The Scientific Temper (2025) Vol. 16 (spl-2): 70-75
E-ISSN: 2231-6396, ISSN: 0976-8653
Doi: 10.58414/SCIENTIFICTEMPER.2025.16.spl-2.11
https://scientifictemper.com/

REVIEW ARTICLE

Unlocking the potential of big data and analytics significance, applications in diverse domains and implementation of Apache Hadoop map/reduce for citation histogram

Madhuri Prashant Pant[1*], Jayshri Appaso Patil[2]

Abstract

Big Data analytics helps every sector that wants to grow. This paper also presents an analysis of Big Data case studies in diverse domains. Various domains include the supply chain, academia, India's unique identification project, the Digital India program, the Smart Cities project, and improved healthcare. The paper presents how Walmart Inc., a global retail chain offering a wide range of products, including groceries, apparel, and financial services, leverages Big Data analytics, processing 2.5 petabytes of data every hour, to enhance sales by predicting customer needs and optimizing product availability across its stores and how Oracle, committed to helping people discover insights and unlock possibilities through data, showcases case studies in manufacturing that use predictive maintenance to enhance operational efficiency, production optimization, and product development. Additionally, Oracle's Big Data solutions enhance customer experience by analyzing data from mobile applications, in-store purchases, and various geographic locations to improve sales and motivate purchase completion. This research paper presents Big Data, its characteristics, the use of Hadoop to handle Big Data challenges, and how it is superior to typical distributed systems and RDBMS. It also explains how Hadoop's HDFS and MapReduce work. It also provides Apache Hadoop's installation instructions and demonstrates the running of the Citation Histogram MapReduce program. It also discusses the Hadoop Ecosystem. This paper contributes to everyone who wants to work in the field of Big Data, providing hands-on practical implementation. Researchers can use different data and derive useful results.

Keywords: Big Data analytics, Hadoop, MapReduce, NameNodes, Jobtracker, Tasktracker, Hadoop ecosystem, Citation histogram.

Introduction

Big data analytics is widely used across various industries, including trading, marketing, manufacturing, telecommunications, financial services, electronic payments, process automation, pharmaceuticals, advertising, and more. Organizations generate and manage massive volumes of multi-structured and unstructured data, posing challenges due to the growing volume, variety, and velocity of data. The size of multi-structured and unstructured data in organizations is huge. Considering the average data growth, the volume, variety, and velocity of data are becoming a problem. "Big Data" refers to datasets whose size or complexity exceeds the capabilities of traditional relational databases to capture, manage, and process with low latency. It has emerged as one of the most significant trends in information technology (IT) (Mucci & Stryker, 2022)".

Every day, we collect large amounts of data from sensors to predict climate, as well as data from social media platforms like Facebook, Twitter, Yahoo, Google, and Amazon, among others. This includes videos, photos, online transaction records, and mobile phone signals. Data increases in size as a result of extensive data collection from mobile devices, remote sensing devices, wireless networks, cameras, microphones, and radio frequency identification (RFID) devices. All of this unstructured data is Big Data (White, 2013).

Data takes the form of "Big Data" when existing techniques cannot store and process it competently

[1]Department of Computer Science, Faculty of Computer Science and Technology, Vishwakarma University, Pune, India

[2]MCA Department, Dr. D. Y. Patil Center Management and Research (MBA & MCA), Chikhali, Pune, India

***Corresponding Author:** Madhuri Prashant Pant, Department of Computer Science, Faculty of Computer Science and Technology, Vishwakarma University, Pune, India, E-Mail: pantmadhuri123@gmail.com

How to cite this article: Pant, M.P., Patil, J.A. (2025). Unlocking the potential of big data and analytics significance, applications in diverse domains and implementation of Apache Hadoop map/reduce for citation histogram. The Scientific Temper, **16**(spl-2):70-75.

Doi: 10.58414/SCIENTIFICTEMPER.2025.16.spl-2.11

Source of support: Nil

Conflict of interest: None.

Received: 18/06/2025 **Accepted:** 29/06/2025 **Published**: 08/07/2025

(Chandhini & Megana, 2013). Big Data is measured in terabytes (10^{12}), petabytes (10^{15}), quintillions (10^{18}), Wand zettabyte (10^{21} bytes). The growth of the internet is responsible for generating this data. Organizations and governments consider that the amount of user-generated content on social media platforms is extremely helpful for gaining insights into people's behavior and improving profitability as more data is analyzed (Borkar & Carey, 2012).

Literature Review and Applications

In the 21st century, Big Data analytics is reaching new heights, where huge datasets are being computed and examined to find insights from the data and use it efficiently for the growth of organizations. Using supply chain decision dynamics, new strategies are proposed in the airline industry, where Big Data analytics provide tailor-made value chains for customers (Himmi *et al*., 2017). In Big Data analytics, "heavyweight approaches"—such as ontologies, knowledge bases, fuzzy logic, and fuzzy knowledge bases—require precise analytical techniques and costly specialized software. These are often contrasted with "lightweight approaches," which rely on freeware and versatile methods, posing minimal challenges in terms of personnel training and required competencies (Globa *et al*., 2016).

Using computerized content analysis methods, academic literature on Big Data case studies has disclosed problems related to the skills possessed by people, technology, organizational culture, and processes for decision-making (Ylijoki *et al*., 2016). In India, Big Data case studies for governance include the Unique Identification Project, the Digital India program, and the Smart Cities project. Under digital India's schemes, three case studies are considered first self-identified, where the use of analytics of Big Data and methods are used to describe scheme policy documents, second publicly identified, where publicly available third-party sources are described which uses big data and third center for internet security assessed schemes that involved the generation of Big Data through various elements of the data flow, paving the way for a quantified society. They focused on decision-making based on Big Data, its amalgamation, unification, Interoperability, and common standards (Hickok *et al*., 2017).

Given the rapid expansion of medical data, enhancing the quality of patient care is essential to reduce healthcare costs and extract value and knowledge from datasets. Big Data analytics is used to achieve this. This paper provides an examination of the content, sources, technologies, tools, and challenges in healthcare related to big data. It also intends to identify the strategies to overcome the challenges (Yousef, 2021). A structured review on healthcare Big Data analytics, grounded in the Resource-Based View (RBV) theory, explores how Big Data resources are leveraged to generate organizational value and capabilities. Through content analysis of selected publications, the study examines the classification of healthcare-related Big Data types, the associated analytical techniques, the value created for various stakeholders, the platforms and tools used for managing Big Health Data, and emerging future directions in the field (Galetsi *et al*., 2020).

Walmart Incorporated is a retail store chain that sells groceries, consumables, health and wellness products, garments, and domestic items at low prices every day. The organization also merchandises energy resources, gift cards, and money management services and goods related to them, such as cash cards, money transfers, money orders, cash cheques, and bill invoices, using e-commerce websites in Asia, Africa, and America. Walmart's case study of big data is used to boost sales by enhancing customer emotional intelligence. Walmart processes 2.5 petabytes every hour. Analysis using Big Data helps retail companies to utilize last year's data to project and estimate the next year's sales. It enables retailers to gain valuable and analytical insights, especially in determining which customers are interested in desired products at a specific time in a particular store across different geographical locations (Singh, 2017).

Oracle, with a mission to help people see data in new ways, discover insights, and unlock endless possibilities, presented case studies in manufacturing that demonstrate predictive maintenance, increased operational efficiency, and production optimization, as well as product development. Big Data also enhances customer experience by identifying customer journey patterns and linking them to various behavioral trends. Many retailers are now analyzing data from mobile apps, in-store purchases, and geolocation tracking to optimize merchandising strategies and encourage purchase completion (Oracle, 2024). Researchers, through the above comprehensive literature review, unlock the perspective of Big Data analytics, its significance, and applications in diverse domains, including supply chain, academia, the Indian government's projects, healthcare, and retail. In the next section, research objectives are formulated, where researchers delve into studying Big Data in depth.

Research Objectives

The objectives of this study are to analyze the characteristics of Big Data, understand the workings of distributed systems, and identify the current problems faced in these areas. Additionally, it aims to explore the characteristics of Hadoop, including its cluster, architecture, Hadoop distributed file system (HDFS), and MapReduce framework. The study further involves the installation and implementation of Apache Hadoop MapReduce for creating a citation histogram. Lastly, it undertakes a systematic investigation of Big Data application sectors, focusing on their uses and the challenges encountered.

Big Data

There are 3 V's for data growth challenges and opportunities, namely volume, velocity, and/or variety of information, which is an asset and enables intellectual decision-making and process optimization (Beyer & Laney, 2012)

Big Data has characteristics

- *Volume*

Enterprises are overloaded with an ever-increasing volume of diverse data, rapidly accumulating petabytes and terabytes of data.

- *Velocity*

A delay of just 2 minutes can be too long to detect fraud sometimes. Such transactions are time-sensitive; Big Data helps provide insight into millions of trades a day.

- *Variety*

Big Data encompasses all types of information, including both unstructured and structured types such as sensor data, log files, text, audio, video, clickstreams, and many more.

- *Veracity*

One in three business leaders lacks confidence in the information they rely on for decision-making. Trust in the data is a significant challenge due to the increasing number of sources from which it originates.

- *Value*

Huge data is of no use till valuable information is extracted from it which is useful. (Mucci & Stryker, 2022)

Big Data Background

As a leader in the Big Data "revolution," Google had to take proactive measures to remain competitive in the search engine industry (Borkar & Carey, 2012). Declining costs of commodity hardware made it clear handling the growing volume of data necessitated the use of many computers working in parallel. Google introduced the MapReduce system alongside the Google file system to enable large-scale computation using these cost-effective computers in 2004. The MapReduce framework facilitates parallel processing for programmers (Dean and Ghemawat, 2004; Ghemawat et al., 2003).

Typical Distributed System and Current Problems Faced

In a typical distributed system, such as in high-performance computing, work is distributed across many machines that access a Storage Area Network. Clients send data to be processed. Then, at the application server, programs will execute and the processed data will be sent back to the clients. In the above system, a lot of data transfer takes place. Therefore, the system requires a substantial amount of bandwidth.

The following problems are there in the above system

- Network bandwidth is a problem
- Nearly 80% of processing power is wasted in transferring data, while only 20% is utilized in business processing.
- Partial failures are difficult to handle as the message-passing interface. The API programmer must explicitly handle checkpoints and data recovery. It is difficult for programmers to handle it.
- Scaling up and down, which involves adding and removing nodes from the system, is difficult.
- A new distributed system, Hadoop, is used for Big Data management.

Hadoop

Based on white papers by the google-on-google file system and MapReduce Dough Cutting at Yahoo, developed an open-source version of MapReduce System called Hadoop. Hadoop is an open-source framework designed for developing and executing distributed applications that handle vast amounts of data. Hadoop provides reliable shared data storage using HDFS.

Materials and Methods

Data Analysis using MapReduce

Characteristics of HADOOP

The following are the main characteristics of Hadoop

- *Accessible*

It operates on extensive clusters of standard hardware or through a cloud.

- *Robust*

Hadoop has an architecture that gracefully handles frequent hardware failures.

- *Scalable*

In a Hadoop cluster, adding and removing nodes is possible easily.

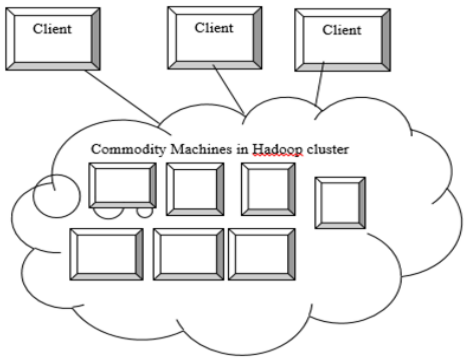

Figure 1: Hadoop cluster

- *Simple*

Parallel code can be written in Hadoop efficiently (Lam, 2010).

Hadoop Cluster

A Hadoop cluster consists of a collection of standard machines connected within a single location. Data storage and processing occur across this network of machines. People using Hadoop can give computational tasks to the Hadoop cluster from remote desktop machines or other client devices. Hadoop programs are written in Java. Figure 1 demonstrates how to interact with a Hadoop cluster.

Features of HADOOP

Hadoop operates on a cluster of standard hardware, commonly available in the market. Hardware failures are handled gracefully; programmers do not have to care. Scalable architecture means adding and removing nodes is easy. Designed for offline use and read many times and write once and analysis applications. Utilizes Move Code to data philosophy.

HDFS

When the size of datasets increases, instead of storing them on one machine, data is distributed and stored on multiple machines. HDFS file system processes distributed data under the framework of MapReduce. HDFS stores files having sizes in gigabytes or terabytes. HDFS processes data based on a write-once, read-many-times pattern. HDFS does not demand high processing and storage machines. It can process data on commonly available hardware. If any machine fails, the HDFS design will continue to function without interruption in processing. The architecture illustrates the master/slave architecture of a Hadoop cluster, where Task Trackers and Data Nodes serve as slaves. The Job Tracker and Name Node behave as the master, and a standalone node with a Secondary Name Node (SNN) is used in case the Name Node fails. Hadoop operates across multiple machines within a Hadoop cluster, where resident programs are executed (Figure 2).

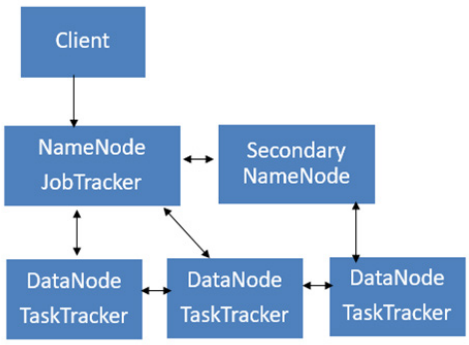

Figure 2: Architecture of HADOOP

Name Node Job Tracker

Each cluster contains only one name node, which serves as the master of HDFS. It tracks the division of HDFS files, determining which file blocks are stored in which data nodes and which blocks are stored on which data nodes. Thus, maintains metadata for all the files. For MapReduce computation, the master node acts as a job tracker. Once data and code are provided for processing, the job tracker determines the execution plan. It decides which file will be processed. Divide the job into different tasks. Assign tasks to different task Trackers. Monitors which are running. If any node fails, the task job tracker automatically relaunches failed tasks on different machines.

Secondary Name Node

Each cluster contains only one secondary name node. Without a name node, the filesystem cannot be used. As the name node is the master and keeps track of all files in the file system if the machine hosting the name node fails, all files in the file system would be lost, as there would be no way to determine how to recover the files from the blocks stored on the data nodes. To avoid loss due to the failure of SNN, a separate physical machine is used in the cluster, as it requires a significant amount of CPU and memory, comparable to that of the name node, to carry out the merging process. The primary function of the SNN is to periodically combine the namespace image with the edit log, ensuring the edit log does not grow excessively large.

Data Nodes Task Trackers

Data nodes serve as slaves in HDFS, responsible for storing and retrieving blocks as instructed by clients or the name node. They regularly report back to the name node with updates on the blocks they are managing. Data nodes handle the task of reading and writing HDFS blocks to actual files within the local file system. When a user wants to read or write a file in HDFS, the file is divided into blocks, typically 64 MB or 128 MB in size. The name node informs the client which data node holds each block, and the client communicates directly with the data nodes to access or modify the corresponding files (Lam, 2010).

Results

In Figure 3, there are two data files: one,/user/chuck/data1, which takes blocks 1, 2, and 3, and another,/user/chuck/data2, which takes blocks 4 and 5. The contents of files are distributed among data nodes. Each block has three replicas. For example, block 1 is replicated over B, C, and D data nodes, while block 4 is replicated over A, C, and D.

Data Nodes

This setup ensures that even if a data node crashes or becomes inaccessible over the network, the files remain accessible. Data nodes continuously communicate with

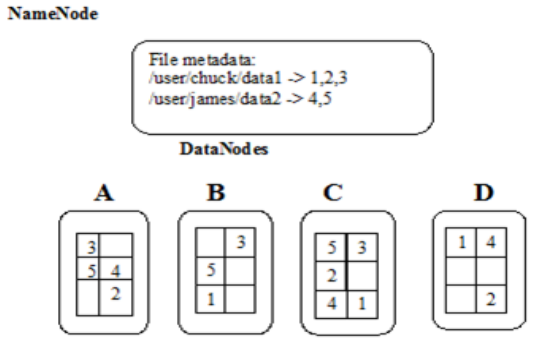

Figure 3: Name node and data nodes and their interaction in HDFS

the Name Node, reporting local changes. In turn, the Name Node issues instructions to data nodes to create, move, or delete blocks on their local disks. The Figure 3 illustrates the roles of the Name Node and data nodes, as well as their interaction within the HDFS (Figure 3).

Task trackers are Java applications whose main class is TaskTracker, which runs tasks that have been split into jobs. One task tracker per slave node, which handles many maps and reduces tasks in parallel. Task trackers always interact with job trackers. If no response is received from the task tracker within a specified time, the job tracker assumes that the task tracker has crashed and resubmits the task to another node in the cluster.

MapReduce

MapReduce ushered in a new era for Big Data technologies. It is a data processing programming model (White, 2013). It is easy scaling that involves adding and removing data nodes for data analysis across multiple computing nodes. To leverage the parallel processing capabilities of Hadoop, we should structure our query as a MapReduce job.

MapReduce works by breaking the processing into two phases 1) Map 2) Reduce. In every phase, key-value pairs <key, value> are used as both input and output, with the programmer having the flexibility to define their types. The programmer also specifies the map and reduce functions. The map function serves as the data preparation phase, where the data is set up and organized (filtering means dropping missing, suspicious, or erroneous data and transforming data) that the reduce function can aggregate.

Word Counting example

For handling multiple files, the input format list "(<String filename, String File_conent >) (<k1, v1>)" is input to map function.

Map function ignores the filename and output a list of "<String word, Integer count>"

with repeated entries. The output of map function is <" seven", 1> <" foo", 1> three times in one document, <" foo", 1> two times in another document.

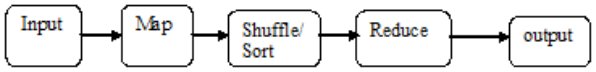

Figure 4: MapReduce logical data flow

The MapReduce framework computes the output from the map function before passing it to the reduce function. During this stage, the key-value pairs are sorted and grouped by key, a process known as shuffling. Our above key-value pairs are get arranged in sorted form as <" foo", list (1,1,1,1,1)> <" seven", 1> that is all pairs sharing the same k2 are grouped together into new key-value pair "(<k2, list<v2>)".

Above list is input to reduce function, in counting words output of reducer is <" foo", 5> <" seven", 1> it represents the total number of times the word occurs in the document. The MapReduce framework automatically gathers all "<k3, v3> "pairs and generates the output. Figure 4 shows MapReduce logical data flow.

Installing Apache HADOOP

Linux is a development and production platform for Hadoop. For Windows, install Cygwin. Download version 0.18.3 of Hadoop from hadoop.apache.org site. Since Hadoop is built in Java, one needs to install Java on the machine. "Configure the JAVA_HOME environment variable to reference the Java installation directory". Check installation with bin/ Hadoop command (Lam, 2010). Compiling, making Jar file and running MapReduce program

First type program with vi editor

Compile using command

"Javac –classpath hadoop-0.18.3-core.jar MyJob.java "

To make jar file give command

"Jar –cvf Myjob.jar MyJob.class Map.class Reduce.class "

Run program using command

"Bin/Hadoop jar MyJob.jar acite75_99.txt" output

MapReduce output will come in specified output directory (Lam, 2010)

Citation Histogram Map/Reduce Program

Acite75_99.txt is a 250MB file containing citation data from the National Bureau of Economic Research, with two columns that are comma-separated values. It contains 16

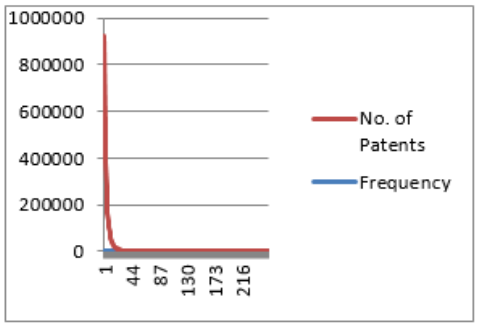

Figure 5: Distribution of citation counts

million rows. The first column lists the patent number, and the second column lists the patent number that the cited patent references. The MapReduce program calculates the number of citations each patient receives, while the Citation Histogram determines the distribution of citation counts. It reveals that many patents are cited only once, whereas a small number are cited hundreds of times. Figure 5 shows the analyzed output of the distribution of citation counts.

Discussion

Hadoop Ecosystem

In relation to Hadoop, other Apache projects include Avro™. A serialization system of data. Cassandra™ a scalable, multi-master database engineered to eliminate any single point of failure. Chukwa™ a data collection system specifically designed for monitoring large-scale distributed systems and HBase™ a distributed and scalable database that allows for structured data storage over extensive tables.

High-level layers were built on top of MapReduce, enhancing programmer productivity for domain-specific tasks. Systems were created to operate on top of Hadoop, utilizing SQL-like languages that also rely on Hadoop as the runtime layer. Hive™ is a data warehouse infrastructure built for data summarization and ad-hoc querying, while Mahout™ provides a scalable library for machine learning and data mining. Pig™ offers a high-level data-flow language and execution framework for parallel processing, and ZooKeeper™ serves as a high-performance service for coordinating distributed applications. Hadoop distributions can be obtained from leading enterprise vendors such as IBM, EMC, Microsoft, and Oracle, as well as from dedicated Hadoop providers like MAPR Technologies, Cloudera, and Hortonworks.

Conclusion

The research paper presents the characteristics of Big Data, the problems faced by distributed systems, and how Hadoop systems overcome these problems. The paper also presents the characteristics of Hadoop, its cluster architecture, and HDFS MapReduce, along with the installation and implementation of Apache Hadoop Map/Reduce for creating a citation histogram. Big Data analytics presents significant opportunities and has diverse applications in various domains, including supply chain management, academia, Indian government projects, healthcare, retail, finance, education, and many more. Big Data analytics presents significant opportunities in various fields, including manufacturing, finance, healthcare, and education, among others, in the future. Thus, the study provides practical implementation insights, enabling researchers to analyze various datasets and derive meaningful results.

Acknowledgments

Madhuri Pant thanks Vishwakarma University, Pune, for support during this research.

References

Beyer, M., & Laney, D. (2012, June 21). *The importance of "Big Data": A definition*. Gartner Research. https://www.gartner.com/doc/2057415

Borkar, V., & Carey M. J. (2012, December). Big Data technologies circa 2012. In *Proceedings of the 18th International Conference on Management of Data (COMAD)*. Pune, Computer Society of India, Mumbai, Maharashtra, IND, 12–14.

Chandhini, C., & Megana, L. P. (2013). Grid computing-a next level challenge with Big Data. *International Journal of Scientific & Engineering Research*, *4*(3), 1-5.

Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1), 107-113.

Galetsi, P., Katsaliaki, K., & Kumar, S. (2020). Big Data analytics in the health sector: Theoretical framework, techniques, and prospects. *International Journal of Information Management*, 50, 206–216. https://doi.org/10.1016/j.ijinfomgt.2019.05.003

Ghemawat, S., Gobioff, H., & Leung, S. T. (2003, October). The Google file system. In *Proceedings of the nineteenth ACM symposium on Operating systems principles*, Bolton Landing, New York, USA (pp. 29-43).

Globa, L., Svetsynska, I., & Luntovskyy, A. (2016). Case studies on Big Data. *Journal of Theoretical and Applied Computer Science*, *10*(2), 41–52.

Hickok, E. E., Chattapadhyay, S., & Abraham, S. (2017). *Big Data in governance in India: Case studies*. The Centre for Internet and Society in India. Retrieved from https://cis-india.org/internet-governance/blog/big-data-in-governance-in-india-case-studies

Himmi, K., Arcondara, J., Guan, P., & Zhou, W. (2017). Value-oriented Big Data strategy: Analysis & case study. In the *Proceedings of the 50th Hawaii International Conference on System Sciences*, 1053–1062.

Lam, C. (2010). *Hadoop in action*. Manning Publications.

Mucci, T., & Stryker, C. (2022, April 5). *What is Big Data Analytics?* https://www.ibm.com/analytics/big-data-analytics

Oracle. (n.d.). Predictive maintenance and asset availability optimization. Oracle. Retrieved August 24, 2024, from https://www.oracle.com/in/data-platform/predictive-maintenance/

Singh, M., Ghutla, B., Jnr, R. L., Mohammed, A. F., & Rashid, M. A. (2017, December). Walmart's Sales Data Analysis-A Big Data Analytics Perspective. In *2017 4th Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE)* (pp. 114-119). IEEE.

White, T. (2013). Hadoop: The definitive guide. O'Reilly Media (pp 37-40).

Ylijoki, O., & Porras, J. (2016). Conceptualizing Big Data: Analysis of case studies. Intelligent systems in accounting, finance and management, 23(4), 295-310. https://doi.org/10.1002/isaf.1393

Yousef, M., M. (2021). Big Data analytics in healthcare: A comprehensive review. *International Journal of Computer Science & Information Technology (IJCSIT)*, *13*(2). 17-28.