International Journal of Scientific Research in Computer Science and Engineering
Vol.**12**, Issue.**1**, pp.**26-33**, February **2024**
E-ISSN: 2320-7639
Available online at: www.isroset.org

Review Article

Big Data Security: Biggest Challenges, Analytics and Use cases

A. Antony prakash[1*]

[1]Dept. of Information Technology, St. Joseph's college, Trichy, India

**Corresponding Author: aantonyprakash@gmail.com*

Received: 15/Dec/2023; **Accepted**: 16/Jan/2024; **Published**: 29/Feb/2024

Abstract— Factors such as population growth and resource depletion are driving manufacturers to improve their productivity and sustainability. Consumers are becoming more informed about the things they use, therefore producers need to be flexible and open to new developments. The manufacturing industry has developed thanks to the expertise of big data analytics, or BDA. Businesses have enhanced the quality of their products while also being more adaptable in their operations. This study looks at the many advantages of big data that manufacturing companies may use to improve. It also goes through well-known manufacturers' examples and how they might set the standard for big data-enabled production. Lastly, it tackles some of the issues that must to be resolved in-order to enable big-data analytics for enterprises of all stripes, not just the best manufacturers**.**

Keywords— Big-data; Hadoop; Map Reduce; Security; HDFS

1. Introduction

Big Data has been a significant topic in the field of Material and Communication Technology (ICT) in recent years. It is clear that big data poses both substantial scientific obstacles and corporate potential. Big Data, according to McKinsey and Co1, is "the upcoming arena of productivity, competition, & innovation ". Big Data has enormous potential for growth and struggle for many enterprises, but when used properly, it can also boost productivity, innovation, and competitiveness for entire economies and industries. Big Data is the next wave of business analysis and data warehousing that has the potential to generate significant profits for businesses at a reasonable cost. [1].

Analysis is necessarily prompted by four major global trends: cloud computing, mobile computing, social networking, and Moore's law, which maintains that technology always gets cheaper. While most large organizations have been handling significant volumes of transactional data for a long time, the floodgates have suddenly opened, allowing for even more volume, speed, and variety. Big Data services and software provide value by facilitating the development of novel solutions and by fostering an inventive environment that was before unattainable [2]. Applications based on sophisticated data analysis across more general Big Data layers, semantic abstractions, or virtualization of networks and physical objects are where the value is found.

In Horizon 2020, the integration of Big Data is evident in both the Industrial Leadership sector, particularly in the action line "Content advances and data administration", as well as in the Societal Challenges sector, which pertains to the necessity of organizing data across various sectors of the economy such as healthcare, climate, transportation, energy, and others [3].

Big data open source technologies have gained significant popularity due to the undeniable capability of handling vast amounts of data in parallel. This has made it feasible to process large datasets quickly using both parallel processing and distributed computing methods. The development of Apache Hadoop, the leading framework for processing big data, was greatly influenced by its essential features and its capability to handle vast amounts of data. It is worth mentioning that the storage and analysis of large datasets have been extensively examined and compared in relation to relational database management systems (RDBMS) and the Hadoop environment, offering valuable insights into this extensively debated technology.

2. Literature Review

Big data technologies play a crucial role in facilitating more precise analysis, thereby enabling more definitive decision-making processes that yield enhanced operational productivities, cost concessions, and diminished risks for businesses. To harness the full potential of big data, it is crucial to have a robust infrastructure that can efficiently handle and analyze large vast amounts of real-time, structured and unstructured data, while also ensuring the protection of data secrecy and security. There are several technologies available in the market, provided by different vendors like Amazon, IBM, and Microsoft, and others that offer effective solutions for managing big data. When evaluating these technologies, we consider two distinct categories of technology [4].

Operational Big Data: NoSQL Big Data systems have been specifically developed to leverage the advancements in cloud computing architectures that have emerged in the last decade. These architectures enable cost-effective and efficient execution of large-scale computations. Consequently, managing operational big data workloads becomes significantly more convenient, cost-efficient, and expeditious. Certain NoSQL systems possess the capability to offer valuable insights into patterns and trends by utilizing real-time data. This can be achieved with minimal coding efforts and without the requirement of data scientists or additional infrastructure [5].

Analytical Big Data: These include technologies like MapReduce and Massively Parallel Processing (MPP) database systems, which provide analytical capabilities for in-depth and thorough analyses that may include all or a substantial amount of the data. MapReduce offers a fresh method of data analysis to enhance the features that SQL provides. Moreover, a system built on MapReduce can easily grow from one server to several powerful and inexpensive workstations. These two technological subcategories complement one another and are frequently used in tandem with one another. [6].

Hadoop has provided a solution for managing and processing large amounts of data, ranging from terabytes to petabytes [7]. Major enterprises such as Google, Facebook, and other internet giants handle numerous client requests and queries every hour. Therefore, transitioning to distribute computing has become essential in order to benefit from cost reduction and improved performance [8]. In addition, a master node can also function as a slave, meaning that the master daemon can run the slave daemons as well. Essentially, the daemons running on the master node are responsible for coordinating and managing the slave daemons on all nodes, which carry out tasks related to data storage and processing [9].

A MapReduce job typically divides the input data into chunks, which are then processed in parallel by the "map stage" and subsequently by the "reduce stage". The Hadoop framework handles the output of the Map stage, which is then used as input for the parallel reduce tasks. These input and output files are stored in a file system, with HDFS being the default choice for input datasets in the MapReduce framework [10].

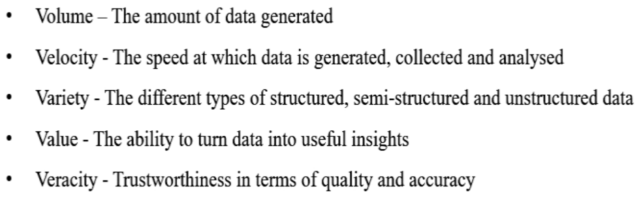
MapReduce programming is used by Hadoop to manage massive volumes of data. FIFO (First In First Out) Scheduler is the default scheduler in Hadoop, which combines many schedulers to run jobs concurrently. Other schedulers that are capable of pre-emption, non-pre-emption, and priority have also been developed. MapReduce has run across certain restrictions over time. YARN (Yet Another Resource Negotiator), the next iteration of MapReduce, was created to get over these restrictions. An introduction to YARN, a summary of Hadoop, and the scheduling techniques it uses are all covered in this paper. Additionally, the research examines MapReduce scheduling algorithms both with and without our suggested approach and with an established data locality optimization method created by Facebook. Experimental results demonstrate that our method frequently results in the fastest reaction time and highest data locality rate for map operations [11]. Moreover, in contrast to the delay algorithm, our method does not require a complex parameter alteration process.

3. Big Data

The field of data management and architecture is currently experiencing an unparalleled level of advancement. On a global scale, an astounding amount of data, exceeding 2.5 quintillion bytes, is generated on a daily basis, with approximately 90 percent of the world's data having been produced within the past few years (Forbes). As data serves as the driving force behind machine learning and the extraction of valuable insights across various industries, organizations are adopting a more diligent approach towards the collection, curation, and management of information [12].
In order to comprehend the complexities associated with big data, it is imperative to have a clear understanding of the term "Big Data" itself. Upon hearing the phrase "Big Data," one may question its distinction from the more prevalent term "data." The term "data" encompasses any unprocessed character or symbol that can be documented on a medium or transmitted through electronic signals by a computer. Nevertheless, raw data remains devoid of value until it undergoes some form of processing.

3.1 The 5 'V's of Big Data

Big Data is just a catch-all term for information that is too complicated and big to be stored in conventional databases. The "5 'V's" of Big Data are:



- Volume – The amount of data generated
- Velocity - The speed at which data is generated, collected and analysed
- Variety - The different types of structured, semi-structured and unstructured data
- Value - The ability to turn data into useful insights
- Veracity - Trustworthiness in terms of quality and accuracy

4. Hadoop Architecture

Written in Java, Hadoop is an open-source platform created by Apache. Its main purpose is to make it possible to use simple programming models to enable distributed processing of big datasets across computer clusters. The Hadoop framework functions inside a setting that enables dispersed computing and storage among computer clusters. It is built to grow from a single server to thousands of devices, each of which may handle local storage and processing.

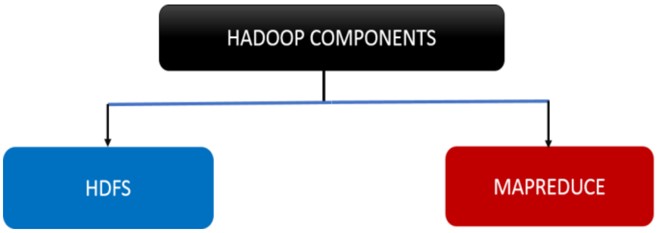


Figure 1 Hadoop Components

Currently, Hadoop has been implemented in various industries to cater to their specific requirements. Yahoo was among the early adopters of Hadoop, and since then, several leading companies such as Facebook, Twitter, and Adobe have integrated it into their architecture to benefit their organizations.

Large amounts of information can be used in the banking and securities business to track fraudulent activity, identify credit risk, maintain audit trails, detect card fraud, and manage customer data analytics to allay security worries in the financial sector. The Securities Exchange Commission (SEC) is currently employing natural language processing and network analytics to track and monitor activity through the use of massive amounts of data. [13].

The Big Data framework in the healthcare sector can help with a thorough examination of data on the premises to pinpoint availability, escalate expenses, and even monitor the progression of chronic illnesses.

Big Data is used in the media and entertainment industry to gather, examine, and derive useful customer information. To further improve corporate processes, it makes use of media material, social media aspects, and real-time analytics to spot trends. Big Data is used by the Grand Slam Wimbledon Tennis Championship to effectively provide real-time sentiment analysis to TV, mobile, and online viewers. When it comes to identifying availability, raising costs, and even tracking the advancement of chronic illnesses, the Big Data framework in the healthcare industry may assist with a comprehensive analysis of data on the premises.

The media and entertainment sector uses big data to collect, analyze, and extract valuable consumer information. It uses real-time analytics to identify trends, social media features, and media content to better enhance organizational procedures. The Grand Slam Wimbledon Tennis Championship uses big data to efficiently give TV, mobile, and online audiences real-time sentiment analysis.

The University of Tasmania, an Australian university, utilized Big Data for Higher Education in order to monitor the actions of 26,000 individuals and oversee their advancement. In a similar vein, it was employed to assess a teacher's efficacy based on the learning experiences, grades, behavior, demographics, and other factors of the students.

Within the Manufacturing and Natural Resources sector, Big Data can boost the productivity of the supply chain. There is a tremendous quantity of untapped data with higher volume and velocity in both industries. By incorporating big data technologies, systems can become more dependable, efficient, and profitable for organizations.

4.1 HDFS

Hadoop is accompanied by a distributed file system known as HDFS. Within HDFS, data is distributed across multiple machines and replicated to guarantee resilience against failures and facilitate parallel application availability. This approach proves to be economically advantageous as it leverages commodity hardware. It incorporates the notion of blocks, data nodes, and node names.

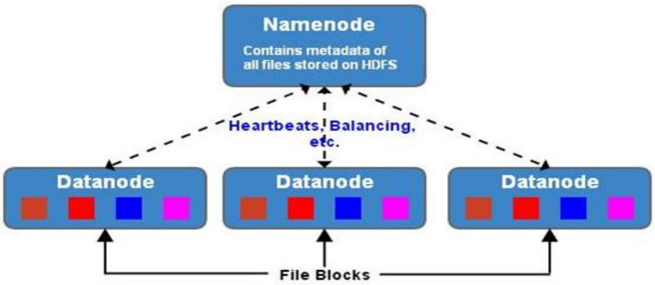

Namenode
Contains metadata of all files stored on HDFS
Heartbeats, Balancing, etc.
Datanode
Datanode
Datanode
File Blocks

Figure 2. Architecture of HDFS

Blocks: A Block represents the smallest unit of data that can be accessed for reading or writing. The default size of a block in the Hadoop Distributed File System (HDFS) is 128 MB, but it can be modified as required. In HDFS, files are divided into chunks that correspond to the block size, and these chunks are treated as separate entities. Unlike conventional file systems, if a file in HDFS is smaller than the block size, it will not occupy the entire block's size. For instance, if a file of 5 MB is stored in an HDFS with a block size of 128 MB, it will only utilize 5 MB of space. The large block size in HDFS is primarily designed to minimize the cost of seeking.

Name Node: The Hadoop Distributed File System (HDFS) operates using a master-worker pattern, in which the name node serves as the master. The name node functions as the controller and manager of HDFS, possessing knowledge of the status and metadata of all files within the system. This metadata includes file permissions, block locations, and names. Due to the small size of the metadata, it is stored in the memory of the name node, enabling faster data access. Additionally, as the HDFS cluster is accessed by multiple clients simultaneously, this single machine efficiently handles all the associated information. Consequently, the name node is responsible for executing file system operations such as opening, closing, and renaming files.

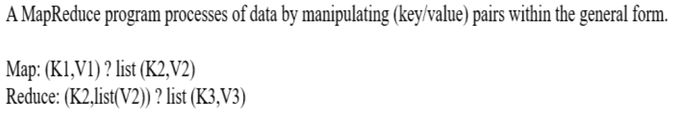
Data Node: The data nodes perform the tasks of storing and retrieving blocks upon receiving instructions from either the client or the name node. They periodically communicate with the name node, providing a list of the blocks they are currently storing. Additionally, the data nodes, being commodity hardware, are responsible for tasks such as block creation, deletion, and replication, as directed by the name node.

4.2 Map Reduce

A MapReduce is a information processing technology that is used for distributed, parallel data processing. Based on a Google paper titled "MapReduce: Simplified Data Processing on Large Clusters," it was created in 2004. The reducer phase and the mapper phase are the two stages of the MapReduce paradigm. The input is given as a key-value pair during the mapper step. The reducer then receives the mapper's output as input. Only once the mapping step is finished does the reduction begin to operate. In a similar vein, the reducer's output is the ultimate output and it too gets input in key-value format.

Steps in Map Reduce

The map function processes data in the form of pairs and produces a list of <key, value> pairs. It should be noted that the keys may not be unique in this particular scenario. The Hadoop architecture applies the sort and shuffle operations to the output of the Map function. This sort and shuffle process operates on the list of <key, value> pairs and generates unique keys along with a list of values associated with each unique key, denoted as <key, list (values)>. The output of the sort and shuffle phase is then sent to the reducer stage. The reducer function performs a specified operation on the list of values corresponding to each unique key. The final output will be in the form of <key, value> pairs, which will be stored or displayed as required.



A MapReduce program processes of data by manipulating (key/value) pairs within the general form.

Map: (K1,V1) ? list (K2,V2)
Reduce: (K2,list(V2)) ? list (K3,V3)

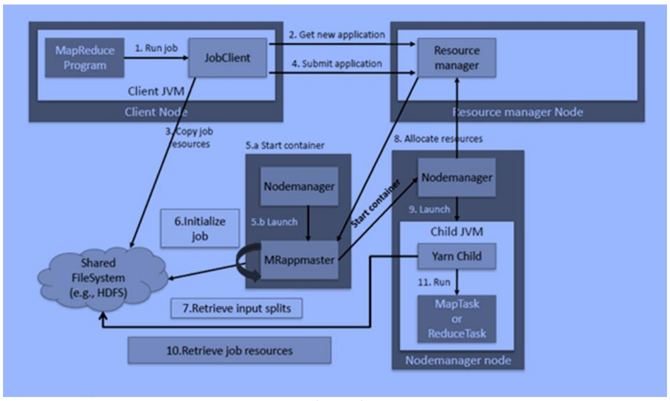


Figure 3. Workflow for MapReduce

5. Security In Hadoop Environment

Hadoop ecosystem includes several tools to support Hadoop security

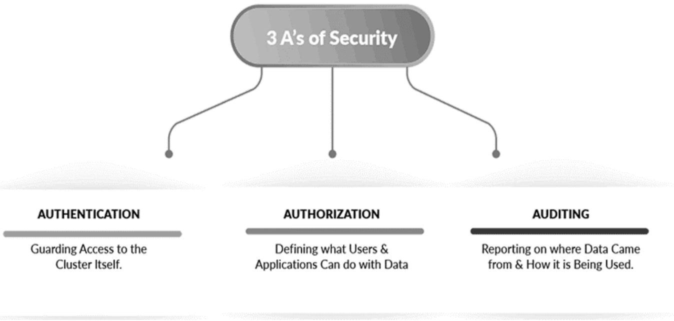

3 A's of Security
AUTHENTICATION
Guarding Access to the Cluster Itself.
AUTHORIZATION
Defining what Users & Applications Can do with Data
AUDITING
Reporting on where Data Came from & How it is Being Used.

Figure 4 Security in Hadoop

5.1 Kerberos

Kerberos serves as an authentication protocol that has become the standard for implementing authentication within the Hadoop cluster. By default, Hadoop lacks any form of authentication, which can have significant repercussions for corporate data centres. To address this limitation, the Hadoop Ecosystem introduced Kerberos, a secure method for authenticating users. Kerberos is a network authentication protocol that originated at MIT and utilizes "tickets" to enable nodes to establish their identities [14In Hadoop, the Kerberos protocol is used to verify that the person submitting a request is who they say they are. Every Hadoop node uses Kerberos for mutual authentication while operating in secure mode. This implies that throughout communication, both nodes confirm the legitimacy of the other node. Secret-key cryptography is used by Kerberos to provide client-server application authentication.

Authentication: In the Kerberos protocol, the client initially undergoes authentication to verify its identity with the authentication server. Subsequently, the authentication server issues a Ticket-Granting Ticket (TGT) to the client, which is timestamped.

Authorization: Following the authentication process, the client employs the TGT to solicit a service ticket from the Ticket-Granting Server. This service ticket enables the client to access specific services within the system.

Service Request: Upon obtaining the service ticket, the client is then able to directly engage with the Hadoop cluster daemons, including the Namenode and Resource Manager.

5.2 Knox

Knox serves as a REST API-based perimeter security gateway that executes authentication, monitoring, auditing, authorization management, and policy enforcement on Hadoop clusters. Its primary function is to authenticate user credentials, typically against LDAP and Active Directory. It permits access to the Hadoop cluster only to those users who have been successfully authenticated [15].

5.3 Ranger

The authorization system is designed to grant or deny access to various Hadoop cluster resources, including HDFS files and Hive tables, based on predetermined policies. It is expected that users requesting access through Ranger have already undergone authentication. Additionally, this system offers distinct authorization capabilities for different Hadoop components, such as YARN, Hive, and HBase.

5.4 Apache Sentry

A module called Apache Sentry is made for Hadoop that allows for multi-tenancy management and fine-grained, role-based authorization. Its main job is to control who can access data and metadata in a Hadoop cluster by giving authorized users and apps different levels of authorization. Sentry is quite standardized and compatible with various Hadoop data models. It is a flexible tool that lets users set up authorization rules to verify user or application requests to access Hadoop resources. Sentry's ultimate objective is to function as a pluggable authentication engine for different Hadoop components, such as HDFS, Apache Hive, Apache Solr, and Impala. [16].

6. Challenges In Big Data In Industry

The amount of information gathered by organizations is constantly increasing at a rapid pace. According to estimates by IDC, the overall quantity of stored data doubles approximately every two years. By the year 2025, the world is projected to generate an astounding 463 exabytes of data on a daily basis [17].

The primary objective of Big Data initiatives is to utilize these vast information repositories to uncover concealed insights and patterns that can aid in making informed business decisions of all types. The potential benefits are significant, but organizations encounter significant obstacles in steering their [18] Big Data strategies towards success. In this article, we will examine the top seven challenges that Big Data faces and explore potential solutions to overcome them.

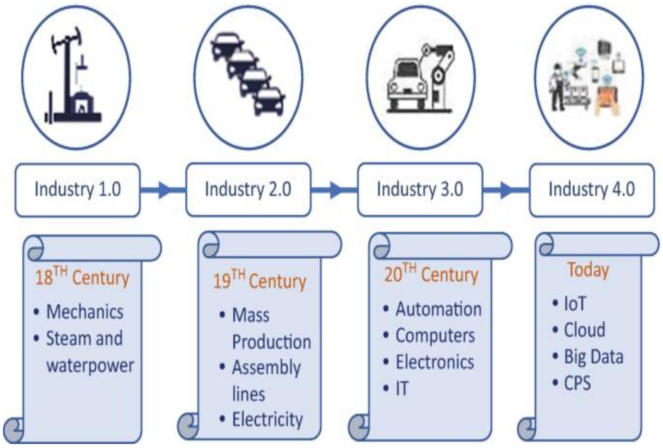


Figure 5 Challenges in big data in industry

6.1 Big Data Challenges in the Healthcare Industry

- Improve the efficacy of diagnostic procedures.
- Prescribe preventive medicine and promote health.
- Delivering results to physicians in a digital format.
- Employing predictive analysis to unveil previously undiscovered patterns.
- Offering real-time monitoring.

Technical Challenges

The objective is to create a data exchange and interoperability architecture that enables the delivery of personalized care to patients. The aim is to develop an AI-based analytical platform that can effectively integrate data from multiple sources. The goal is to propose a predictive and prescriptive modelling platform for physicians, which will help bridge the semantic gap and improve the accuracy of diagnoses.

6.2 Big Data Challenges in Security Management

The security challenges that arise in the context of Big Data storage are multifaceted and stem from the diverse origins of the data sources. It is crucial to acknowledge that the data is collected from a wide range of sources, some of which may not adhere to the established safety and compliance standards of the organization. The combination of data from sources that were not initially meant to be merged can potentially jeopardize privacy and security [19].

Considering the substantial amount of sensitive and confidential data stored in Big Data environments, they become highly appealing to hackers and cybercriminals. Hence, it is of utmost importance to integrate security measures during the initial phases of the architectural planning process. Attempting to implement comprehensive protection at a later stage is a challenging task.

6.3 Big Data Challenges in Hadoop-Delta Lake Migration

Data Migrator is a comprehensive automated solution that facilitates the seamless transfer of on premise HDFS data, Hive metadata, local file system, or cloud data sources to either cloud or on-premises environments, even during ongoing modifications to these datasets. This tool does not necessitate any modifications to applications or business operations [20]. The process of transferring data of any magnitude can commence promptly and be executed without any interruptions to the production system or disruptions to business activities, all while ensuring there is no risk of data loss.

6.4 Big Data Challenges in Cloud Security Governance

Cloud security governance is a management framework that enables an organization to achieve its business objectives while facilitating effective and efficient security operations and management in the cloud environment. The operational procedures, performance standards, executive directives, structures, and metrics that make up this framework's hierarchy are designed to optimize an organization's business value when implemented. Leaders can get help answering questions with cloud security governance such as:

- Are our security investments yielding the anticipated outcomes?
- Do we comprehend our security risks and their impact on our business?
- Are we progressively reducing security risks to acceptable levels?
- Have we instilled a security-conscious culture within the organization?

This list is not exhaustive, and there may be additional questions that can be answered through Cloud Security Governance. It is evident that Cloud Security Governance plays a crucial role in ensuring the security of cloud-based operations.

7. Business Intelligence (Bi) Tools To Help With Big Data

The industry of big data has a significant influence on almost every other industry. In order for companies to create products that are based on big data, they require Business Intelligence (BI) tools to facilitate digital transformation and data discovery. It is more advantageous for business intelligence (BI) tools to be seamlessly incorporated with open-source frameworks like Hadoop. Hadoop, developed by Apache, is a comprehensive set of open-source software utilities that facilitates the efficient processing of massive data volumes across a distributed network of computers. Presented below are some of the most widely used BI tools that can assist with big data.

7.1 Datameer

Datameer utilizes its user-friendly interface to model data, merging various data sets, analytics, and other company system assets into a unified dashboard. The key advantage of Datameer lies in its scalability and interactivity, which facilitates the training of employees even in the absence of coding or analytics experience.

7.2 QlikSense

QlikSense boasts a distinctive touchscreen interface that is particularly suitable for enterprises whose users operate in mobile environments. Its AI-driven search and conversational analytics functionality is exclusive to Qlik, providing companies with a novel means of interacting with their databases to uncover deeper insights beyond conventional oversight.

7.3 Zoho Analytics

Zoho Analytics offers visualizations while seamlessly synchronizing with other interconnected devices and effortlessly integrating with various programs. The synchronization feature holds significant importance in these programs, particularly when data silos persist in a fragmented state. It gathers information to generate comprehensive reports and presents them in a unified and concise dashboard.

8. Use Cases In Several Businesses

8.1 Banking and Finance (Fraud Detection, Risk & Insurance, and Asset Management)

Futuristic banks and financial institutions are harnessing the potential of big data in diverse manners, encompassing the exploration of new markets and market opportunities, as well as the mitigation of fraud and management of investment risks. These entities possess the capability to effectively employ big data analytics as a formidable tool to attain a competitive edge.

Big data analytics is expected to grow at a Compound Annual Growth Rate (CAGR) of 22.97% between 2021 and 2026, based on recent study. This increase might be linked to the growing amount of data being produced and the tightening government restrictions, which are fuelling the need for big data analytics across a range of industries.

8.2 Accounting

Data is the fundamental essence of the accounting profession, and the utilization of big data analytics within the realm of accounting undoubtedly yields enhanced value for accounting enterprises. The accounting sector encompasses a multitude of activities, including diverse forms of audits, ledger verification and maintenance, transaction administration, taxation, financial planning, and so forth.

8.3 Aviation

According to studies, it has been projected that the aviation analytics market will reach a value of 3 billion USD by the year 2025, with a compound annual growth rate (CAGR) of 11.5% over the forecast period.

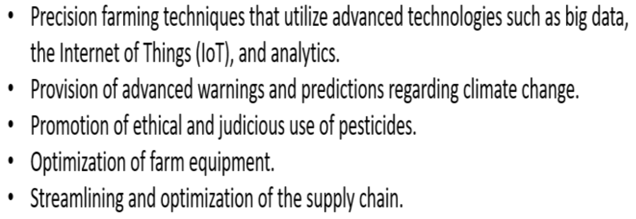
The primary factors driving this growth in the aviation market include the increasing demand for optimized business operations, the impact of the COVID-19 outbreak on normal aviation operations, and the occurrence of mergers, acquisitions, and joint ventures.

Big data analytics has a lot of promise, as seen by recent developments and changes in the Original Equipment Manufacturer (OEM) and user segment of the aviation industry. Cloud-based real-time data collecting and analytics is one especially exciting prospect that requires the use of many data models.

Furthermore, big data analytics holds immense potential within the airline industry as well. It can enhance various aspects of airline operations, ranging from maintenance and resource distribution to flight safety, flight services, loyalty programs, and route optimization.

8.4 Agriculture

According to projections by the United Nations, the global population is expected to reach 9.8 billion by 2050. To address the food requirements of such a large population, it is imperative to make adjustments to the agricultural industry. However, climate change has not only rendered a significant portion of farmland unsuitable for farming, but it has also disrupted rainfall patterns and depleted water sources. As a result, farmers need to not only focus on increasing crop production but also enhance various other farming-related activities. The utilization of big data analytics can offer valuable support to stakeholders in the agriculture and agribusiness sectors in the following ways:



- Precision farming techniques that utilize advanced technologies such as big data, the Internet of Things (IoT), and analytics.
- Provision of advanced warnings and predictions regarding climate change.
- Promotion of ethical and judicious use of pesticides.
- Optimization of farm equipment.
- Streamlining and optimization of the supply chain.

8.5 Biomedical Research and Healthcare: COVID-19 Management, Genomic Medicine, and Cancer

According to recent statistics, it has been projected that the big data analytics market in the healthcare sector will reach approximately 67.82 billion USD by the year 2025. The healthcare industry is a vast sector that generates a substantial amount of data, which holds immense importance for patients, medical institutions, insurance companies, government entities, and research as well.

Table 1 Attributes of Medical dataset

Sl. No	Attributes	Explanation
1	Id	ID Number
2	Diagnosis	The diagnosis of breast tissues where M represents malignant and B represents benign
3.	Radius-mean	Mean of distances from center to pints on the perimeter
4.	Texture-mean	Standard deviation of gray- scale values
5.	Perimeter- mean	Mean size of the core tumor
6.	Area- mean	-
	smoothness mean	Mean of local variation in radius lengths
7.	compactness mean	Mean of perimeter^2 / area - 1.0
8.	concavity mean	Mean of severity of concave portions of the contour
9.	concave points mean	mean for number of concave portions of the contour
10.	symmetry_mean	-
11	Fractal dimension	Coastline approximation

By conducting thorough analysis of these extensive data sets, big data analytics has the potential to not only assist medical researchers in formulating more precise and effective treatment plans, but also in procuring medical supplies from various locations across the globe.

Furthermore, big data analytics can contribute to the improvement of organ donation processes, enhancement of treatment facilities, development of superior medications, and prediction of the occurrence and severity of pandemics or epidemics, thereby enabling effective containment measures.
In conclusion, the implementation of big data analytics in the healthcare industry offers numerous advantages and opportunities for advancement.

8.6 Manufacturing & Supply Chain Management

Supply chains and the manufacturing industry are changing dramatically as the globe gets closer to the fourth industrial revolution. In order to improve their business equity, boost revenues, and find hidden patterns and market trends from massive data sets, manufacturers are looking for efficient ways to use the enormous volumes of data they produce.

There are three key areas in the engineering industry where big data analytics plays a crucial role:
a. Predictive Maintenance - This involves predicting equipment failure, identifying potential issues in manufacturing units and products, and taking proactive measures to prevent them.
b. Operational Efficiency - This involves analysing and assessing production processes, gathering customer feedback, and forecasting future demand to improve operational efficiency.
c. Production Optimization - This involves optimizing production lines to reduce costs, increase revenue, and identifying processes or activities that cause delays in production.

8.7 Media and Entertainment

The application of big data analytics in the media and entertainment sector can yield important information about the various content preferences, brand cost/subscription concepts, and reception. In addition, the analysis of consumer behaviour and content consumption can be used to provide personalized content recommendations and acquire knowledge for the creation of new programs. Furthermore, the evaluation of consumer mood, market segmentation, and market potential can help with critical business decisions that maximize profits and reduce the risk of creating unsatisfactory or uneven content.

Each manuscript should contain a conclusion section within 250-450 words which may contain the major outcome of the work, highlighting its importance, limitation, relevance, application and recommendation. Conclusion should be written in continuous manner with running sentences which normally includes main outcome of the research work, its application, limitation and recommendation. Do not use any subheading, citation, references to other part of the manuscript, or point list within the conclusion. In last paragraph author describes the future Scope for improvement.

9. Conclusion and future work

The future of the business site is characterized by uncertainties and intense competition, and there is no resource more dependable and trustworthy than data. Big data analytics provides businesses with robust capabilities for data mining, management, and processing, enabling them to effectively utilize historical data as well as continuously generated organizational data. With its capacity to inform present and future business decisions, big data analytics stands as one of the most reliable technologies for businesses of all sizes and industries. However, the adoption and implementation of big data analytics pose significant challenges, particularly in terms of resources and capital. Therefore, the most prudent approach to embarking on this revolution is to engage the services of reputable big data consulting firms, such as DataToBiz, which can assist in identifying, comprehending, and addressing your specific big data analytics requirements.

References

[1] Puneet Singh Duggal, Sanchita Paul, “Big Data Analysis: Challenges and Solutions”, In RGPV, Page No.269-276, 2013.

[2] Abdul Ghaffar Shoro and Tariq Rahim Soomro, “Big Data Analysis: Ap Spark Perspective”, Global Journal of Computer Science and Technology: C Software & Data Engineering, Volume 15, Issue 1, and Version 1.0, 2015.

[3] Mr.S.S. Aravinth, Ms.A.Haseenah Begam, Ms.Shanmuga priya, Ms.S. Sowmya and Mr.E. Arun, “An Efficient HADOOP Frameworks SQOOP and Ambari for Big Data Processing”, International Journal for Innovative Research in Science & Technology, Volume 1, Issue 10, 2015.

[4] Chun Wei Tsai, Chin Feng Lai, Han Chieh Chao and Athanasios V. Vasilakos, “Big Data Analytics: A Survey”, Journal of Big Data, https://doi.org/10.1186/s40537-015-0030-3, 2015.

[5] DunrenChe, MejdlSafran, and Zhiyong Peng, “From Big Data to Big Data Mining: Challenges, Issues, and Opportunities”,

DASFAA Workshops 2013, LNCS 7827, pp.1–15, 2013.

[6] Russom, P.: Big Data Analytics. In: TDWI Best Practices Report, pp. 1–40 (2011)

[7] Jeffrey Shafer, Scott Rixner, and Alan L. Cox." The Hadoop Distributed Filesystem: Balancing Portability and Performance". DOP is March 30, 2010.

[8] Kenneth Wottrich, and T. Bressoud, "The performance characteristics of mapreduce applications on scalable clusters", in Proceedings of theMidstates Conference on Undergraduate Research in Computer Science and Mathematics (MCURCSM), Denison University, Granville,USA, Nov. 2011.

[9] S. Loughran, J.M.A. Calero, A. Farrell, J. Kirschnick, and J.Guijarro, "Dynamic deployment of mapreduce architecture in the cloud."IEEEInternet Computing, vol. 16, no. 6, pp. 40-50, Dec. 2012.

[10] T. White, "MapReduce and the hadoop distributed file system", in Hadoop: The definitive guide, 1st edition, O'Reilly Media, Inc., Yahoo press, 2012.

[11] A.Elsayed, O. Ismail, and M.E. El-Sharkawi, "MapReduce: state-of-the-art and research directions", International Journal of Computer andElectrical Engineering, vol. 6, no. 1, pp. 34-39, Feb. 2014.

[12] Kanchan A. Khedikar, "Role of Cloud Computing in Big Data Analytics ,Using MapReduce Component of Hadoop", Maharashtra, India., International Journal of Innovations in Engineering and Technology (IJIET), Volume 4 Issue 1, pp.87 ISSN: 2319 – 1058, 2014.

[13] Lin Wei-wei, Liu Bo, "Hadoop Data Load Balancing Method Based on Dynamic Bandwidth Allocation", Journal of South China University of Technology (Natural Science Edition) , pp. 40(9):42-47, 2012.

[14] B. Saraladevi, N. Pazhaniraja, P. Victer Paul, M.S. SaleemBasha, P. Dhavachelvan, "Big Data and Hadoop-a Study in Security Perspective", Procedia Computer Science, ISSN 1877-0509, Volume 50 , pp. 596-601, 2015.

[15] D. Shehzad, Z. Khan, H. Dağ, and Z. Bozkuş, "A novel hybrid encryption scheme to ensure Hadoop based cloud data security", International Journal of Computer Science and Information Security (IJCSIS), Volume 14, 2016.

[16] J.S. Hurwitz, A. Nugent, F. Halper, M. Kaufman, "Big data for dummies Wiley", USA 2013.

[17] A.Antony prakash, "Data Security On Hadoop Distributed File System Using Rsa Encryption", IJRAR, Volume 6, Year 2019.

[18] Yusuf Perwej," The Hadoop Security in Big Data: A Technological Viewpoint and Analysis", Isroset-Journal (IJSRCSE), Vol.7, Issue.3, pp.10-14, Jun-2019.

[19] A.Antony prakash, "A Novel Hadoop Map Reduce Paradigm Using An Effective Epca Approach" , Journal of Analysis and Computation (JAC), Volume , Year 2019,

[20] Mantripatjit Kaur, Anjum Mohd Aslam, "Big Data Analytics on IOT: Challenges, Open Research Issues and Tools, Vol.6 , Issue.3 , pp.81-85, Jun-2018

AUTHORS PROFILE

Dr. A. Antony Prakash is currently working as an Assistant professor, Department of Information technology, St. Joseph's college Trichy, Tamilnadu, India. He has more than 12 years of teaching experience at PG level and more than 8 years of Experience in Research. He has published more than 20 articles in National and International Journals and also attended many national and international conferences, seminars and workshops. His field of specialization is computer science and the main area of research is Big Data and Data mining