis of mobile entities and their exposure to potentially harmful environmental factors are imperative. (Eiman Al Nuaimi, 2015)

d) Network Optimization

Utilizing BDA methodologies, it is possible to architect a mobile network that provides efficient and reliable services. Content-focused examination, traffic evaluation, and network signaling are vital aspects for ensuring optimal service delivery and superior performance. Network providers can establish a system for collecting, storing, and evaluating user or core network data to enhance signaling effectiveness, forecast traffic fluctuations, prevent congestion, intelligently optimize the network, automate network configuration, and foster intelligent communication. (Khatib, Barco, Muñoz, De La Bandera, & Serrano, 2016)

e) Educational Development

The field of education presents a plethora of data sources that can be leveraged for BDA—these information reservoirs aid in forecasting student performance and achievement. Additionally, using BDA in the educational domain is pivotal in overseeing curriculum content, constructing customized recommendation systems, and facilitating intelligent learning by applying text condensation and computational linguistics. In order to improve teaching and learning, data from massive open online courses (MOOCs) is also used to assist in identifying subject areas that are challenging for students and to support them. (Dobre, 2014)

Big Data is beneficial for customizing educational procedures and raising academic performance. The ability to gather a wealth of data about students—present and past—allows educators to adjust their educational methods and choose the best tools. The results of generated data may influence the development of pedagogical design and teaching methodology. (Nweke, 2019) (Julio Ruiz-Palmero, 2020)

f) Banking Sector

The utilization of client data inevitably raises privacy concerns. BDA may expose private information by revealing hidden links between seemingly unrelated pieces of data. According to research, 62% of bankers cautiously handle big data due to privacy concerns. In addition, spreading consumer data across departments to produce deeper insights or outsourcing data analysis tasks also increase security threats.

For instance, a prominent UK bank's recent security lapse exposed a database containing thousands of customer files. Even though this bank immediately began an inquiry, sensitive documents were found. Such involves misappropriating consumers' wages, savings, mortgages, and insurance policies. (Muhammad Ali Raza, 2023)

g) Finance

Big data carries noticeable ramifications for the financial sector. The essential requirement in the realm of finance revolves around the manipulation of amassed data. The insights extracted from the unprocessed data primarily inform the decision-making process. Given the significance of massive financial data, structured business intelligence is perceived as an advancement. The enterprise leverages organized data to propel decision-making forward, as unrefined and unanalyzed data holds no value to the organization.

Accountants employ diverse methodologies to derive meaningful insights from the data at hand. Harnessing data analysis for pertinent decision-making, efficiency gains, and innovation brings numerous advantages. All business operations can be effectively supported by integrating data with robust analytics.

The convergence of corporate data and burgeoning big data has occurred in finance. This enables the seamless integration of ERP systems with unstructured data repositories. (Kuchipudi Sravanthi, 2015)

Conclusion: A forthcoming scientific revolution is on the horizon as we enter the realm of big data, which represents the next frontier for innovation, competition, and efficiency. We eagerly anticipate the forthcoming technological upheaval. This manuscript presents a comprehensive definition of big data utilizing the "V" approach, aiming to enhance scholars' understanding of this concept. Moreover, this article delves into various challenges and issues associated with big data. To fully harness the potential of big data, it is imperative to foster fundamental research into these technical obstacles. The Hadoop Framework and its diverse components are also examined in this publication. HDFS, designed for standard hardware, stands as a pivotal element within big data.

Furthermore, this study elucidates the various domains where big data finds application. Big data ushers in many novel opportunities and exerts far-reaching impacts. Presently, organizations possess many alternatives that enable the formulation of innovative propositions. Businesses can now leverage big data strategies to attain novel and enhanced outcomes.

REFERENCES

Agrawal D, D. S. (2011). Big data and cloud computing: current state and future opportunities. *Proceeding of the 14th international conference on extending database technology (EDBT/ICDT). ACM, pp 530–533*.

Aguilera G, G. J. (2013). An Accelerated-Time Simulation for Traffic Flow in a Smart City. *FEMTEC*.

Apache Flink. (2019). Retrieved from https://flink.apache.org/

Apache HBase. (2019). Retrieved from http://hbase.apache.org/

Apache Oozie Workflow Scheduler for Hadoop. (2019). Retrieved from http://oozie.apache.org/

Apache Storm. (2019). Retrieved from https://storm.apache.org/

Arockia Panimalar.S 1, V. S. (2017). The 17 V's Of Big Data. *International Research Journal of Engineering and Technology (IRJET)*.

Begoli E, H. J. (2012). Design principles for effective knowledge discovery from big data. *Proceeding of the joint working IEEE/IFIP conference on software architecture (WICSA) and European* *conference on software architecture (ECSA), pp 215–218*.

Bhathal GS, S. A. (2019). Big data computing with distributed computing frameworks. *Saini H, Singh R, Kumar G, Rather G, Santhi K, editors. Innovations in Electronics and Communication Engineering*.

Borkar V, C. M. (2012). Inside “Big Data Management”: ogres, onions, or parfaits? . *In: Proceeding of EDBT/ICDT joint conference. ACM*.

Borne. (2014). *https://www.mapr.com/blog/top-10-big-data-challenges-serious-look-10-big-data-vs*.

Borne, K. (2014). Top 10 Big Data Challenges—A Serious Look at 10 Big Data V’s.

Bu Y, B. V. (2012). Scaling datalog for machine learning on big data. . *Computer research repository (CoRR) Cornell University Library, pp 1–14*.

C.L. Philip Chen, C.-Y. Z. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Retrieved from http://www.elsevier.com/locate/ins

Cartledge, C. (2016). How Many vs Are There in Big Data?

Chaha, P. G. (2020). BIG DATA ANALYTICS FOR IOT. *International Journal of Advanced Research in Engineering and Technology (IJARET)*.

Chaiken R, J. B. (2008). SCOPE: easy and efficient parallel processing of massive data sets. . *Proc VLDB Endow 1(2):1265–1276*.

Chen H, C. R. (2012). Business intelligence and analytics: from big data to big impact. *Manag Inf Syst Q (MIS) Q 36(4):1165–1188*.

D, W. (2012). Linking enterprise data. *Springer, New York*.

Darrin, D. (2016). *https://educationalresearchtechniques.wordpress.com/2016/05/02/characteristics-of-big-data/*.

Dhamodharavadhani, S. a. (2018). Unlock Different V’s of Big. *International Journal of Computer Sciences and Engineering*.

Dobre, C. X. (2014). Intelligent service for Big Data Science. . *Futur. Gener. Comput. Syst.*

E, S. (2012). The changing privacy landscape in the Era of big data. *Mol Syst Biol 8(612):1–3*.

Eiman Al Nuaimi, H. A. (2015). Applications of big data to smart cities. *Journal of Internet Services and Applications*. doi: 10.1186/s13174-015-0041-5

Firican. (2017). The 10 V’S BIG DATA.

Firican, G. (2017). The 10 Vs of Big Data.

Fisher D, D. R. (2012). Interactions with big data analytics.

Flume, A. (2019). Retrieved from https://flume.apache.org/

G. Kapil, A. A. (2016). A study of big data characteristics. *International Conference on Communication and Electronics Systems (ICCES)*.

Gantz J, R. D. (2012). The digital Universe in 2020: big data, bigger digital shadows, and biggest growth in the far east. *IDC IVIEW: IDC Analyze the Future 1414_v3:1–16*.

Gartner. (2013). *http://www.forbes.com/sites/gartnergroup/2013/03/27/gartners-big-data-definiti-consists-of-three-parts-not-to-be-confused-with-three-vs/%235040d1cc3bf6*.

Gewirtz, D. (2016). *http://www.zdnet.com/article/volume-velocity-and-varietyunderstanding-the-three-vs-of-big-data/*.

GoodStratTweet. (2015). *http://www.informationweek.com/big-data/big-data-analytics/big-data-avoid-wanna-v-confusion/d/d-id/1111077?Page_number=1*.

Gurjit Singh Bhathal *, A. S. (2019). Big Data: Hadoop framework vulnerabilities, security issues and attacks. *www.elsevier.com/journals/array/2590-0056/open-access-journal*.

H. B. U. Haq, H. U. (2020). The Popular Tools Of Data Sciences: Benefits, Challenges and Applications. *International Journal of Computer Science and Network Security*.

Hampton SE, S. C. (2013). Big data and the future of ecology.

Hive, A. (2023). (Apache Hive) Retrieved 2019, from http://hive.apache.org

Huidong Sun, M. R. (2020). Identifying Big Data’s Opportunities, Challenges,and Implications in Finance. *mathematics*. Retrieved from http://www.mdpi.com/journal/mathematics

Hussein, A. A. (2020). Fifty-Six Big Data V’s Characteristics and Proposed Strategies to Overcome Security and Privacy Challenges (BD2). *Journal of Information Security, 2020, 11, 304-328*.

Hussein, A. A. (2020). How Many Old and New Big Data V's Characteristics, Processing Technology, And Applications (BD1). *International Journal of Application or Innovation in Engineering & Management (IJAIEM)*.

J, C. (2014).) Visualizing big data with compressed score plots: approach and research challenges. *Chemometr Intell Lab Syst 135:110–125*.

Julio Ruiz-Palmero, E. C.-M.-A. (2020). Big Data in Education: Perception of Training Advisors on Its Use in the Educational System. *social sciences*. Retrieved from http://www.mdpi.com/journal/socsci

Khatib, E. B., & Muñoz, P. D. (2016). Self-Healing in Mobile Networks with Big Data.

Kuchipudi Sravanthi, T. S. (2015). Applications of Big data in Various Fields. *International Journal of Computer Science and Information Technologies.*

Kumari, Y. K. (2020). A Study of Big Data Analytics using Apache Spark. *Proceedings of the Third International Conference on Intelligent Sustainable Systems [ICISS 2020]*.

Laney, D. (2012). http://blogs.gartner.com/doug-laney/deja-vvvue-others-claiming-gartners-volume-velocity-variety-construct-for-big-data/.

Light RP, P. D. (2014). Open data and open code for big science studies.

Lisbeth Rodríguez-Mazahua, C.-A. R.-E. (2015). A general perspective of Big Data: applications, tools, challenges and trends. doi: 10.1007/s11227-015-1501-1

Manyika J, C. M. (2011). Big data: the next frontier for innovation, competition and productivity. . *McKinsey Global Institute, New York*.

McAfee A, B. E. (2012). Big data: the management revolution. *Harv Bus Rev 90(10):60–68*.

Mircea Răducu TRIFU, M. L. (2014). Big Data: present and future.

Muhammad Ali Raza, e. a. (2023). Review: Big Data Trends, Tools and Applications in Education.

Nweke, I. A. (2019). Big Data and Business Analytics: Trends, Platforms, Success Factors and Applications . *Big data and cognitive computing*.

Olston C, R. B. (2008). Pig Latin: a not-so-foreign language for data processing. *In: Proceeding of the SIGMOD conference, pp 1099–1110*.

S, C. (2012). What next? A Half-Dozen data management research goals for big data and the cloud. *Proceeding of the symposium on principles of database systems (PODS)*.

S, M. (2012). From databases to big data.

S, S. (2014). Big data classification: problems and challenges in network intrusion prediction with machine learning.

Schadt E, L. M. (2010). Computational Solutions to Large-Scale Data Management and Analysis.

Sepúlveda, A. C. (2021). Use and Adaptations of Machine Learning in Big Data—Applications in Real Cases in Agriculture. *electronics*.

Shen Z, W. J.-L. (2012). Visual analysis of massive web session data. *IEEE symposium on large data analysis and visualization (LDAV)*, 65-72.

Sivarajah, U. K. (2017). Critical Analysis of Big Data Challenges and Analytical Methods. *Journal of Business Research*.

Spark. (2019). *https://spark.apache.org/*. Retrieved from https://spark.apache.org/

Thusoo A, S. J. (2010). Hive-A petabyte scale data Warehouse using Hadoop. *Hive-A petabyte scale data Warehouse using Hadoop*.

V, M. (2013). Biology: The Big Challenges of Big Data. *Nature*.

Vorhies. (2014). How Many V'S in Big Data. Retrieved from http://www.aimspress.com/journal/Math

Walker DW, D. J. (1996). MPI: a standard message passing interface.

Wu X, Z. X.-Q. (2014). Data mining with big data.

Zafar, S. &. (2021). Big Data: Challenges, Popular Tools Of Big Data -Benefits And Applications. *International Journal of Scientific & Technology Research*.

Zhang X, Y. L. (2014). A scalable two-phase top-down specialization approach for data anonymization using mapreduce on cloud. *IEEE Trans Parallel Distrib Syst 25(2):363–373*.

ZHANG Yaoxue1, R. J. (2017). A Survey on Emerging Computing Paradigms for Big Data. *Chinese Journal of Electronics*.