RESEARCH ARTICLE OPEN ACCESS

Reviewing the Security, Optimization, and Scalability Challenges in Big Data Ecosystems: Insights into Hadoop and Cloud Computing Solutions

Vedaprada Raghunath*
Visvesvaraya Technological University (VTU), Public university in Belgaum, Belgaum, Karnataka, India
email id: vedapradaphd@gmail.com

Abstract:

This research paper explores the integration of big data, cloud computing, the Internet of Things (IoT), and artificial intelligence (AI) across various industries, highlighting their transformative impact on fields such as healthcare, urban planning, environmental sustainability, and business management. It examines how big data frameworks like Hadoop and Spark, along with cloud-based infrastructures, facilitate the storage, processing, and analysis of vast datasets, enabling real-time decision-making and innovative solutions. The study also addresses the key challenges associated with these technologies, including data privacy, security, interoperability, and efficient load balancing. It emphasizes the importance of emerging solutions such as differential privacy, secure data management, and machine learning models to address these concerns. Furthermore, the paper discusses the environmental implications of big data processing and proposes sustainable computing strategies to mitigate the carbon footprint of large-scale data operations. The findings underscore the need for continued advancements in cloud frameworks, data management systems, and machine learning algorithms to unlock the full potential of these technologies while ensuring their responsible application across industries.

***Key words**: Big Data, Cloud Computing, Security, Hadoop*

I. INTRODUCTION

Cloud computing has revolutionized how businesses and individuals handle data storage, processing, and accessibility, providing unmatched scalability, flexibility, and cost efficiency. Yet, as cloud technology has advanced, so have the security challenges associated with managing vast volumes of data on these platforms [1,2]. This extensive accumulation of big data within cloud environments introduces specific vulnerabilities, complicating efforts in data protection, access control, and security monitoring. Traditional security methods often struggle to effectively secure this data due to the large scale and shared nature of cloud infrastructure, where data from multiple sources frequently coexists and interconnects [3,4].

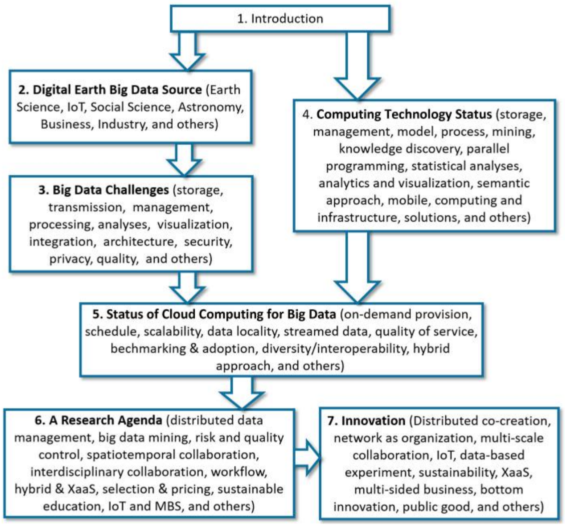

1. Introduction
2. Digital Earth Big Data Source (Earth Science, IoT, Social Science, Astronomy, Business, Industry, and others)
4. Computing Technology Status (storage, management, model, process, mining, knowledge discovery, parallel programming, statistical analyses, analytics and visualization, semantic approach, mobile, computing and infrastructure, solutions, and others)
3. Big Data Challenges (storage, transmission, management, processing, analyses, visualization, integration, architecture, security, privacy, quality, and others)
5. Status of Cloud Computing for Big Data (on-demand provision, schedule, scalability, data locality, streamed data, quality of service, bechmarking & adoption, diversity/interoperability, hybrid approach, and others)
6. A Research Agenda (distributed data management, big data mining, risk and quality control, spatiotemporal collaboration, interdisciplinary collaboration, workflow, hybrid & XaaS, selection & pricing, sustainable education, IoT and MBS, and others)
7. Innovation (Distributed co-creation, network as organization, multi-scale collaboration, IoT, data-based experiment, sustainability, XaaS, multi-sided business, bottom innovation, public good, and others)

Fig 1. Tackling of Big Data challenges with cloud computing for innovation.[3]

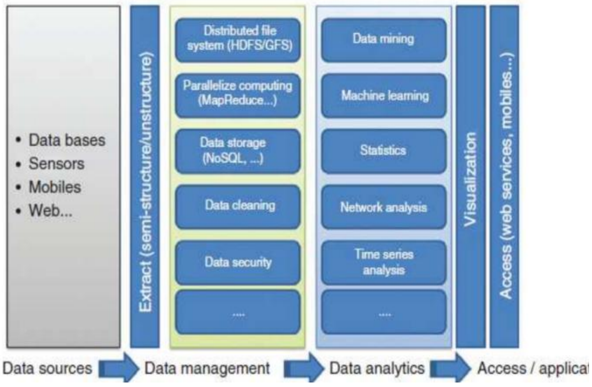


Fig 2. Big Data framework [4].

Big data analytics within cloud computing now offers promising solutions to address security concerns, harnessing the extensive processing capabilities of cloud systems. These analytical tools can monitor for suspicious activity, detect anomalies, and identify potential threats before they become critical [3-5]. Nonetheless, managing big data within the cloud also brings challenges around data privacy, the risk of data breaches, and the intricate task of securing data across distributed networks [5,6]. To navigate these complexities, there is an increasing need for advanced security frameworks that integrate big data analytics and machine learning, allowing for proactive monitoring and data protection while ensuring compliance with rigorous regulatory standards [7-9].

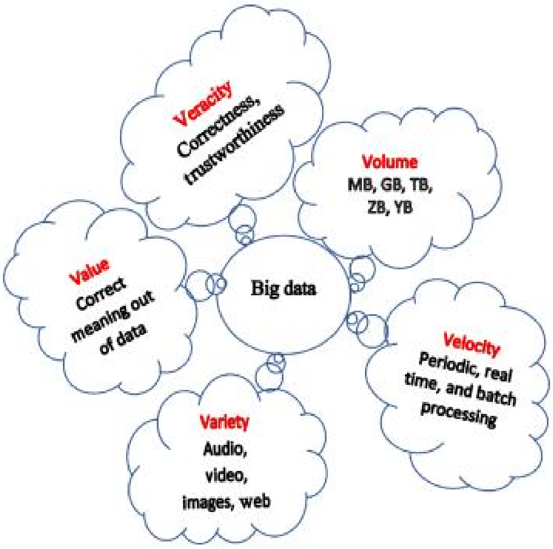

Veracity
Correctness, trustworthiness
Volume
MB, GB, TB, ZB, YB
Value
Correct meaning out of data
Big data
Velocity
Periodic, real time, and batch processing
Variety
Audio, video, images, web

Fig 3. Five Vs of big data.[7]

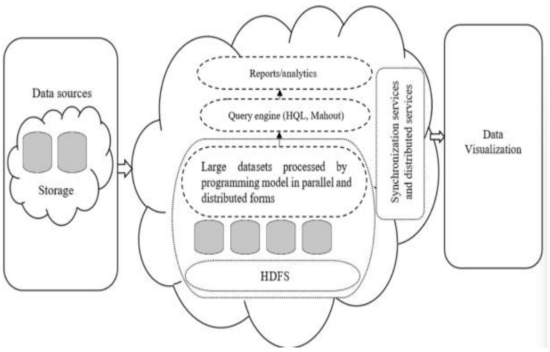

Data sources
Storage
Reports/analytics
Query engine (HQL, Mahout)
Large datasets processed by programming model in parallel and distributed forms
HDFS
Synchronization services and distributed services
Data Visualization

Fig 4. Big data and cloud computing [7]

Big data is characterized by five primary attributes: volume, velocity, variety, value, and veracity, which capture its unique complexities[9-14]. Initially defined by the three Vs—volume, velocity, and variety—these characteristics were introduced by Gartner to emphasize the distinct challenges inherent to big data. Over time, data architecture has evolved, and with it, the scope of big data characteristics expanded to include value and veracity [15-17]. Volume refers to the immense quantities of data generated by various sources such as sensors, social media, and smartphones. For example, in 2012, approximately 2.5 exabytes (EB) of data were created every day. By 2013, this volume doubled to 4.4 zettabytes (ZB), and by 2020, it reached an astonishing 40 ZB. Velocity highlights the rapid speed at which data is produced and processed, a trend driven by millions of connected devices coming online daily. Platforms like YouTube showcase the high-speed data generation characteristic of today's digital landscape[18-20].

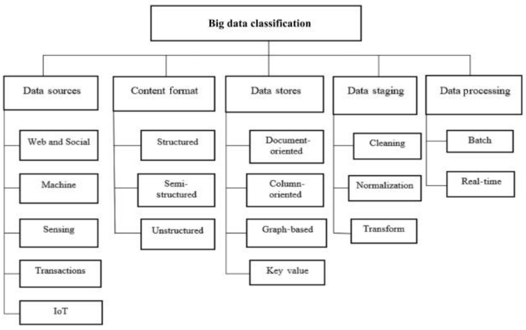


Fig 5. Classification of big data [12]

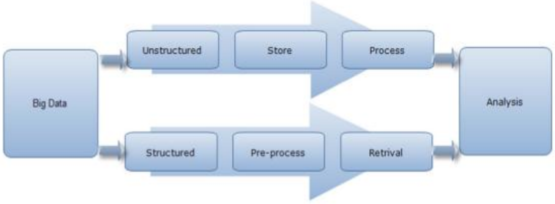


Fig 6. Transforming big data for analysis.[12]

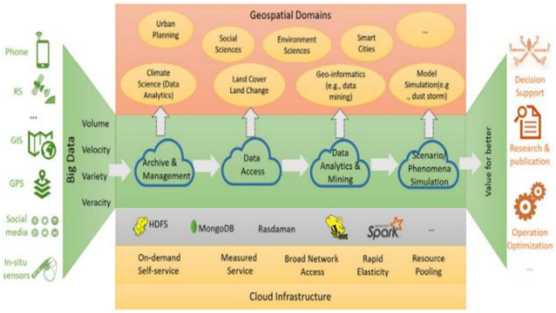


Fig 7. Big Data to address the 4Vs to obtain Value for better decision support, research, and operations for various geospatial domains.[14]

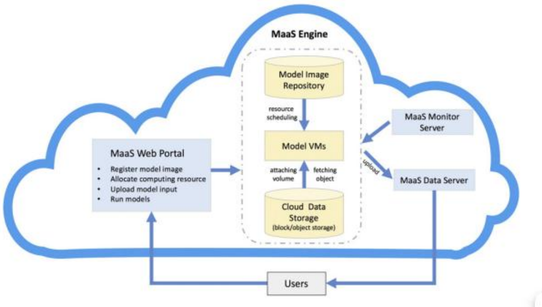


Fig 8. The cloud-based service-oriented workflow system for climate model study.[14]

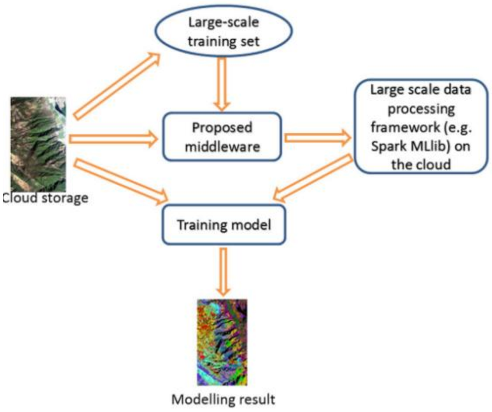


Fig 9. The role the proposed middleware plays in the model building process[14]

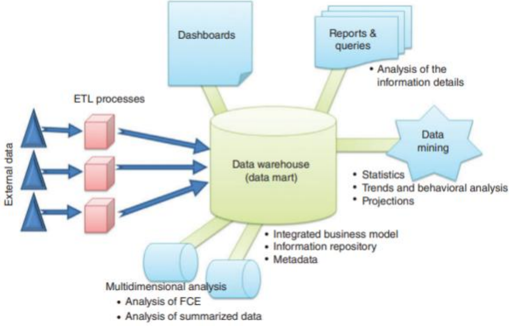

Dashboards
Reports & queries
• Analysis of the information details
ETL processes
External data
Data warehouse (data mart)
Data mining
• Statistics
• Trends and behavioral analysis
• Projections
• Integrated business model
• Information repository
• Metadata
Multidimensional analysis
• Analysis of FCE
• Analysis of summarized data

Fig 10. Business Intelligence structure.[18]

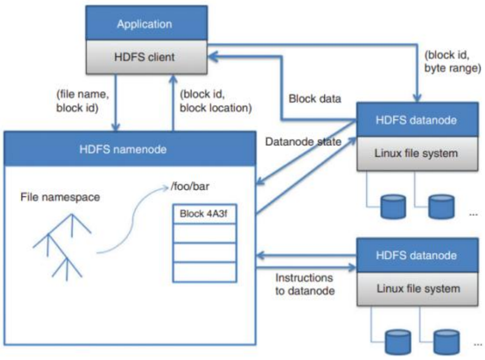
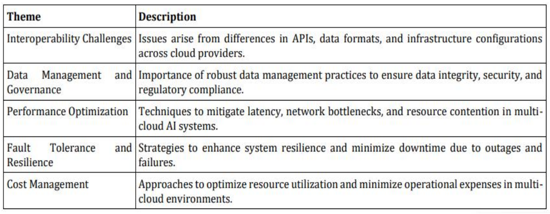


Fig 11. The architecture of Hadoop-distributed file system (HDFS).[18]



Theme	Description
Interoperability Challenges	Issues arise from differences in APIs, data formats, and infrastructure configurations across cloud providers.
Data Management and Governance	Importance of robust data management practices to ensure data integrity, security, and regulatory compliance.
Performance Optimization	Techniques to mitigate latency, network bottlenecks, and resource contention in multi-cloud AI systems.
Fault Tolerance and Resilience	Strategies to enhance system resilience and minimize downtime due to outages and failures.
Cost Management	Approaches to optimize resource utilization and minimize operational expenses in multi-cloud environments.

Fig 12. Key Issues and Strategic Solutions for AI Implementation in Multi-Cloud Environments[19]

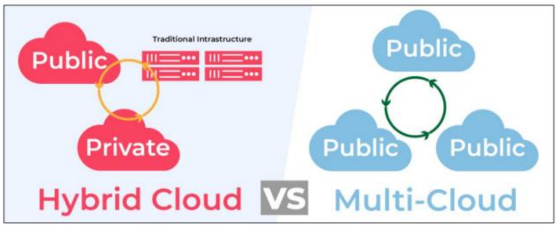


Fig 13. Comparison between Hybrid Cloud and Multi-cloud[19]

Big data also demonstrates high variety, as it is generated in diverse formats such as images, videos, audio, documents, and text across multiple platforms.

This data can be structured, semi-structured, or unstructured, adding complexity to its management and analysis. Beyond these attributes, value and veracity are critical in determining data's utility and reliability [21-22].

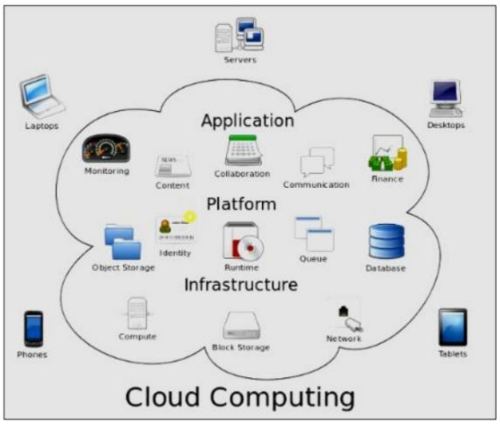


Fig 14. Cloud Computing [21]

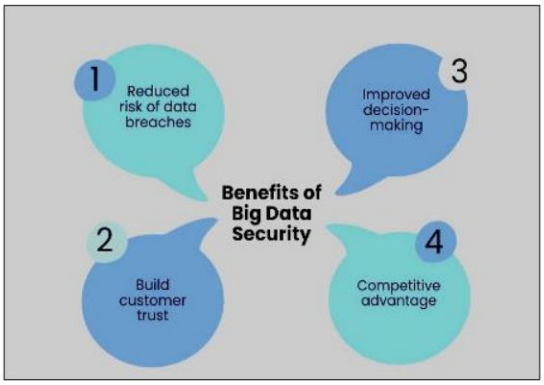


Fig 15. Benefits of Big Data security[21]

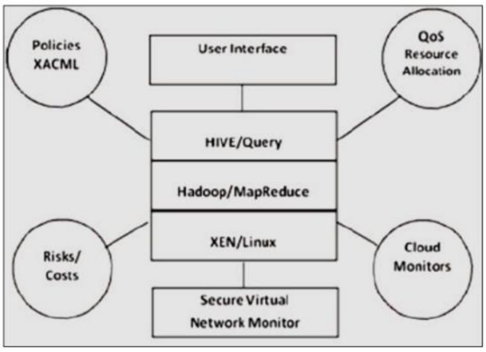


Fig 16. Layered framework for assuring cloud[21]

Value speaks to the ability of big data to yield meaningful insights, driving its worth in decision-making processes. Veracity, on the other hand, concerns the quality, accuracy, and trustworthiness of data [23-28]. High veracity is essential for ensuring that big data provides a solid foundation for analysis and is a reliable source for insights. As data continues to grow in both volume and complexity, understanding these characteristics is fundamental to unlocking its potential while addressing the challenges it presents [28-30].

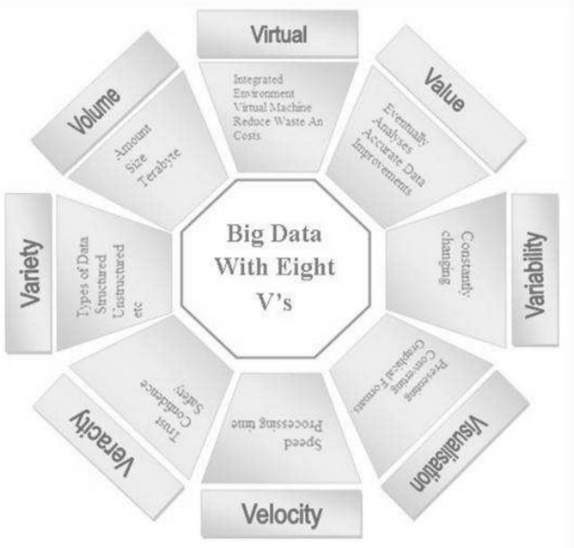
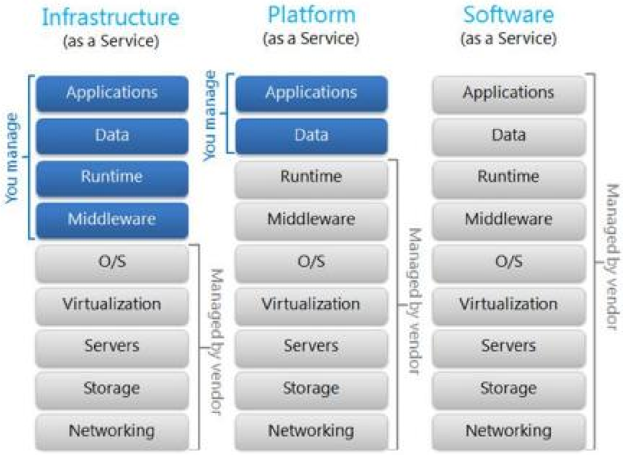

Virtual
Integrated Environment Virtual Machine Reduce Waste An Costs
Value
Eventually Analyses Accurate Data Improvements
Variability
Constantly changing
Visualisation
Presenting Converting Graphical Formats
Velocity
Speed Processing time
Veracity
Trust Confidence Safety
Variety
Types of Data Structured Unstructured etc
Volume
Amount Size Terabyte
Big Data With Eight V's

Fig 17. Characteristics Of Big Data.[24]



Infrastructure (as a Service)	Platform (as a Service)	Software (as a Service)
Applications	Applications	Applications
Data	Data	Data
Runtime	Runtime	Runtime
Middleware	Middleware	Middleware
O/S	O/S	O/S
Virtualization	Virtualization	Virtualization
Servers	Servers	Servers
Storage	Storage	Storage
Networking	Networking	Networking

Fig 18. Cloud computing Models.[24]

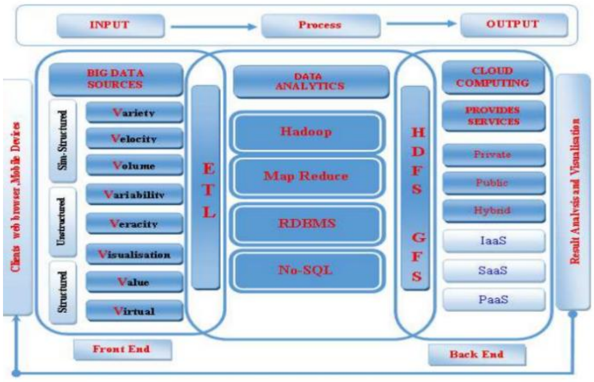
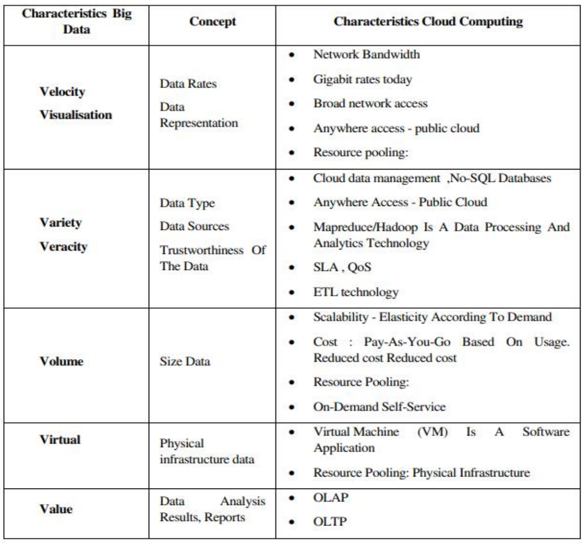


Fig 19. Model Showing The Relationship Between Big Data And Cloud Computing[24]



Characteristics Big Data	Concept	Characteristics Cloud Computing
Velocity Visualisation	Data Rates Data Representation	• Network Bandwidth • Gigabit rates today • Broad network access • Anywhere access - public cloud • Resource pooling:
Variety Veracity	Data Type Data Sources Trustworthiness Of The Data	• Cloud data management ,No-SQL Databases • Anywhere Access - Public Cloud • Mapreduce/Hadoop Is A Data Processing And Analytics Technology • SLA , QoS • ETL technology
Volume	Size Data	• Scalability - Elasticity According To Demand • Cost : Pay-As-You-Go Based On Usage. Reduced cost Reduced cost • Resource Pooling: • On-Demand Self-Service
Virtual	Physical infrastructure data	• Virtual Machine (VM) Is A Software Application • Resource Pooling: Physical Infrastructure
Value	Data Analysis Results, Reports	• OLAP • OLTP

Fig 20. Compatibility between big data and cloud computing in terms of characteristics.[24]

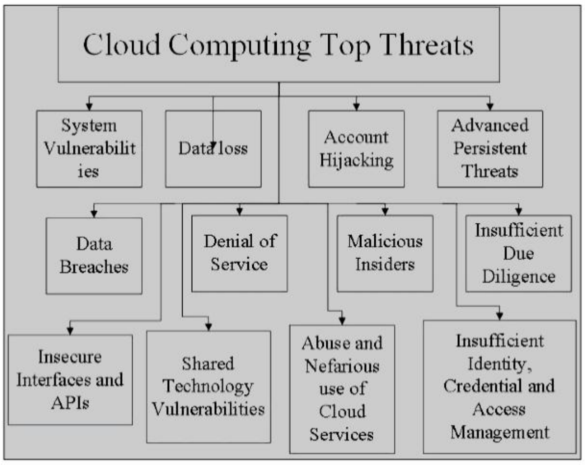

Cloud Computing Top Threats
System Vulnerabilities
Data loss
Account Hijacking
Advanced Persistent Threats
Data Breaches
Denial of Service
Malicious Insiders
Insufficient Due Diligence
Insecure Interfaces and APIs
Shared Technology Vulnerabilities
Abuse and Nefarious use of Cloud Services
Insufficient Identity, Credential and Access Management

Fig 21. Top Cloud Security threat. [26]

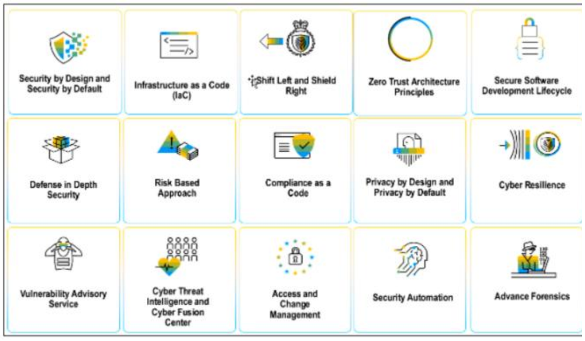

Security by Design and Security by Default
Infrastructure as a Code (IaC)
Shift Left and Shield Right
Zero Trust Architecture Principles
Secure Software Development Lifecycle
Defense in Depth Security
Risk Based Approach
Compliance as a Code
Privacy by Design and Privacy by Default
Cyber Resilience
Vulnerability Advisory Service
Cyber Threat Intelligence and Cyber Fusion Center
Access and Change Management
Security Automation
Advance Forensics

Fig 22. SAP's Comprehensive Strategy for Countering Cyber Threats[26]

This paper offers a novel contribution by addressing the critical security challenges associated with Big Data and cloud computing, particularly focusing on the Hadoop ecosystem. While previous research has explored the potential of Big Data and AI/ML technologies, there is a significant gap in analyzing the specific security risks within distributed cloud environments. We extend existing studies by proposing solutions to enhance data privacy, integrity, and encryption in the context of Big Data processing and cloud storage. This work integrates emerging security practices and frameworks to mitigate vulnerabilities in Hadoop and cloud infrastructures, offering actionable insights for improving data protection and ensuring reliable, scalable performance in modern enterprise system.

II. Methodology

Types of big data Data are produced at unprecedented rates from various sources, such as financial, government, health, and social networks. Such rapid growth of data can be attributed to smart devices, the Internet of Things, etc. In the last decades, companies have failed to store data efficiently and for long periods. This drawback relates to traditional technologies that lack adequate storage capacity and are costly. Meanwhile, big data require new storage methods backed by powerful technologies.

Cloud computing is a transformative technology that enables large-scale and complex computations without the need for maintaining costly hardware, dedicated infrastructure, or specialized software. With the rapid expansion of cloud computing, there has been a significant increase in the volume of data, commonly referred to as big data. Managing and processing big data has become a complex and resource-intensive task, necessitating vast computational power to handle and analyze the data effectively. This study reviews the intersection of big data and cloud computing, exploring their definitions, key characteristics, and classifications [31].

As distributed computing systems continue to evolve, modern Big Data analysis platforms have become increasingly varied in their features and capabilities. This diversity can make it challenging for users, particularly those who are new to Big Data, to make well-informed decisions when first engaging with these platforms. In this paper, we examine the design principles and emerging research directions in contemporary Big Data platforms, drawing on recent advancements in Big Data technologies.

We offer a thorough review and comparison of several leading frameworks, ultimately proposing a typical architecture with five horizontal layers and one vertical layer. Building on this structure, the paper highlights the key components and cutting-edge optimization techniques developed for Big Data, providing guidance for selecting the most appropriate components and architectural design based on specific requirements.[32]

The quick growth of genetic data has become a disruptive force in the Big Data age, affecting almost every business, but the pharmaceutical industry is most affected. Scientists are using this abundance of genetic data to investigate human origins and migratory patterns more and more as nations throughout the world create their own gene banks. Big Data has also been essential in improving cancer research and treatment, giving cancer patients fresh hope.

The research of a variety of illnesses has benefited greatly from big data as well. Notably, compared to more conventional approaches, genetic analysis of disorders has shown greater efficacy in determining available treatments. This review provides an overview of the role of Big Data in medical research, focusing on its applications in the study of tumors, neurological and psychiatric disorders, cardiovascular diseases, and other medical fields. It also highlights the ongoing developments and future directions of Big Data in medicine[33]

The term "Big Data" describes the enormous amount of data with intricate and diverse structures that conventional data management techniques are unable to efficiently handle. Big data analytics is being utilized by businesses more and more these days, and it is essential in many different fields. This study examines the use of big data in important domains including human resources management (HRM), business process management (BPM), and telecommunications [34]

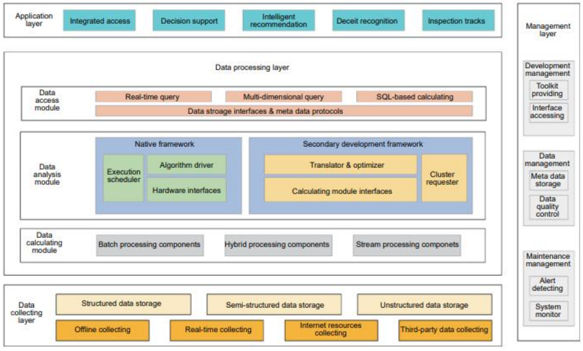


Fig 23. Five-horizontal-and-one-vertical structure[34]

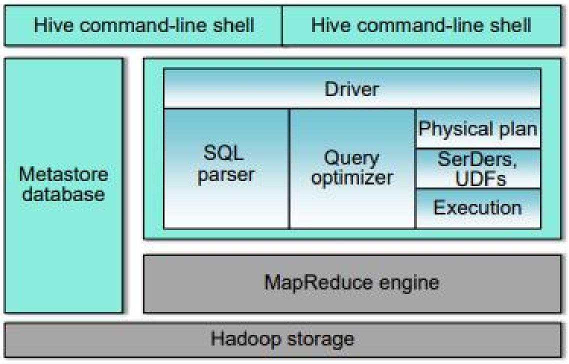


Fig 24. Architecture of Hive system.[34]

There are several categories into which big data may be divided. The large data categorization is shown in Figure 2. The definitions of the different kinds of big data are compiled in Table 4. 3. Big Data with Machine Learning Finding knowledge and making wise selections are the primary purposes of machine learning techniques. Machine learning is used in various realworld applications, such as data mining, recognition systems, recommendation engines, and autonomous control systems. The machine learning domain can be divided into three areas, namely, supervised learning, unsupervised learning, and reinforcement learning [35].

The research of a variety of illnesses has benefited greatly from big data as well. Notably, compared to more conventional approaches, genetic analysis of disorders has shown greater efficacy in determining available treatments. With an emphasis on its applications in the study of cancers, neurological and psychiatric illnesses, cardiovascular diseases, and other medical domains, this article gives a broad overview of the importance of big data in medical research. It also emphasizes the current advancements and potential paths of big data in medicine. This paper provides a comprehensive overview of different approaches to pattern mining in the context of Big Data.

This paper examines pattern mining techniques like Apache Hadoop, Apache Spark, and parallel and distributed processing, focusing on four main types: parallel frequent itemset mining, high utility itemset mining, sequential pattern mining, and frequent itemset mining in uncertain datasets. It reviews advancements in parallel, distributed, and scalable pattern mining, identifies challenges in algorithm design, and discusses open research issues and opportunities.

Finding significant associations in datasets is accomplished using the fundamental data mining approach of pattern mining. There are several types of pattern mining, such as sequence mining, high utility itemset mining, and frequent itemset mining. The goal of high utility itemset mining, a new data science endeavor, is to extract information according to domain-specific goals. The "utility" of a pattern refers to its effectiveness or benefit, which is determined by user priorities and domain-specific insights.

Sequential pattern mining (SPM) has been extensively studied and expanded in various directions. It involves identifying sequential patterns within a collection of sequential data. In recent years, there has been increasing interest in frequent pattern mining over uncertain transaction datasets. Furthermore, mining itemsets in Big Data environments, particularly using Apache Hadoop and Apache Spark, has garnered significant attention.[36]

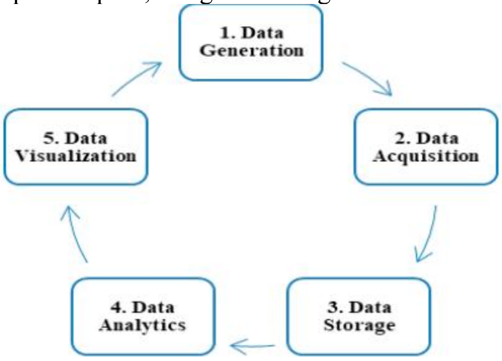


Fig 25. Big Data life cycle[36]

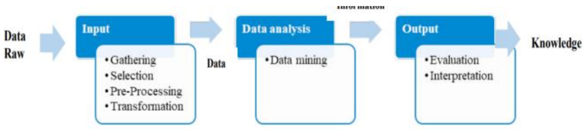


Fig 26. Data analytics process[36]

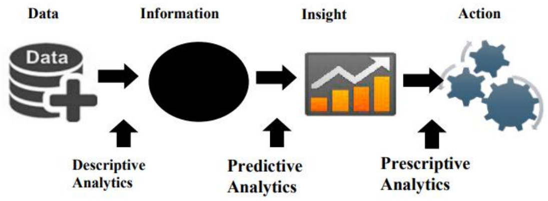


Fig 27. Types of Analytics Techniques[36]

With an emphasis on urban data—data associated with cities and intrinsically related to both place and time—I define large data mainly in terms of its scale. This data, which is mostly broadcast from sensors, represents a substantial change in the type of data that is now accessible about urban settings, specifically with regard to what occurs, where, and when in cities. I argue that this transformation in data collection represents a fundamental change in how we understand and manage urban spaces. The rapid growth of big data is driving a shift from long-term strategic planning to more immediate, short-term thinking about the functioning and management of cities. However, there is the potential for this data, over much longer periods, to provide insights into various time horizons, offering valuable information for both current and future urban development.

Finally, by analyzing six months of smart trip card data from the public transport system in Greater London, I draw attention to the need for fresh ideas and analytical techniques. This data, which monitors individual journeys, is an illustration of how big data may be utilized to better understand urban mobility and guide urban planning choices [37].

Data mining and big data analytics are crucial methods for examining enormous databases and drawing insightful conclusions. However, because of the complexity and enormous volume of big data, traditional methods of analysis and extraction are sometimes useless. Data clustering, a crucial data mining approach, divides data into categories to make information extraction simpler. However, existing clustering algorithms, such as k-means and hierarchical clustering, often fall short in terms of efficiency, as the quality of the clusters they produce is compromised.

This study presents Hybrid Clustering, a novel clustering technique that addresses these constraints by overcoming the drawbacks of current methods. We compare the performance of the proposed hybrid algorithm against traditional clustering algorithms based on key metrics such as precision, recall, F-measure, execution time, and accuracy. Experimental results demonstrate that the hybrid clustering algorithm outperforms existing methods, offering superior accuracy, precision, recall, and F-measure values, making it a more efficient and reliable solution for clustering large-scale data [38].

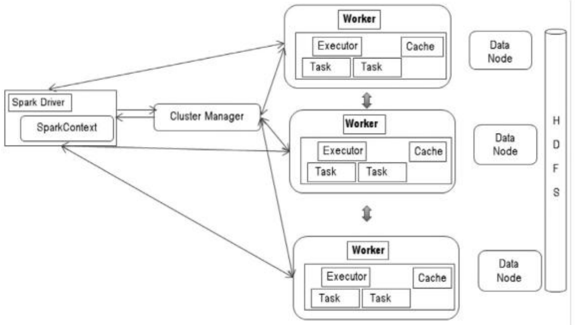


Fig 28. Working model of spark framework[38]

Big data processing and analysis are mostly conducted on shared-nothing computer clusters. Two essential techniques for improving processing performance and scalability in cluster computing are data partitioning and sampling. In the context of big data processing and analysis, this study provides an extensive overview of the strategies and tactics utilized for data sampling and partitioning.

We start by giving a summary of the big data frameworks that are most frequently utilized in Hadoop clusters. Next, we explore various data partitioning techniques, starting with the three classical horizontal partitioning schemes: range partitioning, hash partitioning, and random partitioning. We also discuss data partitioning strategies specific to Hadoop clusters, including the novel Random Sample Partition (RSP) distributed model. After that, the study examines traditional data sampling techniques such reservoir sampling, stratified sampling, and simple random sampling. We look at record-level sampling and block-level sampling, two popular sampling strategies in big data settings. While record-level sampling is less efficient for large distributed datasets, block-level sampling, which operates on data blocks produced by classical partitioning methods, does not always yield representative samples for approximate computing of big data.

Additionally, we summarize existing strategies and related work on sampling-based approximation in Hadoop clusters. The paper emphasizes that data partitioning and sampling should be considered together in order to develop reliable approximate cluster computing frameworks that balance both computational efficiency and statistical accuracy.[39]

Big Data applications are crucial for organizations as they rely on vast amounts of data for insights. Traditional methods struggle with slow responsiveness, scalability, and performance limitations. Advancements in Big Data technologies have addressed these challenges, with new distribution models and technologies emerging to handle complexity. This paper reviews these developments to guide organizations in selecting the right combination of Big Data technologies based on their specific needs and application requirements. It provides an overview of key technologies, their features, advantages, limitations, and typical use cases, assisting in informed decision-making for Big Data adoption and implementation [40].

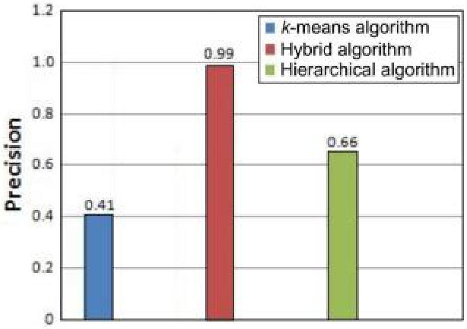


Fig 29. Comparison of three algorithms based on precision.[40]

Big Data mining, based on the "5 V's" (Volume, Variability, Velocity, Variety, and Value), is crucial for managing large, complex datasets. As Big Data evolves, its challenges will become significant areas of research in the coming years.

In order to extend Information and Communication technologies (ICT) applications beyond conventional LAN and WAN contexts to the cloud and the wider Internet, researchers and developers worldwide are utilizing Big Data technologies. An overview of several ICT applications that profit from data mining and big data analytics is given in this article.

The article highlights the existing platforms, languages, and tools available for these goals while examining the vast array of ICT applications that leverage data mining and analytics for Big Data. It also looks at the difficulties encountered in various application areas. The study concludes by providing a brief overview of current developments in Big Data technology research that seek to solve these issues and enhance ICT applications, perhaps providing future answers [41]

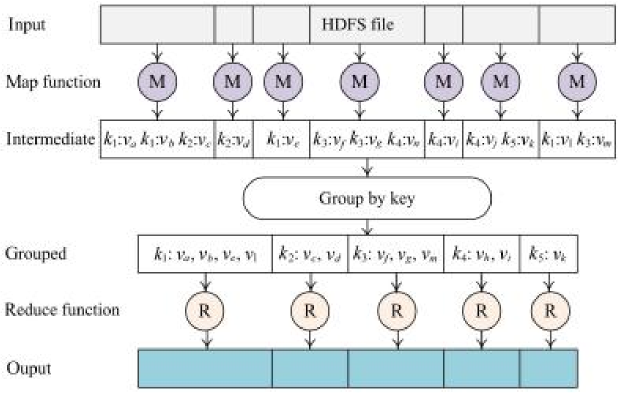
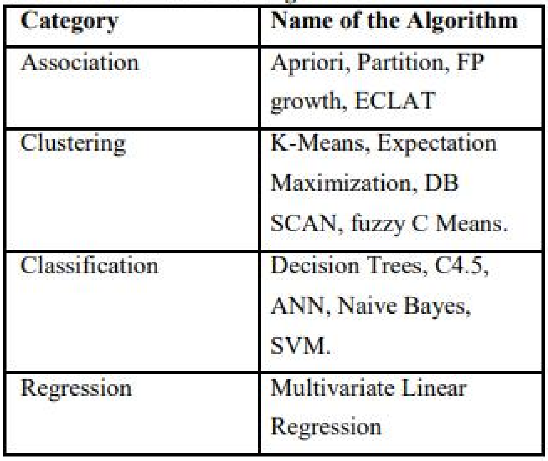


Fig 30. MapReduce model.[41]

The difficulties posed by large data are discussed in this article, with an emphasis on security concerns in the Hadoop environment. It specifically looks at the security issues with the central Hadoop architectural layer, the Hadoop Distributed File System (HDFS). Three methods are investigated to improve HDFS security: Kerberos authentication, encryption algorithms, and NameNode security mechanisms [42].

The goal of data mining is to extract novel insights and patterns from huge databases. Data mining has been researched in many different application areas throughout the years, which has resulted in the creation and use of many different techniques. However, the recent surge in data volume, along with the increased computational and analytical demands, has rendered many traditional data mining methods impractical for handling big data.

In order to fulfill the scalability and performance requirements of large-scale data mining jobs, effective parallel and concurrent algorithms are crucial. The usage of threads, MPI (Message Passing Interface), MapReduce, and mash-up or workflow technologies are some of the parallelization approaches that have been created. Depending on the particular needs of the application, these methods provide different performance and usability features.The MPI model has proven to be effective for computationally intensive problems, particularly in simulation, but it can be difficult to implement in practice. On the other hand, MapReduce, which originated from information retrieval techniques, has become a popular cloud-based technology for processing big data. Over time, several MapReduce frameworks have been developed, with Google's MapReduce being one of the most well-known. Another widely adopted implementation is Hadoop, an open-source MapReduce framework that has gained significant traction among major IT companies, including Yahoo, Facebook, and eBay. This paper specifically focuses on Hadoop and its implementation of MapReduce for analytical processing, exploring its capabilities and the impact it has had on big data analysis.[43]



Category	Name of the Algorithm
Association	Apriori, Partition, FP growth, ECLAT
Clustering	K-Means, Expectation Maximization, DB SCAN, fuzzy C Means.
Classification	Decision Trees, C4.5, ANN, Naive Bayes, SVM.
Regression	Multivariate Linear Regression

Fig 31. Classification of Algorithms[43]

The growing use of big data in cloud computing raises privacy concerns. However, the increasing volume of data also presents challenges, such as the execution time required for data encryption. This paper proposes the Dynamic Data Encryption Strategy (D2ES), which selectively encrypts data based on privacy classification methods while adhering to timing constraints. The strategy aims to maximize privacy protection by encrypting only the most sensitive data within the required execution time, balancing security and performance. Experiments show D2ES effectively enhances privacy without compromising data processing and transmission performance [44].

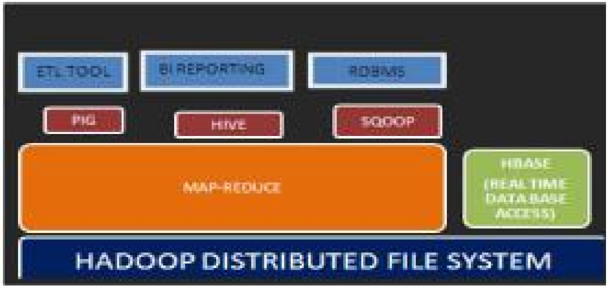

ETL TOOL
BI REPORTING
RDBMS
PIG
HIVE
SQOOP
MAP-REDUCE
HBASE (REAL TIME DATA BASE ACCESS)
HADOOP DISTRIBUTED FILE SYSTEM

Fig 32. Hadoop Architecture [44]

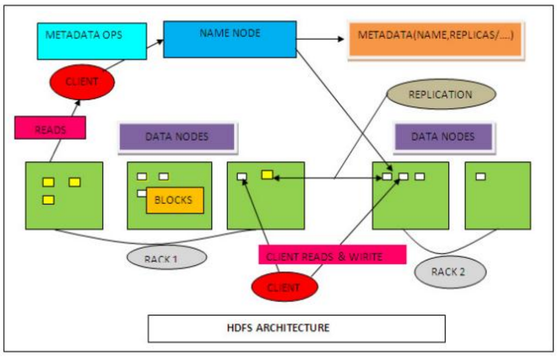


Fig 33. HDFS Architecture[44]

The rapid rise in smart device usage has led to a surge in data generation, posing challenges for processing and managing large-scale datasets. Service providers like Google, Amazon, and Microsoft have deployed geographically distributed data centers to efficiently process massive volumes and ensure quick response times for users using technologies like Hadoop and Spark.However, less attention has been given to the underlying network infrastructure, which is a critical component for the successful implementation of any solution in this environment. Heavy network traffic, particularly data migrations across multiple data centers, can overwhelm the network infrastructure, leading to performance degradation. In extreme cases, the network may fail to transfer data packets from source to destination, negatively impacting service quality.

The authors propose a novel SDN-based big data management approach that optimizes network resource consumption, focusing on bandwidth and data storage units, enhancing the efficiency of big data analytics across multiple cloud data centers. The solution uses Bloom-filter-based insertion and deletion techniques in the flow table managed by the OpenFlow controller, enabling real-time traffic analysis and optimization of network resources for big data applications in multi-cloud environments. This approach improves network resource management, addressing challenges of scaling data transfer across distributed cloud data centers and enhancing big data performance[45].

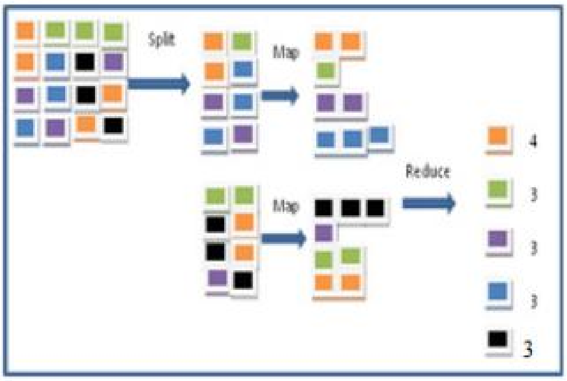


Fig 34. Steps in MapReduce to process the database[45]

The term "big data" is increasingly used in various fields, including everyday life and traditional research. The proliferation of networks has diversified the issues, solutions, and approaches associated with big data. This paper reviews recent research on network big data, including data types, storage models, privacy concerns, data security, analysis methods, and applications, offering insights into current trends and predicting future directions[46].

Data streaming learning Various real-time world technologies, such as stock management, network traffic, and credit card transactions, generate huge datasets. Data mining plays an important role in finding interesting patterns, 5 Research Issues in Big Data As data are growing at exponential rates, a number of issues and problems emerge during the processing and storage of big data. Few tools are available to resolve these issues and problems in a cloud environment. Technologies, such as PigLatin, Dryad, MongoDB, Cassandra, and MapR, are not able to resolve these issues in big data processing. Even with the help of Hadoop and MapR, users cannot execute queries on databases, and they have low-level infrastructures for data processing and management. Some issues and problems in big data are summarized as follows[47]: (1) Distributed database storage system: Numerous technologies are used to store and retrieve huge amounts of data. Cloud computing is an important aspect of big data. Big data are generated by multiple devices on a daily basis. At present, the main issue in distributed frameworks is the storage of data in a straightforward manner and the processing and migration of data between distributed servers. (2) Data security: Security threats are an important issue in a cloud computing environment. Cloud computing has been transformed with modern information and communication technologies, and several types of unresolved security threats exist in big data. Data security threats are magnified by the variety, velocity, and volume of big data. Meanwhile, various issues and threats, such as the availability of data, confidentiality, real-time monitoring, identity and access authorization control, integrity, and privacy, exist in big data when used with cloud computing frameworks. Therefore, data security must be measured once data are outsourced to cloud service providers [48].

The increasing amount of data in computer technologies presents challenges for users. Cloud computing services offer a powerful environment for storing large volumes of data, eliminating the need for dedicated space and expensive hardware and software maintenance. However, handling big data requires large computational clusters. This work discusses the definition, classification, and characteristics of big data, compares cloud-based big data frameworks, and addresses research challenges in distributed database storage, data security, heterogeneity, and data visualization [49].

Heterogeneity: Big data are heterogeneous in nature because data are gathered from multiple devices in different formats, such as images, videos, audio, and text. Before loading data into a warehouse, they need to be transformed and cleaned, and the processes present challenges in big data [49]. Combining all unstructured data and reconciling them for use in report creation are incredibly difficult to achieve in real-time. (4) Data processing and cleaning: Data storage and acquisition require preprocessing and cleaning, which involves data merging, data filtering, data consistency, and data optimization. Thus, processing and cleaning data are difficult because of the wide variety of data sources[50]. Moreover, data sources may contain noise and errors, or they may be incomplete. The challenge is how to clean large amounts of data and how to determine whether such data are reliable. (5) Data visualization: Data visualization is a technique to represent complex data in a graphical form for clear understanding. If the data are structured, then they can be easily represented in the traditional graphical way. If the data are unstructured or semistructured, then they are difficult to visualize with high diversity in realtime.

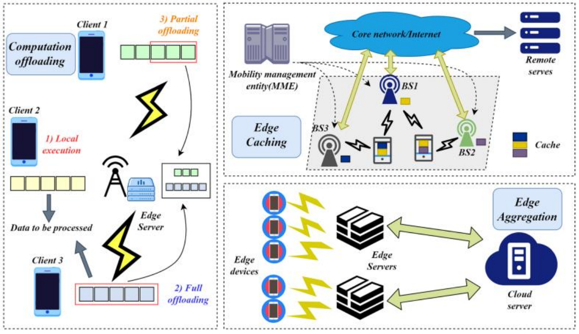


Fig 35. The frameworks of computation ofoading, edge caching and edge aggregation[50]

Very Large (VL) data, or big data, refers to datasets that are too large to fit into a computer's working memory. While this definition is not entirely objective, it is practical and easy to understand, as it captures the essence of data that exceeds the capacity of typical computational resources. Clustering is a key task in pattern recognition and data mining, used to explore VL databases (including VL images). Therefore, clustering algorithms that scale well to VL data are both important and highly useful.

This paper compares three approaches to extend fuzzy c-means (FCM) clustering to VL data. The methods include sampling followed by non-iterative extension, incremental techniques, and kernelized versions of FCM. Numerical experiments were conducted on both loadable and VL datasets to compare time and space complexity, speed, approximation quality, and alignment with ground truth. Results showed random sampling combined with extension FCM, bit-reduced FCM, and approximate kernel FCM are effective for VL data approximation [51]. A powerful approach that allows consumers and businesses to obtain services on-demand based on their unique needs is cloud computing. Storage, deployment platforms, and simple access to web services are just a few of the many things it provides. Load balancing (LB), which is essential for preserving optimal application performance while satisfying Quality of Service (QoS) standards and abiding by Service Level Agreements (SLAs) issued by cloud providers, is one of the primary difficulties in cloud computing.

Since cloud providers frequently struggle to divide workloads equally among servers, it is crucial to put in place an effective load balancing strategy that optimizes resource usage and guarantees high customer satisfaction. This study offers a thorough analysis of several load balancing strategies for cloud settings, emphasizing static, dynamic, and nature-inspired methods to tackle issues like system performance and data center response time. The review provides an analytical comparison of these algorithms, identifies gaps in current research, and suggests directions for future studies. Additionally, the paper includes graphical representations of the reviewed algorithms to illustrate their operational flow. Furthermore, it discusses fault-tolerant frameworks and explores other relevant frameworks in recent literature, offering valuable insights for further advancements in the field of cloud load balancing.[52]

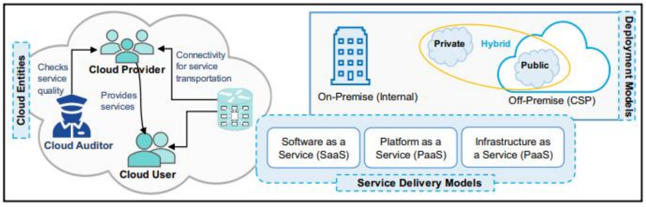


Fig 36. Overview of Cloud Computing[52]

The development of public cloud infrastructure relies heavily on effective data protection, which ensures that the infrastructure is properly implemented and secure, benefiting both cloud users and providers. However, data protection has not always received the necessary attention in this context. This paper addresses the potential risks that data may face during transfer and recovery in the cloud. We explore various known attacks that can compromise data security and, in turn, examine the advantages and disadvantages of the different techniques proposed in the literature to mitigate these risks [53].

Concerns about privacy and security have grown in importance as cloud computing develops, causing both industry and academics to take notice. By looking at several privacy protection methods, this study examines the advancements achieved in tackling privacy security issues in cloud computing.

We start by outlining the privacy threats connected to cloud computing and offer a thorough methodology for safeguarding privacy. Access control, ciphertext-policy attribute-based encryption (CP-ABE), key-policy attribute-based encryption (KP-ABE), trace mechanisms, fine-grain multi-authority revocation mechanisms, proxy re-encryption (PRE), hierarchical encryption, searchable encryption (SE), and multi-tenant trust models are some of the important technologies that we then go over in order to mitigate these risks. We examine and contrast the features and areas of use of common schemes for every technology.

Finally, we highlight the current challenges in the field and suggest [54].

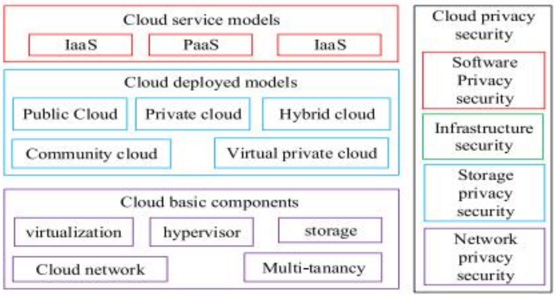


Fig 37. Cloud computing framework[54].

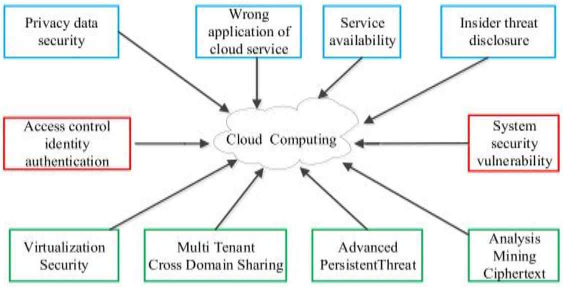

Privacy data security
Wrong application of cloud service
Service availability
Insider threat disclosure
Access control identity authentication
Cloud Computing
System security vulnerability
Virtualization Security
Multi Tenant Cross Domain Sharing
Advanced PersistentThreat
Analysis Mining Ciphertext

Fig 38. Privacy security risk in cloud computing.[54]

Since data creation has increased significantly in recent years, software solutions made to support data-intensive applications have become widely used. Although many businesses use data-driven procedures, it can still be difficult to completely adopt a data-centric worldview. It takes a lot of time and effort to set up a production-ready and efficient deployment for Big Data applications. In order to simplify deployment setup for Big Data applications, this scenario necessitates the use of creative models and methodologies. Furthermore, a lot of businesses are using cloud-deployed clusters as an affordable substitute for on-premises deployments. Accurately forecasting the execution time of Big Data applications is crucial for managing cloud utilization since it enhances design-time choices, lowers the costs associated with cloud over-allocation, and aids in budget management. In this paper, we propose analytical models based on Stochastic Activity Networks (SANs) to model the execution of popular Big Data frameworks such as Hadoop, Tez, and Spark. These models account for both the applications themselves and the underlying cluster infrastructure, providing accurate execution time predictions. We evaluate the accuracy of our proposed SAN models using the TPC-DS industry benchmark across different configurations. With an average prediction error of just 4.5% for MapReduce (MR), 5.8% for Tez, and 2.7% for Spark applications, our numerical analysis outperforms current approaches in terms of accuracy. Furthermore, compared to previously described methodologies, the time needed to solve the suggested models is substantially shorter, indicating both efficiency and increased forecast accuracy [55]. Genomic sequence data has increased at a never-before-seen pace since the Human Genome Project was completed at the turn of the century. Future medical advancements will therefore be more and more reliant on our capacity to handle and evaluate massive genetic data sets, which are only getting bigger as sequencing costs come down. This research examines how big data and cloud computing technologies may be used to manage the enormous amounts of data found in biology. In particular, we concentrate on big data technologies like the Apache Hadoop project, which makes it possible to analyze data at the petabyte (PB) scale in parallel and distributed. We also look at how Hadoop is currently being used in the bioinformatics field and how it is helping to advance genomic analysis and research[56]

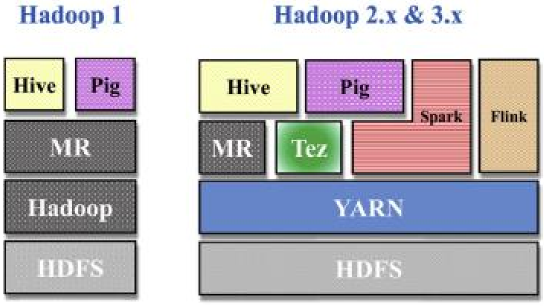


Fig 39. Hadoop evolution[55].

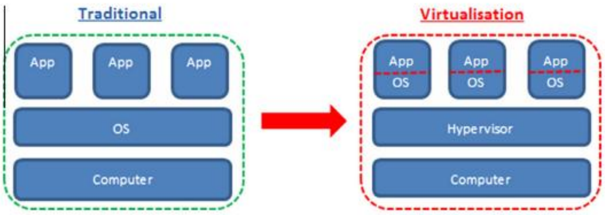


Fig 40. Conceptualisation of the transition from traditional computing[56]

The deep sea is a dynamic environment where benthic ecosystems interact with the water column and surface layers. In the aphotic zones, organisms rely on external cues to regulate their biological clocks, responding to cyclic hydrodynamic patterns and fluctuations in variables like temperature, salinity, phytopigments, and oxygen concentration. Diel Vertical Migration (DVM) is a key component of this synchronization, influencing benthic species' behavioral rhythms and providing essential food for deep-sea communities. Bioluminescent species in migrating deep scattering layers may play a crucial role in benthopelagic coupling, highlighting the need for enhanced methods of detecting and quantifying bioluminescence. Integrating bioluminescence studies into long-term monitoring programs using deep-sea neutrino telescopes could enhance our ability to track carbon and nutrient transfer from the upper ocean into the deep-sea Benthic Boundary Layer (BBL), a critical component of the ocean's biological pump and vital for understanding climate change impacts [57].

A study in Utah found that the air-fuel equivalence ratio significantly predicts emissions from natural gas-fueled pumpjack engines. Higher ratios resulted in lower nitrogen oxide emissions but higher organic compound emissions. For engines with higher ratios, 57% of fuel gas passed through uncombusted, with a median NO_x emissions of 3 ppm. Conversely, engines with lower ratios showed less fuel slip, higher NO_x emissions, and more reactive organic compounds.On average, NO_x emissions from the engines in this study were only 9% of the levels reported in a regulatory oil and gas emissions inventory for natural gas-fueled pumpjack engines. However, volatile organic compound (VOC) emissions in our study were 15 times higher than those in the inventory. We hypothesize that these discrepancies arise from variations in emissions as engines operate at lower loads and age under field conditions. This research has important implications for improving emissions inventories and the effectiveness of related regulatory measures. Additionally, our findings will enhance the ability of photochemical models to better simulate the atmospheric impacts of oil and gas development, contributing to more accurate assessments of environmental impacts.[58]

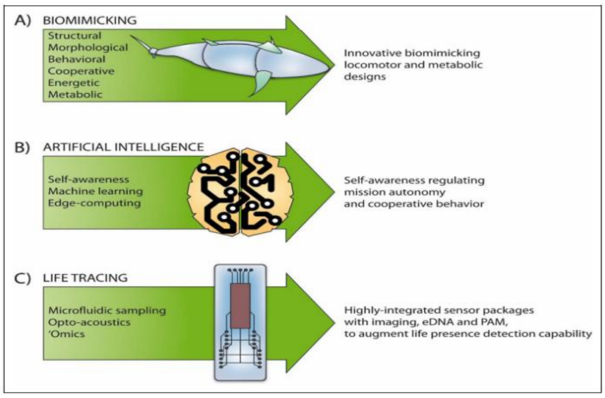

A) BIOMIMICKING
Structural
Morphological
Behavioral
Cooperative
Energetic
Metabolic
Innovative biomimicking locomotor and metabolic designs
B) ARTIFICIAL INTELLIGENCE
Self-awareness
Machine learning
Edge-computing
Self-awareness regulating mission autonomy and cooperative behavior
C) LIFE TRACING
Microfluidic sampling
Opto-acoustics
'Omics
Highly-integrated sensor packages with imaging, eDNA and PAM, to augment life presence detection capability

Fig 41. Major lines of action within each field for research development.[58]

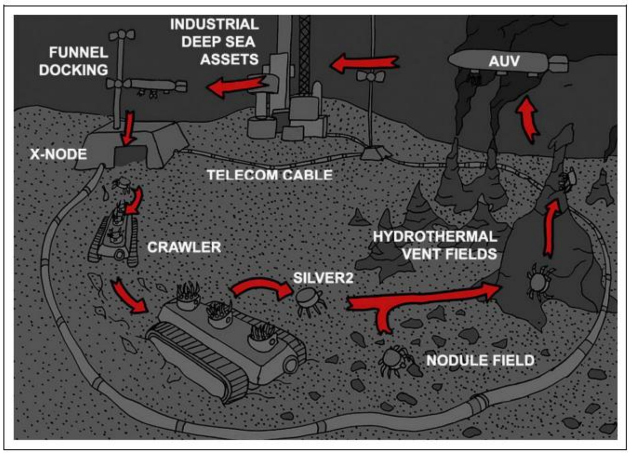


Fig 42. Innovative biomimicking applications in a scenario of extreme environment exploration[58]

Climate science is rapidly becoming a Big Data domain, experiencing unprecedented growth. To address the challenges posed by this data explosion, we are advancing the concept of Climate Analytics-as-a-Service (CAaaS). Our focus on analytics stems from the understanding that the true value of Big Data in climate science lies in the knowledge gained from analyzing it, which ultimately leads to societal benefits.

We advocate for CAaaS because it offers a structured approach to tackling these challenges, akin to the concept of business process-as-a-service—an evolving extension of Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS), and Software-as-a-Service (SaaS), all enabled by Cloud Computing. While Cloud Computing plays a critical role, we see it as just one component in a broader ecosystem of capabilities necessary for delivering climate analytics as a service. Collectively, these capabilities promote generativity—a capacity for self-assembly—that is key to solving many Big Data challenges in climate science.

A prime example of cloud-enabled CAaaS is MERRA Analytic Services (MERRA/AS), which is built on this generative principle. MERRA/AS enables MapReduce analytics over NASA's Modern-Era Retrospective Analysis for Research and Applications (MERRA) dataset. MERRA integrates observational data with numerical models to create a globally consistent synthesis of 26 key climate variables, which is critical for climate change research and decision support applications.

MERRA/AS demonstrates the full potential of CAaaS by integrating several essential generative elements:

1. High-performance, data-proximal analytics
2. Scalable data management
3. Software appliance virtualization
4. Adaptive analytics
5. Domain-harmonized API

These elements work together to provide end-to-end CAaaS capabilities, enabling more efficient and scalable climate analytics.

The effectiveness of MERRA/AS has been proven through various applications. In our experience, Cloud Computing lowers barriers to organizational change, reduces risks, fosters innovation, and supports technology transfer. It also provides the agility necessary to meet the evolving needs of climate science. By offering a new tier in the data services stack, Cloud Computing connects enterprise-level data and computational resources with new customers and mobility-driven applications.

For climate science, the greatest value of Cloud Computing lies in its ability to engage communities in building new capabilities, making it a crucial link between Big Data and the future of climate science research.[59]

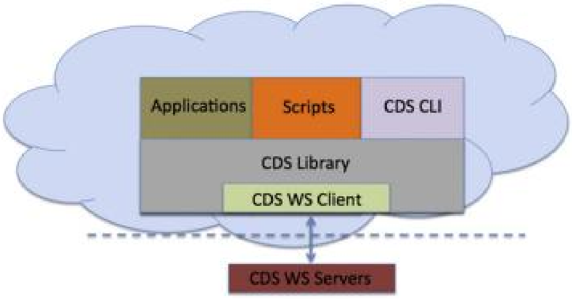


Fig 43. Climate Data Services client stack built on the capabilities enabled by the CDS Reference Model, Web ServicesClient, Library, and API.[59]

In the era of big data, vast amounts of data are generated and collected rapidly from a wide range of sources. Embedded within this massive volume of data is valuable information and knowledge, especially in fields such as healthcare and epidemiology. For instance, data related to patients affected by viral diseases like COVID-19 can provide critical insights. By applying data science techniques to analyze such epidemiological data, researchers, epidemiologists, and policymakers can gain a deeper understanding of the disease, which can inform strategies for detection, control, and prevention.

This paper presents a data science solution designed to analyze large-scale COVID-19 epidemiological data. The solution provides users with a clearer understanding of key information, such as the number of confirmed COVID-19 cases. Evaluation results highlight the effectiveness of our solution in uncovering valuable knowledge from big COVID-19 data, demonstrating its potential to aid in the fight against the pandemic.[60]

On the other hand, Big Data also brings significant challenges, such as difficulties in capturing, storing, analyzing, and visualizing vast amounts of data. This paper provides an in-depth exploration of Big Data, examining its applications, opportunities, and challenges. It also discusses the state-of-the-art techniques and technologies currently employed to address Big Data issues. Furthermore, we explore several emerging methodologies for managing the data deluge, including granular computing, cloud computing, bio-inspired computing, and quantum computing, all of which are key to advancing our ability to cope with the challenges of Big Data.[61]

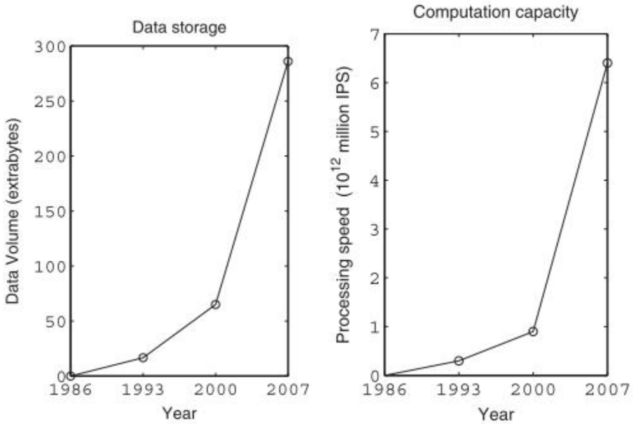


Fig 44. Data deluge: the increase of data size has surpassed the capabilities of computation[61]

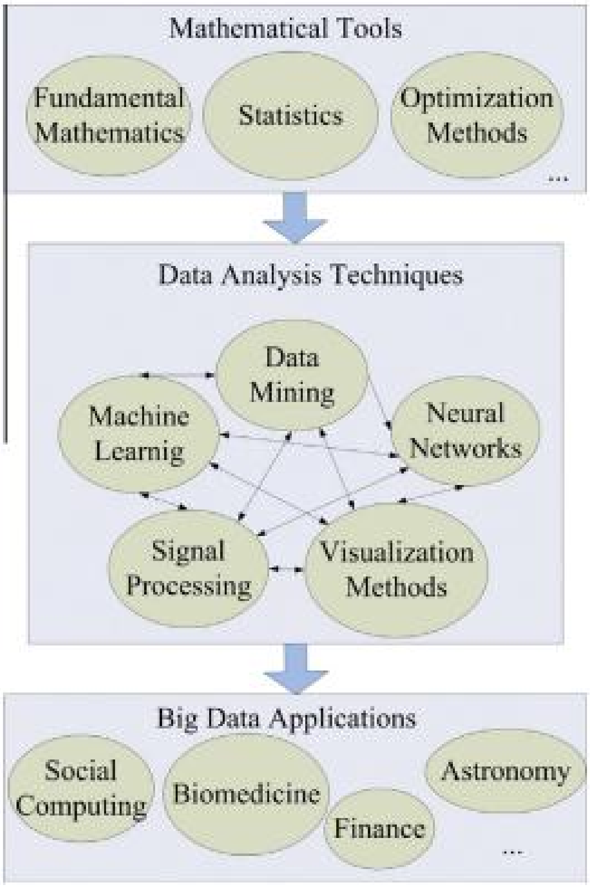


Fig 45. Big Data techniques.[61]

In today's world, medical technology has become central to societal development. With the rapid growth of biomedical data, the medical field is encountering various challenges, such as the overwhelming volume of data and the intensive computational demands required to process it. As a result, large-scale data processing in biomedicine has garnered significant attention. This approach has become a primary method for analyzing various biological and biomedical experiments, including next-generation sequencing and mass spectrometry analysis.

Given its potential, the development of big data processing in biomedicine is moving in a positive direction. A key enabler of this progress is cloud computing, which has become an essential tool in handling the vast amounts of biomedical data. This paper explores the role of cloud computing in biomedical data processing and offers practical recommendations for leveraging cloud technologies to address current challenges in the field.[62]

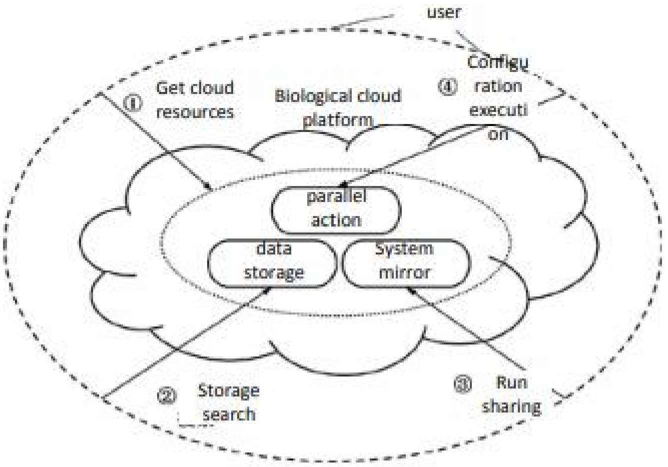
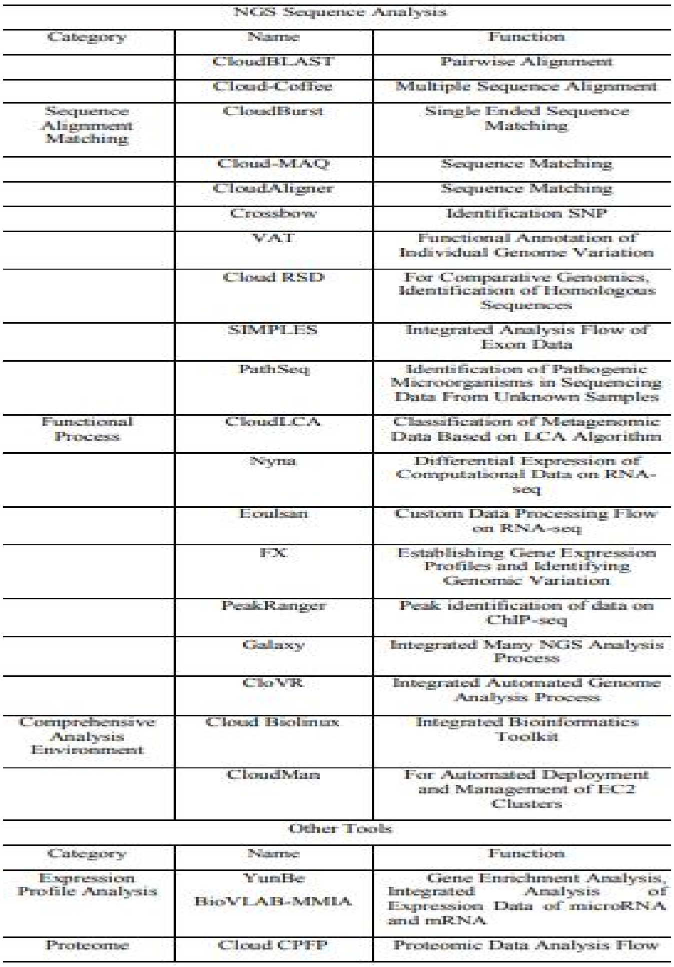


Fig 46. basic models for biological cloud solutions[62]



NGS Sequence Analysis

Category	Name	Function
	CloudBLAST	Pairwise Alignment
	Cloud-Coffee	Multiple Sequence Alignment
Sequence Alignment Matching	CloudBurst	Single Ended Sequence Matching
	Cloud-MAQ	Sequence Matching
	CloudAligner	Sequence Matching
	Crossbow	Identification SNP
	VAT	Functional Annotation of Individual Genome Variation
	Cloud RSD	For Comparative Genomics, Identification of Homologous Sequences
	SIMPLES	Integrated Analysis Flow of Exon Data
	PathSeq	Identification of Pathogenic Microorganisms in Sequencing Data From Unknown Samples
Functional Process	CloudLCA	Classification of Metagenomic Data Based on LCA Algorithm
	Nyna	Differential Expression of Computational Data on RNA-seq
	Eoulsan	Custom Data Processing Flow on RNA-seq
	FX	Establishing Gene Expression Profiles and Identifying Genomic Variation
	PeakRanger	Peak identification of data on ChIP-seq
	Galaxy	Integrated Many NGS Analysis Process
	CloVR	Integrated Automated Genome Analysis Process
Comprehensive Analysis Environment	Cloud Biolinux	Integrated Bioinformatics Toolkit
	CloudMan	For Automated Deployment and Management of EC2 Clusters

Other Tools

Category	Name	Function
Expression Profile Analysis	YunBe BioVLAB-MMIA	Gene Enrichment Analysis, Integrated Analysis of Expression Data of microRNA and mRNA
Proteome	Cloud CPFP	Proteomic Data Analysis Flow

Fig 47. Biomedical Applications in Cloud Computing[62]

Big Data encompasses a range of hardware and software technologies with heterogeneous infrastructures, and the Hadoop framework plays a key role in storing and processing this vast volume of data. It offers a fast, cost-effective solution for managing Big Data and is widely used across various sectors, including healthcare, insurance, and social media. Hadoop is an open-source, distributed computing framework designed for storing and processing data on a cluster of commodity hardware.

Despite the flexibility and scalability offered by Hadoop, the framework also introduces certain vulnerabilities, which can become potential threats to data security. These vulnerabilities expose the system to attacks that can compromise data integrity and performance. This paper discusses various types of vulnerabilities within the Hadoop ecosystem and presents potential solutions to mitigate or eliminate these risks.

The paper also details an experimental setup where common attacks are performed to demonstrate the impact of these vulnerabilities and test the effectiveness of proposed security solutions. The results of these experiments highlight the adverse effects of attacks on system performance, emphasizing the need for a robust security strategy, such as defense-in-depth, to safeguard data in Hadoop environments.[63]

Big data refers to the massive volume of data that requires new technologies and architectures to effectively capture, store, and analyze in order to extract meaningful insights. The sheer size and complexity of big data make it difficult to apply traditional analysis techniques, which were not designed to handle such vast amounts of information. The unique characteristics of big data—such as volume, velocity, variety, variability, value, and complexity—present significant challenges in its processing and analysis.

As big data technology continues to emerge, it offers substantial benefits to business organizations. However, adapting to this technology also brings its own set of challenges. This paper introduces the concept of big data, emphasizing its importance in the modern world, and explores existing projects that are transforming scientific research and societal practices through the use of big data. The paper further discusses the various challenges and issues associated with adopting big data technologies, particularly the Hadoop ecosystem, and highlights the problems that Hadoop currently faces. The paper concludes by outlining best practices for effectively utilizing big data and addressing the challenges in its adoption.[64]

The Industrial Internet of Things (IIoT) is poised to play a crucial role in enabling Industry 4.0. However, achieving reliable and low-latency communication in networked control automation remains a significant challenge, as industrial networks often require strict timing to ensure proper responses. Modern wireless protocols for industrial networks, such as IEEE 802.15.4-2015 with time-slotted channel hopping (TSCH), rely on a precisely scheduled transmission plan to prevent collisions and ensure deterministic end-to-end traffic.

Despite its benefits, guaranteeing bounded end-to-end latency in TSCH is challenging, as transmissions must be temporally coordinated. This becomes even more complex when link quality degrades, potentially requiring a full reconstruction of the TSCH schedule along the communication path.

To address these challenges, we propose a Low-Latency Distributed Scheduling Function (LDSF). This approach organizes the TSCH slotframe into smaller segments called "blocks." Each transmitter selects the appropriate blocks based on its hop distance from the border router, ensuring that retransmission opportunities are automatically scheduled. Additionally, to conserve energy, nodes can power down their radios as soon as their packet is successfully acknowledged.

Mathematical analysis and simulation evaluations demonstrate the effectiveness of the proposed LDSF algorithm. When compared to three state-of-the-art scheduling functions (SFs)—1) Minimal SF (MSF), 2) Low-Latency SF (LLSF), and 3) Stratum—LDSF shows superior performance in terms of latency and efficiency.[65]

The Internet of Things (IoT) is revolutionizing everyday life by enabling new services that enhance convenience and efficiency. This technology, in turn, integrates with other emerging technologies like Big Data, Cloud Computing, and Monitoring, offering a powerful framework for data-driven solutions. In this work, we explore the synergies between these four technologies to identify common operations and combine their functionalities in ways that create beneficial use cases.

While the broader concept of smart cities is widely discussed, our focus is on developing innovative systems for collecting and managing sensor data in a smart building operating within an IoT environment. We propose using a cloud server as the core technology for the sensor management system. This server would collect data from various sensors installed throughout the building, enabling centralized management and control.

Through the IoT-enabled network, users can remotely monitor and control the system via mobile devices, making it easy to access and manage data from anywhere. The integration of these technologies in a smart building can lead to enhanced energy efficiency and sustainability, ultimately contributing to the creation of a Green Smart Building.[66]

This work proposes an innovative infrastructure for secure management of big data (BD) in smart buildings (SBs) within a wireless-mobile 6G network. As the telecommunications field continues to rapidly evolve, new challenges and opportunities arise. The sixth-generation (6G) wireless network not only builds on the benefits of its predecessors but also addresses some of the limitations encountered in earlier versions. Moreover, technologies related to telecommunications, such as the Internet of Things (IoT), cloud computing (CC), and edge computing (EC), can seamlessly operate within the 6G network, further enhancing its capabilities.

Building on these advancements, we propose a scenario that integrates IoT, CC, EC, and BD to create a smart and secure environment in smart buildings. The primary goal of this work is to develop a novel secure cache decision system (CDS) within a wireless 6G network for SBs. This system aims to provide users with a safer and more efficient environment for browsing the Internet, as well as for sharing and managing large-scale data in the fog computing layer.

The proposed CDS consists of two types of servers: a cloud server and an edge server. To support this proposal, we examine and compare existing cache decision systems, outlining their strengths and weaknesses. Our work aims to improve upon these systems by offering a secure, efficient, and scalable solution for managing big data in the context of 6G-enabled smart buildings.[67]

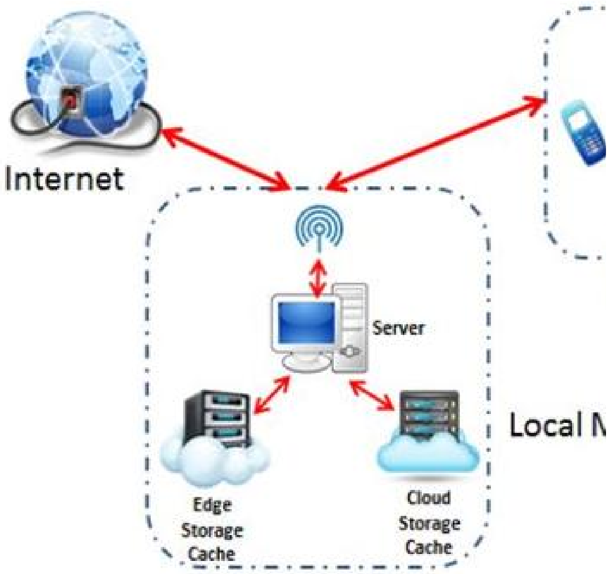


Fig 48. Proposed System Architecture.[67]

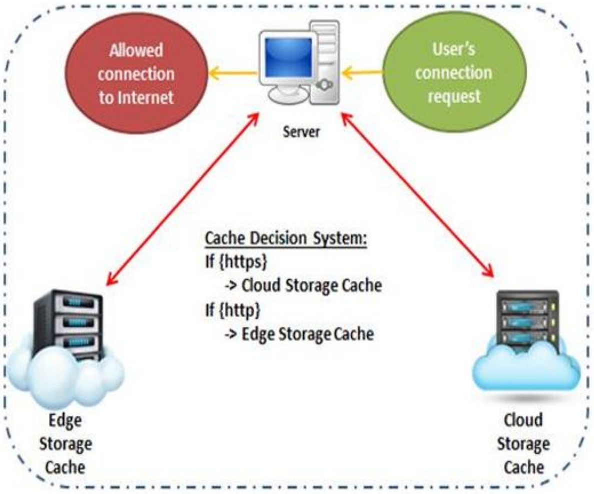


Fig 49. Proposed Cache Decision System (CDS).[67]

Energy-saving (ES) systems based on the Internet of Things (IoT) play a crucial role in enhancing smart homes by automating the understanding of human behavior and activity recognition. However, traditional approaches often struggle to capture the relationships between users' contexts and the energy-saving actions of appliances. These methods also face challenges in handling large volumes of metering data and time-varying user context datasets. Additionally, privacy concerns—both from residents and utility providers—are a significant issue when dealing with sensitive data.

To address these challenges, we propose a privacy-preserving, residential context-aware online energy-saving system (PRCOES) for IoT-enabled smart homes. In this system, we model the repeated interaction between energy-saving actions of appliances and user activity recognition as a contextual multi-armed bandit (CMAB) problem. Using this approach, a context-aware online learning algorithm can predict the most appropriate energy offers (EOs) that meet user satisfaction, task completion rates, and energy-saving goals for appliances.

Our system employs a tree-based structure that expands from top to bottom to recommend EOs, which is scalable and capable of handling large metering datasets while maintaining user context-awareness. Theoretical analysis demonstrates that our approach achieves sublinear regret and ensures differential privacy for both residents and utility providers.

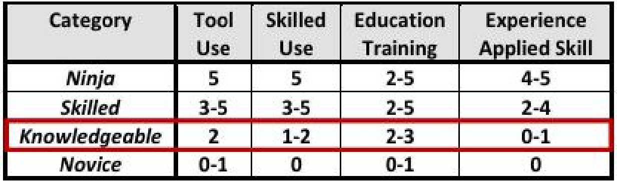
Experimental results validate that PRCOES not only enhances the user experience by improving energy savings but also promotes sustained engagement with the system. Furthermore, it ensures privacy protection for both residents and utility providers, addressing a key concern in the deployment of smart home energy systems.[68]

Protecting software from malicious reverse engineering remains a significant challenge for commercial software companies that invest heavily in developing their products. To safeguard their investments against attacks such as illegal copying, tampering, and reverse engineering, many companies rely on protection software, known as obfuscators, to create variants of their products that are more resilient to adversarial analysis.

In this paper, we assess the effectiveness of different commercial obfuscators against traditional man at the end (MATE) attacks, where an adversary, acting as a legitimate end-user, employs tools such as debuggers, disassemblers, and decompilers to analyze binary executables. Our case study involves four benchmark programs, each with specific adversarial goals classified into comprehension or change tasks.

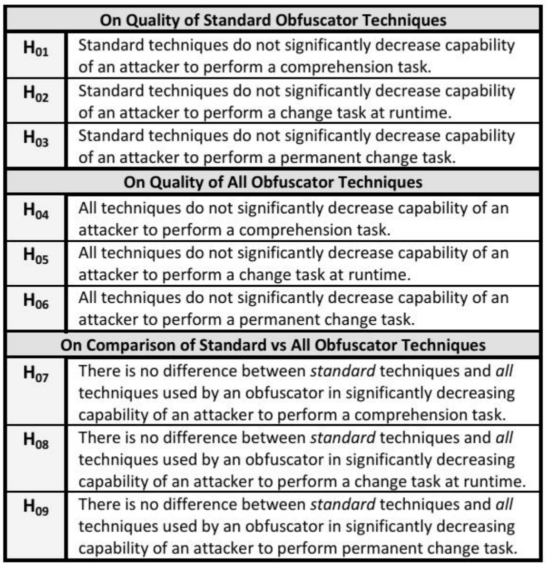
We use both static and dynamic analysis techniques to evaluate the adversarial workload and outcomes before and after each program is obfuscated with three different commercial obfuscators. Our findings confirm the commonly held assumption: an adversary with a reasonable background in computing can easily comprehend and modify completely unprotected programs using standard tools. While obfuscated programs are more resilient, they can still be probed by an adversary to leak certain information. However, none of the protected programs could be successfully altered and saved to create a cracked version.

A key contribution of this study is our unique methodology for evaluating obfuscation effectiveness. Unlike prior research, we categorize adversarial skills and divide program goals into comprehension and change abilities. Additionally, we consider the load time and performance overhead of obfuscated variants, providing a more comprehensive assessment of obfuscation techniques.[69] Software is essential in modern-day life, and software companies often embed secrets into their programs to manage licensing and protect their intellectual property (IP). These secrets, typically in the form of passwords, PINs, or activation codes, are evaluated through point functions, where only the correct input will allow an end-user to legally install or use a product. However, Man at the End (MATE) attacks pose a serious threat to the security of such systems. In these attacks, legitimate software owners, who have full access to the software and its execution environment, can use static and dynamic analysis tools to retrieve or alter sensitive data directly from the program's executables.



Category	Tool Use	Skilled Use	Education Training	Experience Applied Skill
Ninja	5	5	2-5	4-5
Skilled	3-5	3-5	2-5	2-4
Knowledgeable	2	1-2	2-3	0-1
Novice	0-1	0	0-1	0

Fig 50. Proposed Categories of Malicious Reverse Engineers[69]



On Quality of Standard Obfuscator Techniques	
H_{01}	Standard techniques do not significantly decrease capability of an attacker to perform a comprehension task.
H_{02}	Standard techniques do not significantly decrease capability of an attacker to perform a change task at runtime.
H_{03}	Standard techniques do not significantly decrease capability of an attacker to perform a permanent change task.
On Quality of All Obfuscator Techniques	
H_{04}	All techniques do not significantly decrease capability of an attacker to perform a comprehension task.
H_{05}	All techniques do not significantly decrease capability of an attacker to perform a change task at runtime.
H_{06}	All techniques do not significantly decrease capability of an attacker to perform a permanent change task.
On Comparison of Standard vs All Obfuscator Techniques	
H_{07}	There is no difference between *standard* techniques and *all* techniques used by an obfuscator in significantly decreasing capability of an attacker to perform a comprehension task.
H_{08}	There is no difference between *standard* techniques and *all* techniques used by an obfuscator in significantly decreasing capability of an attacker to perform a change task at runtime.
H_{09}	There is no difference between *standard* techniques and *all* techniques used by an obfuscator in significantly decreasing capability of an attacker to perform permanent change task.

Fig 51. Formulated Null Hypotheses[69]

While software companies use legal measures to deter theft, many also rely on technical solutions such as software protection techniques to safeguard their IP against MATE attacks. This paper presents a novel approach to software protection that involves virtualizing software into an alternate, semantically equivalent form. Unlike traditional virtualization, which acts as an interpreter that converts one instruction set into another, our approach focuses on virtualizing instructions into a hardware or circuit representation.

Although the concept of secure multi-party computation also involves circuit-based code realization, we are the first, to our knowledge, to apply this idea in the context of software protection. Our approach builds upon the fundamental principle that "every program can be converted to a circuit and every circuit can be represented as a program." This leads to the development of a technique we call Software based Hardware Abstraction (SBHA).

The contributions of this work are as follows:

1. Realization of SBHA: We propose a novel, low-overhead obfuscation technique that transforms point functions in C programs into hardware abstractions.
2. Resilience and Effectiveness: We demonstrate SBHA's potency against well-known dynamic symbolic analysis tools (such as Klee and Angr), its resilience to compiler and circuit-based deobfuscation methods, and its stealth (when combined with supporting obfuscations), all while maintaining minimal cost.
3. Efficiency: Our approach introduces a new form of software virtualization, with only 3x code overhead significantly more efficient than traditional virtualization transformations, while completely thwarting symbolic execution-based attackers.
4. Unification of Research Areas: We bridge two historically separate fields software obfuscation and hardware (circuit) obfuscation by integrating ideas from both domains. SBHA harnesses over a decade of prior research from both communities to improve software protection.

In summary, this paper introduces SBHA, a novel approach to software protection that combines software obfuscation with hardware abstraction, offering a new level of security against MATE attacks while maintaining high efficiency and low overhead.[70]

III. Results and Discussion

In real-world scenarios, the volume of data increases linearly over time. Social networking sites like Facebook and Twitter have already identified the potential for data growth that could become uncontrollable in the future. To manage this massive influx of data, the proposed method processes the data in parallel, breaking it into smaller chunks across distributed clusters, and then aggregates the results from these clusters to generate the final processed data.

Within the Hadoop framework, MapReduce is employed to handle tasks such as filtering, aggregation, and maintaining efficient storage structures. The data is refined using collaborative filtering techniques, which help predict and identify the specific data requested by users. Additionally, the proposed method is enhanced

with sentiment analysis, leveraging natural language processing (NLP) to parse data into tokens. Emoticon-based clustering is then applied, grouping data according to user emotions to match the needs of individual users.
The results demonstrate that this approach significantly improves the performance of complexity analysis, enhancing the overall efficiency and accuracy of data processing.[71]

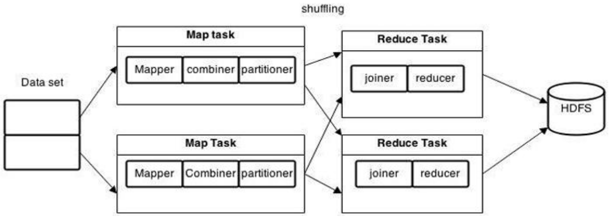
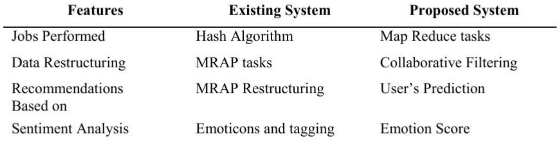


Fig 52. Execution Flow of a MapReduce process.[71]



Features	Existing System	Proposed System
Jobs Performed	Hash Algorithm	Map Reduce tasks
Data Restructuring	MRAP tasks	Collaborative Filtering
Recommendations Based on	MRAP Restructuring	User's Prediction
Sentiment Analysis	Emoticons and tagging	Emotion Score

Fig 53. Comparison between Existing and Proposed Systems [71]

To meet class-specific Quality of Service (QoS) requirements in next-generation wireless networks, it is crucial to manage broadband services efficiently. In this context, a self-optimization technique has emerged as a practical solution for controlling and managing large-scale data networks. This technique enables autonomous control of resources and key performance indicators (KPIs) without human intervention, relying solely on network intelligence.

This study introduces a Big Data-based Self-Optimization Networking (BD-SON) model for wireless networks, where KPI parameters affecting QoS are controlled through a multi-dimensional decision-making process. The model incorporates a Resource Management Center (RMC) that allocates the necessary resources to different parts of the network, based on decisions made by the SON engine. This allocation process ensures that the QoS constraints of multicast sessions are met, with the primary challenge being the management of interference constraints.

Additionally, a Load-Balanced Gradient Power Allocation (L-GPA) scheme is applied to the QoS-aware multicast model. This scheme adjusts the transmission power levels based on link capacity requirements, helping to optimize network performance. Experimental results demonstrate that the proposed power allocation techniques significantly increase the likelihood of achieving an optimal solution. Furthermore, the results confirm that the BD-SON model delivers notable improvements in QoS, capacity, and load-balancing optimality, while maintaining low complexity in network management.[72]

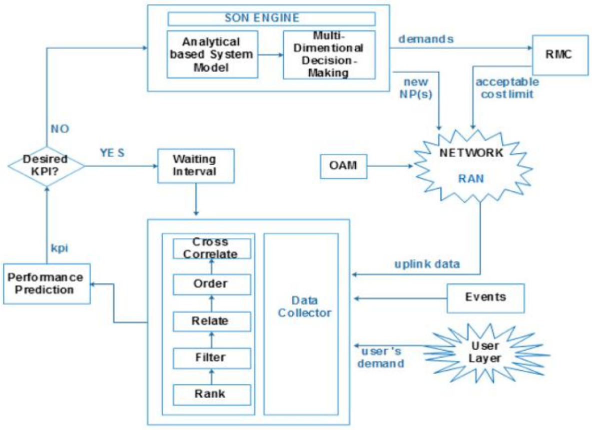


Fig 54. the proposed BD-Son Architecture[72]

Cloud storage systems have evolved to handle the massive volume of heterogeneous and rapidly changing data, commonly referred to as Big Data. However, due to the large-scale hardware components that make up these systems, failures are inevitable, posing significant challenges to ensuring the reliability and fault tolerance of cloud storage for Big Data applications.

Replication and erasure coding are two critical data reliability techniques commonly used in cloud storage systems. Each method comes with its own trade-offs in terms of durability, availability, storage overhead, network bandwidth and traffic, energy consumption, and recovery performance. This survey examines the challenges associated with employing both techniques in cloud storage systems for Big Data applications, considering the aforementioned parameters.

Furthermore, we propose a conceptual hybrid technique designed to enhance the reliability, latency, bandwidth usage, and storage efficiency of Big Data applications in cloud computing environments. This approach aims to address the shortcomings of each individual technique while improving overall system performance.[73]

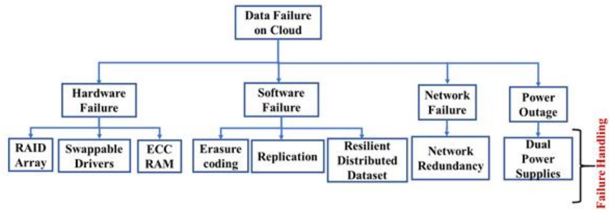
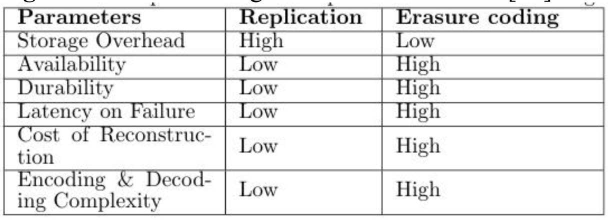


Fig 55. Failure Handling in Cloud Data Centres[73]



Parameters	Replication	Erasure coding
Storage Overhead	High	Low
Availability	Low	High
Durability	Low	High
Latency on Failure	Low	High
Cost of Reconstruction	Low	High
Encoding & Decoding Complexity	Low	High

Fig 56. Comparison between replication and erasure coding[73]

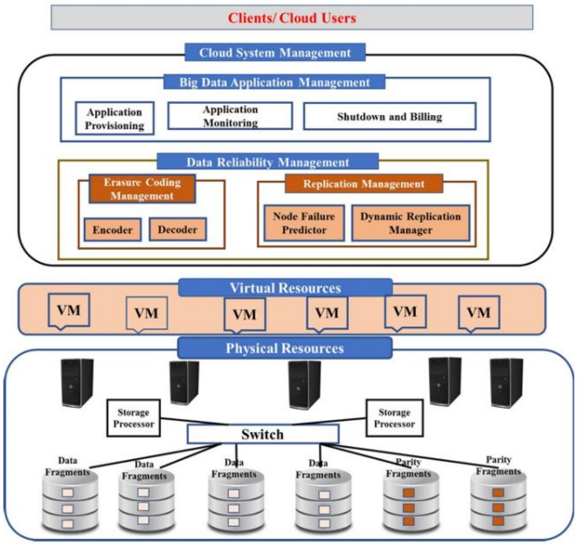
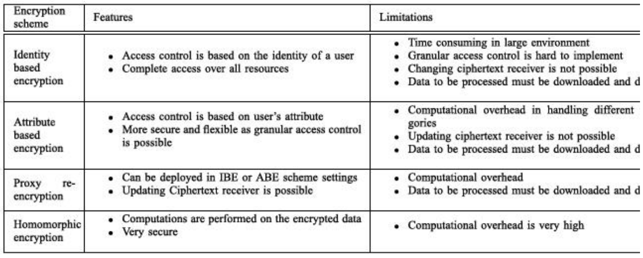


Fig 57. Reliability management of Big Data applications on cloud computing: concep tual architecture[73]

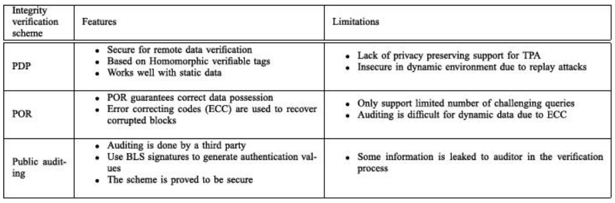
In recent years, Big Data has become a prominent research topic, driven by the exponential growth in data volumes. However, this rapid expansion also increases the risk of privacy breaches, particularly as Big Data requires substantial computational power and storage capacity, often relying on distributed systems. The involvement of multiple parties in these systems further heightens the risk of privacy violations. To address these concerns, various privacy-preserving mechanisms have been developed to protect privacy at different stages of the Big Data lifecycle, including data generation, storage, and processing.

This paper provides a comprehensive overview of these privacy-preserving mechanisms and highlights the challenges faced by existing solutions. Specifically, we examine the infrastructure of Big Data and the state-of-the-art privacy protection techniques employed at each stage of the Big Data lifecycle. Additionally, we discuss the ongoing challenges and identify key areas for future research in privacy preservation within the context of Big Data.[74]



Encryption scheme	Features	Limitations
Identity based encryption	• Access control is based on the identity of a user • Complete access over all resources	• Time consuming in large environment • Granular access control is hard to implement • Changing ciphertext receiver is not possible • Data to be processed must be downloaded and d
Attribute based encryption	• Access control is based on user's attribute • More secure and flexible as granular access control is possible	• Computational overhead in handling different gories • Updating ciphertext receiver is not possible • Data to be processed must be downloaded and d
Proxy re-encryption	• Can be deployed in IBE or ABE scheme settings • Updating Ciphertext receiver is possible	• Computational overhead • Data to be processed must be downloaded and d
Homomorphic encryption	• Computations are performed on the encrypted data • Very secure	• Computational overhead is very high

Fig 58. Comparison of encryption schemes.[74]



Integrity verification scheme	Features	Limitations
PDP	• Secure for remote data verification • Based on Homomorphic verifiable tags • Works well with static data	• Lack of privacy preserving support for TPA • Insecure in dynamic environment due to replay attacks
POR	• POR guarantees correct data possession • Error correcting codes (ECC) are used to recover corrupted blocks	• Only support limited number of challenging queries • Auditing is difficult for dynamic data due to ECC
Public auditing	• Auditing is done by a third party • Use BLS signatures to generate authentication values • The scheme is proved to be secure	• Some information is leaked to auditor in the verification process

Fig 59. Comparison of integrity verification schemes[74]

In recent years, a vast amount of structured, unstructured, and semi-structured data has been generated by institutions worldwide, collectively referred to as *Big Data*. The healthcare sector, in particular, faces the challenge of managing this large volume of heterogeneous data produced by various sources, including electronic health records, medical imaging, and genomic databases. To address this challenge, a range of big data analytics tools and techniques have been developed, particularly within the Hadoop ecosystem.

This paper explores the impact of big data in healthcare and reviews the tools available within the Hadoop ecosystem for processing and analyzing this data. Additionally, we examine the conceptual architecture of big data analytics in healthcare, which encompasses data from various sources, including clinical decision support systems, genomic databases, electronic health records, and text or imagery data. By understanding these tools and frameworks, we can better address the challenges of managing and utilizing big data in healthcare to improve patient care and decision-making.[75]

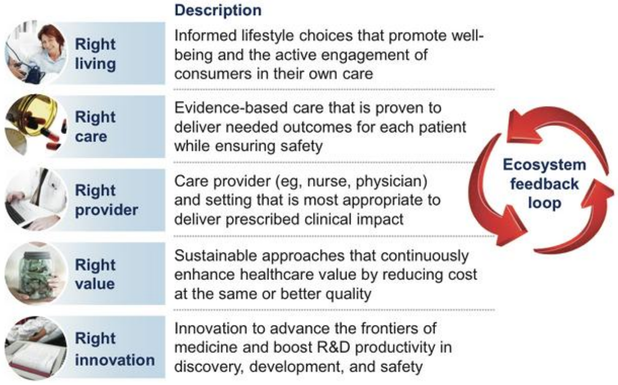

Description
Right living
Informed lifestyle choices that promote well-being and the active engagement of consumers in their own care
Right care
Evidence-based care that is proven to deliver needed outcomes for each patient while ensuring safety
Right provider
Care provider (eg, nurse, physician) and setting that is most appropriate to deliver prescribed clinical impact
Right value
Sustainable approaches that continuously enhance healthcare value by reducing cost at the same or better quality
Right innovation
Innovation to advance the frontiers of medicine and boost R&D productivity in discovery, development, and safety
Ecosystem feedback loop

Fig 60. Newpathways that have impact on big data.[75]

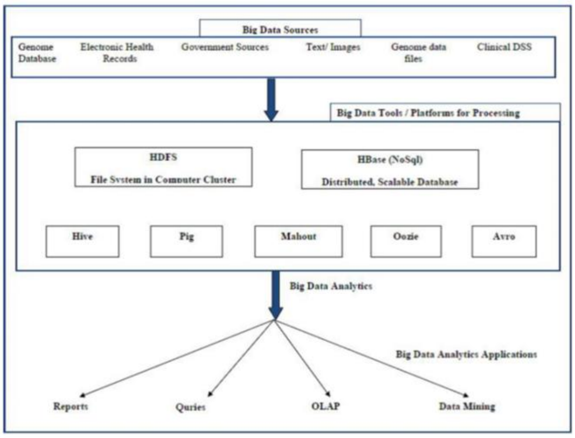


Fig 61. Conceptual architecture of big data analytics for health informatics.[75]

Cloud computing provides a flexible and scalable model that allows users and organizations to purchase services based on their specific needs, offering a wide range of services such as storage, deployment platforms, and easy access to web services. However, one of the common challenges in cloud environments is load balancing (LB), which is critical for maintaining application performance while ensuring adherence to Quality of Service (QoS) metrics and Service Level Agreements (SLA) set by cloud providers.

Load balancing becomes particularly challenging for cloud providers who must efficiently distribute workloads across servers to maintain optimal performance. An effective LB technique should not only optimize resource usage but also ensure high user satisfaction by efficiently utilizing virtual machine (VM) resources.

This paper provides a comprehensive review of various load balancing techniques in cloud environments, focusing on static, dynamic, and nature-inspired approaches. These techniques are evaluated in terms of their impact on data center response time and overall system performance. The paper also includes an analytical review of key algorithms and identifies research gaps, offering a perspective for future developments in this area. To aid in understanding, a graphical representation of the reviewed algorithms is provided to illustrate their operational flow. Additionally, the paper presents a fault-tolerant framework and explores other existing frameworks in the recent literature.[76]

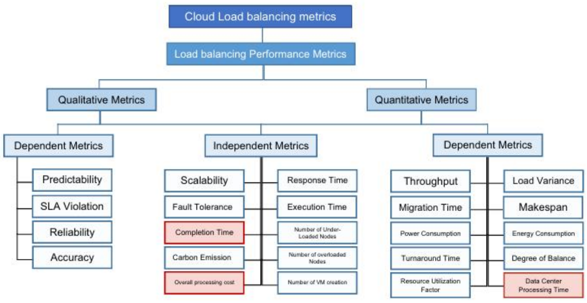


Fig 62. Classification of Load Balancing Metrics[76]

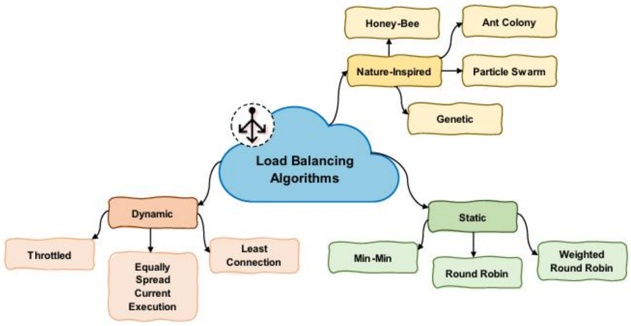


Fig 63. Taxonomy of Load Balancing Algorithms[76]

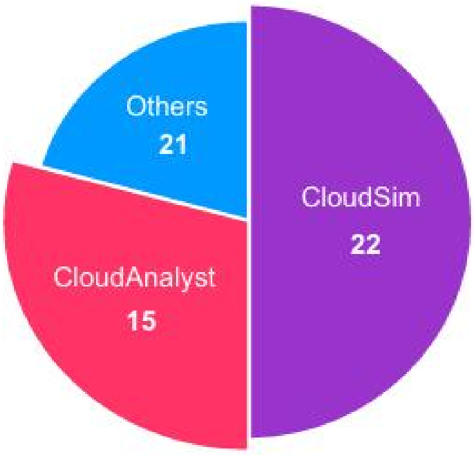


Fig 64. Simulation Tools of Reviewed Articles[76]

Clustering algorithms have become a powerful tool in meta-learning, enabling accurate analysis of the massive volumes of data generated by modern applications. Their primary goal is to categorize data into clusters, grouping similar objects together based on specific metrics. Despite the vast body of knowledge in the field of clustering, a significant challenge remains in the application of these algorithms to big data. A major issue is the lack of consensus on the definition of clustering properties and the absence of a formal categorization framework, which often leads to confusion among practitioners.

This paper addresses these challenges by introducing key concepts and algorithms related to clustering. It provides a concise survey of existing clustering techniques and offers both theoretical and empirical comparisons. From a theoretical standpoint, we propose a categorization framework based on the key properties identified in previous studies. Empirically, we conduct extensive experiments comparing the most representative algorithms from each category, using a large set of real-world big data.

The effectiveness of these clustering algorithms is evaluated using various internal and external validity metrics, as well as tests for stability, runtime, and scalability. Additionally, we identify the best-performing clustering algorithms for big data applications.[77]

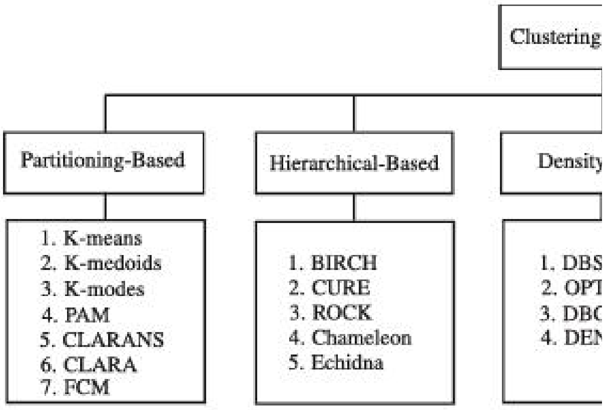
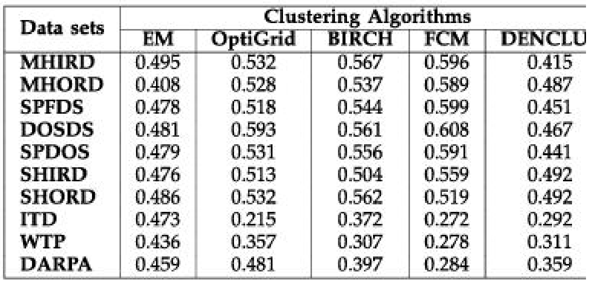
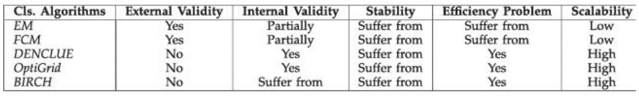


Fig 65. An overview of clustering taxonomy.[77]



Data sets	Clustering Algorithms				
	EM	OptiGrid	BIRCH	FCM	DENCLU
MHIRD	0.495	0.532	0.567	0.596	0.415
MHORD	0.408	0.528	0.537	0.589	0.487
SPFDS	0.478	0.518	0.544	0.599	0.451
DOSDS	0.481	0.593	0.561	0.608	0.467
SPDOS	0.479	0.531	0.556	0.591	0.441
SHIRD	0.476	0.513	0.504	0.559	0.492
SHORD	0.486	0.532	0.562	0.519	0.492
ITD	0.473	0.215	0.372	0.272	0.292
WTP	0.436	0.357	0.307	0.278	0.311
DARPA	0.459	0.481	0.397	0.284	0.359

Fig 66. Stability of the candidate clustering algorithms.[77]



Cls. Algorithms	External Validity	Internal Validity	Stability	Efficiency Problem	Scalability
EM	Yes	Partially	Suffer from	Suffer from	Low
FCM	Yes	Partially	Suffer from	Suffer from	Low
DENCLUE	No	Yes	Suffer from	Yes	High
OptiGrid	No	Yes	Suffer from	Yes	High
BIRCH	No	Suffer from	Suffer from	Yes	High

Fig 67. Compliance summary of the clustering algorithms based on empirical evaluation metrics.[77]

In 2004, Walmart boasted the world's largest data warehouse, with a storage capacity of 500 terabytes—equivalent to 50 printed collections of the U.S. Library of Congress. By 2009, eBay's storage had reached eight petabytes, roughly equivalent to 104 years of HD TV video. Just two years later, Yahoo's data center grew to 170 petabytes—about 8.5 times the total hard drive storage produced in 1995. As digitalization has expanded, organizations across various sectors have accumulated vast amounts of data, capturing trillions of bytes about their customers, suppliers, and operations. The volume of data continues to grow exponentially due to the increase in machine-generated data (e.g., log files, sensor data) and the surge in human-generated content on social media platforms.

Given the sheer scale of data involved, processing such massive volumes has become a significant challenge, which has led to the rise of "big data" technologies. This paper explores key aspects of big data, including its core components and the associated cost considerations. In Section II, we discuss the important aspects of big data, with a focus on its costing issues. Section III addresses the challenges organizations face when transitioning from traditional database systems to big data architectures.[78]

As emerging technologies and associated devices continue to evolve, it is predicted that a massive volume of data will be generated in the coming years. In fact, nearly 90% of the current data has been created in just the past few years, and this trend is expected to persist for the foreseeable future. Sustainable computing focuses on designing computers and their subsystems efficiently and effectively, with minimal environmental impact.

However, current machine learning systems, particularly in the context of intelligent applications, are primarily performance-driven. The emphasis is on predictive or classification accuracy based on known properties derived from training samples. For example, many nonparametric machine-learning models require substantial computational resources to identify global optima. As datasets grow larger, the number of hidden nodes in the model typically increases, resulting in an exponential rise in computational complexity.

This paper reviews both theoretical and experimental data modeling literature from large-scale, data-intensive fields. It covers (1) model efficiency, including the computational requirements for learning and the structure and design of data-intensive systems, and (2) introduces new algorithmic approaches that minimize memory requirements and processing demands. These approaches aim to reduce computational costs while maintaining or improving predictive accuracy, classification performance, and overall system stability.[79]

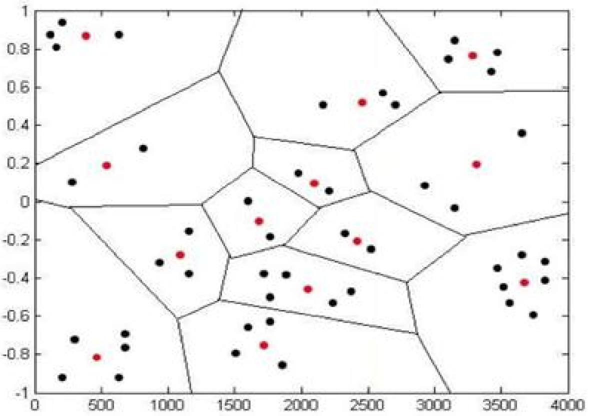


Fig 68. Two-dimensional (2D) vector quantization[79]

As the volume of data continues to grow exponentially, traditional strategies for handling big data are becoming increasingly inadequate. To effectively analyze such large datasets, alternative approaches are needed to partition the data into statistically representative blocks that can be used for analysis. This paper introduces the Random Sample Partition (RSP) distributed data model, which divides a big data set into disjoint data blocks, referred to as RSP blocks. Each RSP block is designed to have a probability distribution that closely mirrors that of the entire data set.

These RSP blocks can be leveraged to estimate the statistical properties of the data and build predictive models without the need to process the entire dataset. The paper highlights the advantages of using the RSP model for sampling from big data and presents a new RSP-based method for approximate big data analysis. This method offers significant reductions in computational complexity, making it particularly useful in real-world industrial applications. By using the RSP model, data scientists can increase their productivity and efficiently manage the computational demands of big data analysis.[80]

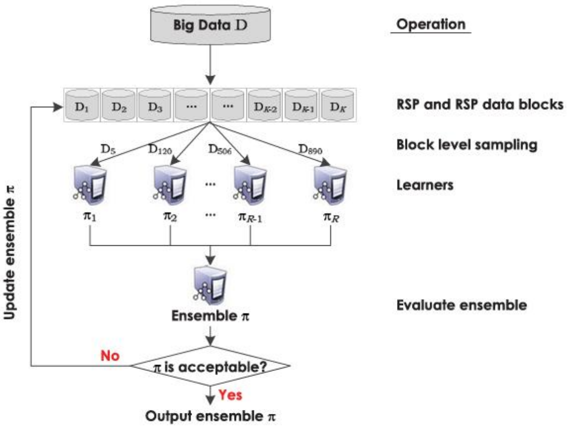


Fig 69. Asymptotic Ensemble Learning Framework[80]

The future of humanity increasingly relies on the performance and sustainability of cities. Big data offers powerful tools to study and improve interconnected urban systems—spanning environmental, social, and economic factors—to achieve urban sustainability. However, the definition of big data and its application to urban studies have been inconsistent, with research on this topic remaining sporadic in both focus and geographical scope. Therefore, a comprehensive review of big data-based urban environment, society, and sustainability (UESS) research is essential.

This study aims to provide a systematic review of big data-based UESS research, combining bibliometric and thematic analyses. Our findings reveal a sharp increase in the number of publications and citations in recent years. Human behavior data stands out as the most commonly used type of big data in UESS research. The primary analytical methods employed include classification, clustering, regression, association rules, and social network analysis.

Key research topics in big data-based UESS include urban mobility, land use and planning, environmental sustainability, public health and safety, social equity, tourism, energy and resource utilization, as well as sectors like real estate, retail, accommodation, and catering. Big data enhances UESS research by providing people-oriented perspectives, real-time information, and fine-resolution spatial dynamics.

However, challenges persist in applying big data to UESS research, including issues with data quality and acquisition, storage and management, security and privacy, as well as data cleaning, preprocessing, and analysis. To advance the field, future research should focus on integrating multiple big data sources, utilizing emerging methods such as deep learning and cloud computing, and expanding research into the interactions between human activities and urban environments.

This review aims to provide a clearer understanding of the current state of big data-based UESS research and offer insights for future studies in this growing field.[81]

In recent years, numerous business cases leveraging big data have been realized. Prominent examples include social networking giants like Twitter, LinkedIn, and Facebook, as well as companies in other sectors such as Netflix, which focuses on streaming movies, and various organizations improving network traffic monitoring or enhancing processes in manufacturing. Additionally, implementation architectures for these big data use cases have been published. However, conceptual work that integrates these diverse approaches into a unified reference architecture remains limited.

This paper addresses this gap by proposing a technology-independent reference architecture for big data systems, derived from an analysis of published implementation architectures across different big data use cases. A key contribution of this paper is also the classification of related implementation technologies and products/services, based on an analysis of the published use cases and a review of relevant literature.

The reference architecture and classification provided in this paper aim to assist in the architecture design process and guide the selection of technologies or commercial solutions when constructing big data systems. By consolidating these insights, the paper offers a cohesive framework that can be applied to a wide range of big data applications.[82]

The growing popularity of big data analytics (BDA) has prompted organizations to explore the potential of their large-scale data. BDA has become a strategic tool for organizations, offering opportunities to achieve competitive advantage and sustainable growth. However, much of the previous research on BDA has focused primarily on the traits of big data, known as the "Vs" (volume, velocity, variety, etc.), while giving less attention to the quality of the data used in BDA applications.

To address this gap, this study aims to investigate the impact of both big data traits and data quality dimensions on the application of BDA. The study formulated 10 hypotheses that examine the relationships between big data traits, key data quality dimensions (such as accuracy, believability, completeness, timeliness, and ease of operation), and BDA application outcomes.

A survey was conducted using a questionnaire to collect data, and partial least squares structural equation modeling (PLS-SEM) was employed to analyze the relationships between the constructs. The results revealed that big data traits significantly influence all data quality dimensions, and that ease of operation has a notable effect on the application of BDA.

This study contributes to the BDA literature by providing new insights into how big data traits and data quality dimensions interact to influence BDA application. The findings may serve as valuable guidance for both future researchers and practitioners seeking to better understand and implement BDA in real-world settings.[83]

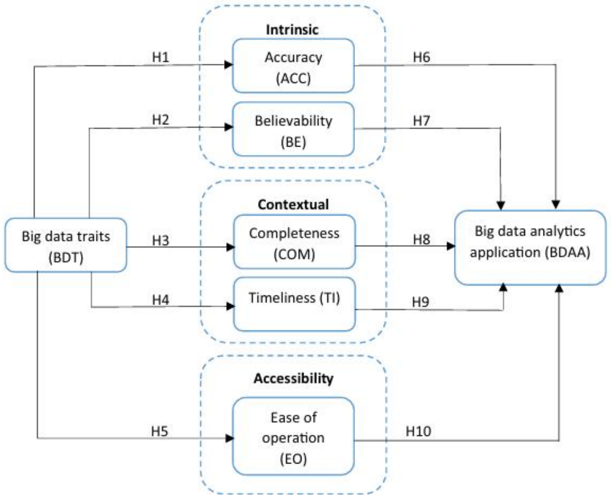


Fig 70. The proposed conceptual model for BDA application [83]

A cloud framework refers to the collection of components such as development tools, middleware, and database services that are essential for cloud computing. These components help in the development, deployment, and management of cloud-based applications, making the cloud an effective paradigm for scaling dynamically allocated resources and handling complex computations. Big Data Analytics (BDA) plays a crucial role in cloud architecture by providing data management solutions for storing, analyzing, and processing large volumes of data.

This paper offers a performance-based comparative analysis of cloud-based big data frameworks from leading enterprises such as Amazon, Google, IBM, and Microsoft. The aim is to provide researchers, IT analysts, and business users with valuable insights to help them select the most suitable framework for their specific needs, ultimately ensuring successful outcomes.[84]

The recent surge of big data has significantly impacted Malaysia, prompting both industry and academic communities to actively address the challenges of insight, hindsight, and foresight. This effort aims to position Malaysia among the global leaders in the big data information economy over the next decade. The rapid advancement of Information and Communication Technology (ICT) has been pivotal, as the growing number of users accessing data continues to increase. This phenomenon has come to be known as "big data."

So, what exactly is big data? We define big data as a valuable asset that requires a specialized platform to manage, mine, and analyze datasets with characteristics beyond the capacity of traditional data storage systems. The unique behavior of big data is driven by three key attributes: volume, velocity, and variety. To handle these attributes, new architectures, techniques, algorithms, and analytics are essential to uncover the valuable insights hidden within large volumes of data.

In this paper, we share our experiences in setting up a Data Science/Big Data platform, developing algorithms, and creating tools that align with the "plug-and-play" model for big data. Our focus is on integrating these solutions within the academic environment while also providing services to the wider community and industry.[85] In this paper, we propose a novel secure role re-encryption system (SRRS) designed to prevent privacy data leakage in cloud environments, while also enabling authorized deduplication and supporting dynamic privilege updates and revocations. The system efficiently ensures proof of ownership and supports ownership verification for authorized users.

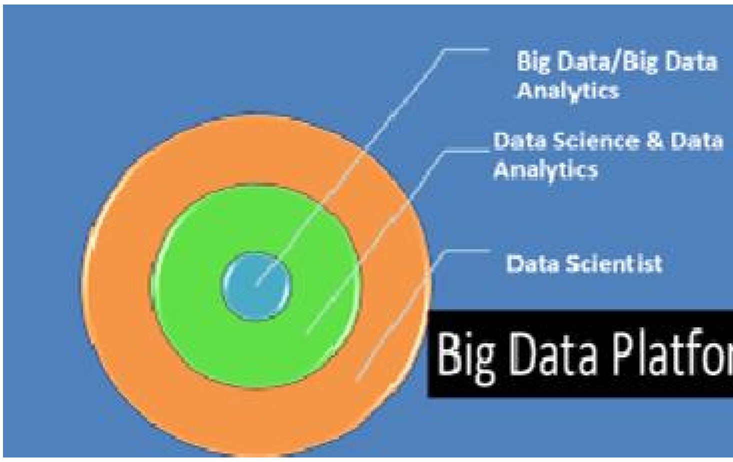


Fig 71. Correlations between Data Science and Big platform g Data under big data [85]

Our SRRS utilizes convergent encryption and a role re-encryption algorithm to achieve secure data access. Specifically, we introduce a management center that handles authorized requests and establishes a role-authorized tree (RAT) to map the relationship between roles and encryption keys. By leveraging convergent encryption and role re-encryption techniques, we ensure that only authorized users with the corresponding re-encryption key can access specific files, preventing unauthorized data access and leakage.

The system also supports dynamic privilege management by enabling the updating and revoking of role re-encryption keys, allowing for the flexible and secure management of user privileges. Additionally, we implement dynamic count filters (DCF) to facilitate efficient data updates and enhance ownership verification during retrieval.

We conduct both security analysis and simulation experiments to demonstrate the effectiveness and efficiency of our proposed system in ensuring data privacy, security, and proper access control in cloud environments.[86]

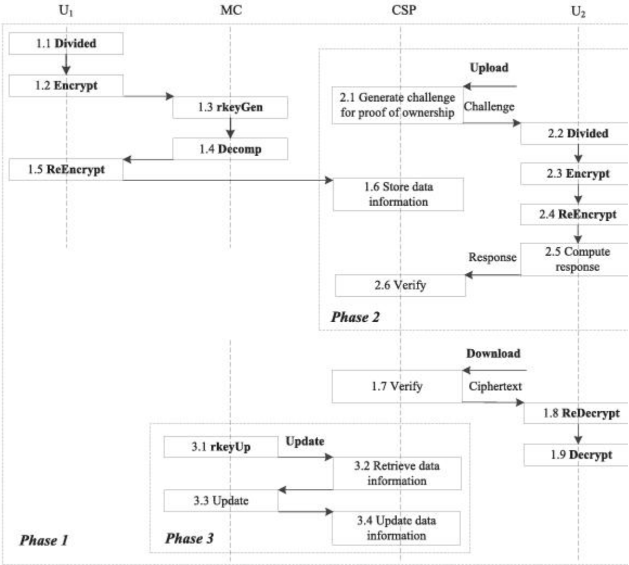


Fig 72. Overview of the SRRS system.[86]

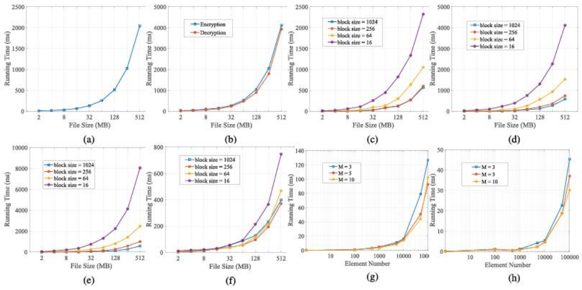


Fig 73. Performance analysis of system. (a) The computational cost of generating file token. (b) The computational cost of encryption and decryption. (c) The computational cost of role re-encryption (rkey 256bit). (d) T computational cost of role re-encryption (rkey (e) The computational cost of role re-encryption (rkey 512bit). 1024bit). (f) The computational cost of re-decryption. (g) T computational cost of RAT initialization. (h) The computational cost of RAT retrieval.[86]

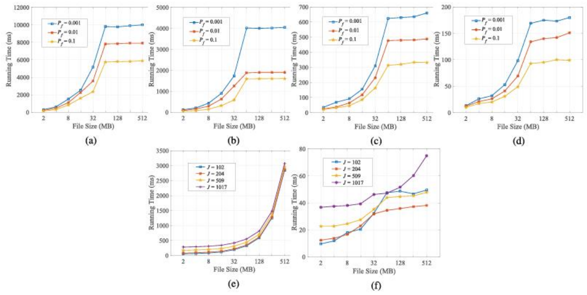


Fig 74. Performance analysis of system. (a) The computational cost of DCF initialization (token DCF initialization (token 64bit). (c) The computational cost of DCF initialization (token 16bit). (b) The computational cost of 256bit). (d) The computational cost of DCF initialization (token 1024bit). (e) The computational cost of PoW challenge generation. (f) The computational cost of PoW verification.[86]

Our clustering analysis of international collaborations identified two major research clusters: one linking Asia and North America, and another centered in Europe. From this analysis, three key research areas emerged: (i) energy provision, primarily relying on microbial fuel cells to power marine robots; (ii) biomaterials, which are still in development for fully operational soft-robotic solutions; and (iii) design and control, with a focus on optimizing locomotor designs.

Despite advancements in these areas, marine biomimetic robotics still faces significant challenges, particularly in providing long-lasting energy solutions that hinder operational autonomy. Consequently, there is a pressing need in the research community to identify natural processes by which living organisms obtain and manage energy. By understanding these processes, researchers can develop more efficient systems to sustain energy-demanding tasks, while also optimizing natural designs to minimize energy consumption.[87]

The invention relates to systems, methods, and computer program products designed for the authorization and use of cross-linked resource instruments. This system enhances the flexibility of resource transfers by enabling a user to establish a virtual link between an account and a resource instrument that was not originally associated with the account. As a result, the user can complete transactions using their preferred account, even if the original resource instrument is lost, damaged, or otherwise rendered ineffective. Additionally, the system allows for real-time activation of cross-link requests, enabling users to quickly and efficiently complete transactions from any location.[88]

This predictive correlational quantitative research study investigates the effectiveness of existing Information System (IS) tools, such as e-learning platforms, in U.S. high schools. The study highlights the gap between the promising performance of these IS tools and the tangible educational outcomes they deliver, with a specific focus on educators.

Using Gable's IS Impact Measurement Model, the research employs Partial Least Squares Structural Equation Modeling (PLS-SEM) via SmartPLS 4 to validate a formative IS impact model within the context of high school education. The study gathered valid responses from 237 participants, including educators, administrators, and IT personnel across diverse U.S. high schools.

Quantitative statistical analysis in this study includes rigorous assessments of both the measurement and structural models for reliability and validity. The overall results reveal that several key variables such as Information Quality, Individual Impact and Organizational Impact have a significant positive effect on the Overall Impact experienced by educators. However, System Quality was found to have no significant impact on the overall outcomes.

The study aims to lay the groundwork for future research on the adoption and implementation of the IS impact model as an assessment framework in educational settings, encouraging data-driven decisions for IS integration. The findings provide empirical evidence to help bridge the IS gaps in U.S. high schools and offer insights into rationalizing the success and impact of IT in educational environments.[89]

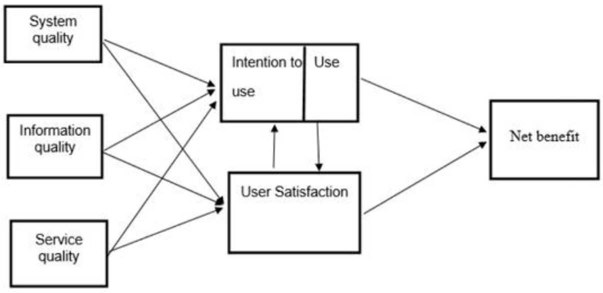


Fig 75. the updated version of the IS Success Model[89]

Technological advancements have significantly transformed product management by integrating artificial intelligence (AI) across various organizations. This research paper examines the impact of AI on automating routine product management tasks, thereby improving efficiency. The study also incorporates Resource-Based Theory (RBT) as a theoretical framework, exploring how AI capabilities contribute to identifying critical resources necessary for developing strong AI capabilities in product management.

The research methodology involved developing AI tools to measure AI capabilities and their relationship with organizational creativity, ultimately enhancing overall performance. The findings provide compelling evidence that AI can significantly boost organizational creativity and performance, highlighting its potential as a valuable resource in modern product management.[90]

Since the 1950s, both the plastic industry and consumer demand have surged, leading to a dramatic increase in plastic waste, particularly in the oceans, which has grown tenfold since 1980. This pollution poses significant threats to marine life, with many animal species unable to survive in these polluted environments. The effects are far-reaching, impacting plankton populations and disrupting the carbon cycle, which in turn contributes to global warming. This is just one example of how inefficient use of scarce resources, coupled with poor planning, generates waste and harms the natural world.

In response to these challenges, the concept of a circular economy (CE) has emerged as a promising alternative to the traditional, wasteful linear model of production. CE focuses on maximizing the value of raw materials throughout a product's entire life cycle, emphasizing reuse, recycling, and remanufacturing. This approach aims to minimize waste and reduce the environmental impact of industrial processes.

Two cutting-edge technologies that are poised to accelerate the adoption and implementation of CE are artificial intelligence (AI) and machine learning (ML). This research explores how AI applications are being integrated into CE, offering innovative solutions for waste reduction, resource efficiency, and sustainability in manufacturing practices.[91]

Data analysis has become crucial for the survival and success of organizations in today’s digital and competitive landscape. To store vast amounts of data, various OLTP ERP systems such as SAP S/4HANA, Amazon S3, and Oracle NetSuite are commonly used. However, accessing this data efficiently can be challenging due to the high costs associated with extraction, transformation, and loading (ETL) processes, as well as the complexities of data maintenance, monitoring, and visualization in data warehousing.

Serverless systems have emerged as a solution to simplify data analysis. These systems offer key advantages, including ease of operation, automatic scalability, and lower costs for accessing large datasets. Popular serverless platforms like Amazon Athena, Amazon Glue, Microsoft Azure, and Google BigQuery offer these features. Among them, Amazon Athena stands out for its cost-effectiveness, seamless integration with Amazon S3, scalability, security, and overall simplicity.

This paper explores the serverless architecture of Amazon Athena and provides insights on how to troubleshoot common issues encountered when accessing large data volumes using the platform.[92]

Artificial Intelligence (AI) has emerged as a transformative force across various industries, and higher education is no exception. This critical review explores the evolving role of AI in Science, Technology, Engineering, and Mathematics (STEM) education. It examines the impact of AI on various aspects of STEM higher education, including teaching and learning methodologies, curriculum design, student engagement, assessment practices, and institutional strategies. The paper also discusses the potential benefits and challenges of integrating AI into STEM education, while identifying key areas for future research and development. Ultimately, this article provides valuable insights into how AI can revolutionize STEM higher education and offers recommendations for fully leveraging its potential.[93]

Organizations are increasingly relying on vast amounts of data to drive business decisions. However, the data collected from various sources is often "dirty," which can compromise the accuracy of predictions and analyses. Data cleansing is essential for improving data quality, ensuring that the data is accurate and ready for analysis. As the volume of data continues to grow annually, many traditional data cleansing methods are becoming inadequate for handling big data.

The data cleansing process typically involves identifying errors, detecting issues, and correcting them. While this process is critical for ensuring high-quality data, it is also complex and time-consuming. Given the need for quick analysis, organizations face significant challenges in maintaining data quality. Furthermore, the role of domain experts is crucial in the data cleansing process, as their input is vital for verification and validation of the cleaned data.

This paper reviews the data cleansing process, explores the challenges posed by big data, and examines the available methods for data cleansing.[94]

Big Data plays a crucial role in the decision-making process by uncovering valuable insights that drive business and engineering advancements. However, managing and processing such large volumes of data presents significant challenges. Cloud computing has emerged as a key enabler, offering the necessary computational, networking, and storage capabilities to support Big Data operations. This paper provides a review of how cloud computing resources facilitate the transformation of Big Data, while also highlighting the opportunities and challenges involved in this process.[95]

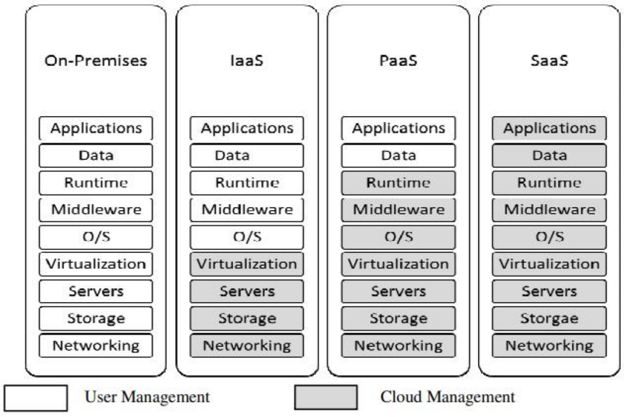


Fig 76. Summary of Key Differences[95]

Fig 77. Primary Cloud Computing Services[95]

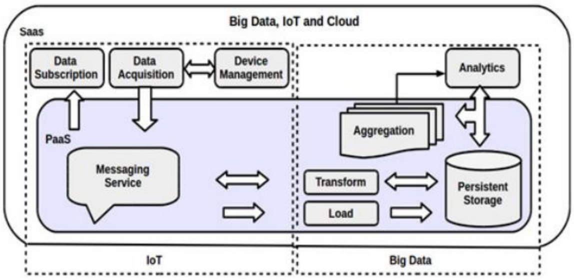
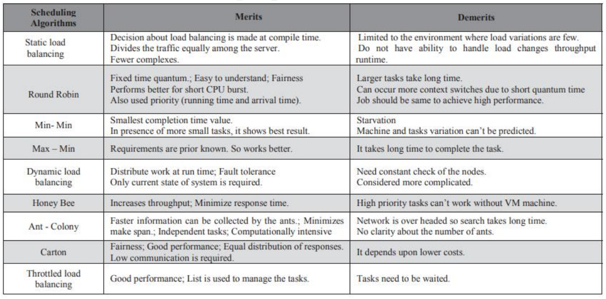


Fig 78. Overview of IoT, Big Data processing and Cloud Computing[95]

Cloud computing has gained significant popularity due to its appealing features, and as the development of new applications accelerates, the load on cloud systems is increasing rapidly. Load balancing, a crucial aspect of cloud computing, ensures that all devices or processors distribute the workload evenly, allowing each to perform an equal amount of work in a similar time frame. Over the years, various models and algorithms for load balancing have been developed to optimize cloud resource allocation, making cloud services more accessible and convenient for end users.

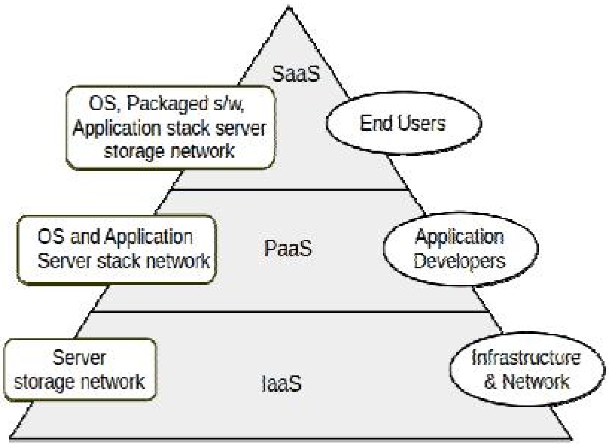
This paper provides a structured and comprehensive overview of research on load balancing algorithms in cloud computing. Specifically, it surveys the state-of-the-art tools and techniques developed between 2004 and 2015. We categorize the existing approaches based on their load balancing models, offering a clear and concise view of the underlying principles adopted by each method. This categorization helps in understanding the different strategies used to achieve fair and efficient load balancing in cloud environments.[96]



Scheduling Algorithms	Merits	Demerits
Static load balancing	Decision about load balancing is made at compile time. Divides the traffic equally among the server. Fewer complexes.	Limited to the environment where load variations are few. Do not have ability to handle load changes throughput runtime.
Round Robin	Fixed time quantum.; Easy to understand; Fairness Performs better for short CPU burst. Also used priority (running time and arrival time).	Larger tasks take long time. Can occur more context switches due to short quantum time Job should be same to achieve high performance.
Min- Min	Smallest completion time value. In presence of more small tasks, it shows best result.	Starvation Machine and tasks variation can't be predicted.
Max – Min	Requirements are prior known. So works better.	It takes long time to complete the task.
Dynamic load balancing	Distribute work at run time; Fault tolerance Only current state of system is required.	Need constant check of the nodes. Considered more complicated.
Honey Bee	Increases throughput; Minimize response time.	High priority tasks can't work without VM machine.
Ant - Colony	Faster information can be collected by the ants.; Minimizes make span.; Independent tasks; Computationally intensive	Network is over headed so search takes long time. No clarity about the number of ants.
Carton	Fairness; Good performance; Equal distribution of responses. Low communication is required.	It depends upon lower costs.
Throttled load balancing	Good performance; List is used to manage the tasks.	Tasks need to be waited.

Fig 79. Merits and demerits of load balancing algorithms[96]

The era of big data is upon us, but the enormous volumes of data involved may be too much for conventional data analytics techniques to handle. The creation of high-performance systems that can effectively analyze large amounts of data and the implementation of efficient mining algorithms to get insightful information from it provide a significant challenge. This article starts with an introduction to data analytics and then dives into a detailed discussion of big data analytics in order to thoroughly examine this problem.


SaaS
OS, Packaged s/w, Application stack server storage network
End Users
PaaS
OS and Application Server stack network
Application Developers
IaaS
Server storage network
Infrastructure & Network

Additionally, we identify important open issues and propose potential directions for future research in the field of big data analytics [97].

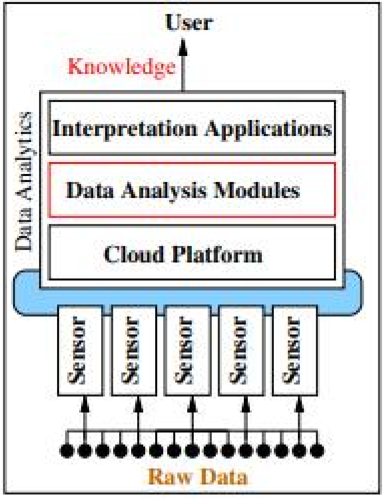


Fig 80. The basic idea of big data analytics on cloud system[97]

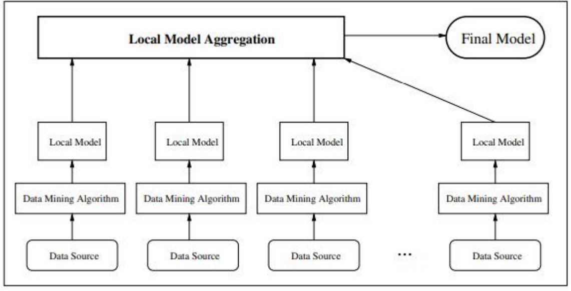


Fig 81. A simple example of distributed data mining framework[97]

The usage of cloud data centers has significantly increased as a result of the quick expansion of today's IT requirements. In this context, lowering these data centers' computational power usage is one of the main problems. Reducing the amount of resources allotted to a job can assist reduce energy usage since power consumption is directly proportional to the number of resources given to tasks.

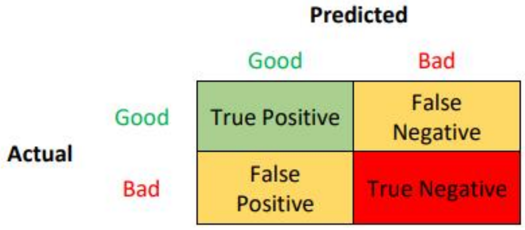
In this paper, we examine energy consumption in cloud environments across various services and propose strategies to promote green cloud computing. By optimizing resource allocation, we aim to reduce overall system energy consumption. Task allocation in cloud computing is a well-known challenge, and addressing it effectively can contribute to greener, more energy-efficient cloud operations.

An approach for adaptive work allocation that is tailored for varied cloud settings is what we suggest. This method aims to minimize the system's make span and lower energy usage at the same time. Our simulation findings show that the suggested approach performs better than current methods in terms of energy efficiency when we implemented it in the CloudSim simulation environment. The suggested method presents a viable way to maximize energy use in cloud data centers, supporting more environmentally friendly cloud computing procedures. [98].

Businesses are developing sophisticated artificial intelligence systems that can monitor security and infrastructure data to swiftly and automatically identify possible risks as a result of the increased threat of cyberattacks on a global scale. Machine learning-powered security analytics is a promising field in cybersecurity that seeks to mine security data and lower the high maintenance costs related to static rules and procedures.

Nevertheless, choosing the right method is a significant obstacle when using machine learning for security log analytics. This continues to be a major obstacle to the effective application of AI in cybersecurity, particularly in expansive or international Security Operations Centers (SOCs), where there is a high risk of false-positive detections. To reduce these mistakes and increase detection accuracy in such settings, selecting the appropriate machine learning strategy is essential.

A machine learning method that can reduce false positives is a crucial component of a successful cyber threat exposure system. Threat detection systems frequently use logistic regression, one of the most used machine learning techniques. As part of supervised learning, logistic regression is a fundamental classification algorithm that beginners in machine learning encounter early in their studies. It is essential for threat identification, but how well it is used in the larger framework of security log analytics and threat detection systems determines how successful it is[99].



		Predicted	
		Good	Bad
Actual	Good	True Positive	False Negative
	Bad	False Positive	True Negative

Fig 82. Confusion Matrix showing theaccuracy of the model on cyber threat detection[99]

Enterprise Resource Planning (ERP) systems have evolved significantly with the integration of emerging technologies like Machine Learning (ML), Artificial Intelligence (AI), and the Internet of Things (IoT). In particular, the use of SAP HANA as an in-memory database has been instrumental in enabling faster and more efficient processing of large data sets. Researchers such as Kunkulagunta and Raghunath have explored how ERP systems can leverage these technologies to optimize data analytics, anomaly detection, and predictive maintenance within industrial processes. By embedding machine learning models, ERP platforms can identify patterns, make real-time decisions, and predict outcomes that would otherwise be missed by traditional data processing methods. This application enhances system reliability and supports more informed decision-making for businesses, thereby increasing their operational efficiency [100-104].

Moreover, the shift to cloud-based ERP systems has opened up opportunities for scalability and remote access. Kunkulagunta's studies on IoT and cloud computing in ERP systems show how these technologies facilitate interconnectedness, enabling more dynamic resource planning and management. The cloud allows for decentralized data access, which is essential for modern, distributed work environments, and IoT enables real-time monitoring and coordination of industrial resources. Additionally, the adoption of AI-driven analytics and multi-agent systems in Industry 4.0 settings has shown promise in managing complex supply chain operations, as demonstrated by Raghunath's work on SAP HANA and Google Cloud integration. This combination of ERP with ML and cloud-based technologies creates an intelligent ecosystem where data-driven insights can continuously improve process efficiency and resilience [104-108].

By providing remote access and continuous, real-time monitoring, health monitoring systems are transforming patient care. The absence of continuous monitoring in traditional healthcare might cause delays in the detection of medical problems. Real-time data is collected by wearable sensors including heart rate and eye blink sensors, and it is then sent to the cloud-based ThingSpeak platform for analysis. Non-intrusive eye blink sensors, in particular, assist in diagnosing conditions like dry eye syndrome, blepharospasm, and Parkinson's disease. With remote monitoring, healthcare professionals can provide timely interventions, offering customizable alerts to ensure prompt action [108-113].

This method may improve patient outcomes and minimize problems by reducing the need for frequent hospital visits, particularly for chronic patients, by utilizing Internet of Things (IoT) technologies. The main objective is to provide smooth sensor communication to the ThingSpeak platform so that consumers may view their health information on mobile devices and PCs. In addition to improving overall monitoring capabilities and promoting a more patient-centric, effective healthcare model, this gives individuals the confidence to actively manage their health. [113-115].

Conclusion

In conclusion, the transformative power of big data, cloud computing, IoT, and AI across numerous industries has redefined how data is stored, processed, and analyzed, offering significant advancements in fields like healthcare, climate science, business, and urban planning. The ability to manage vast amounts of data in real-time has enabled innovative solutions, such as the Climate Analytics-as-a-Service (CAaaS) model for climate change and big data analytics in healthcare, which are critical for informed decision-making and policy planning. Technologies such as Hadoop and Spark frameworks, together with IoT integration, support these advancements by providing scalable infrastructure and enabling data-driven insights that shape effective responses to complex challenges.

However, despite the benefits, significant challenges remain in fully realizing the potential of these technologies. Issues surrounding data privacy, security, heterogeneity in data formats, and the need for effective load balancing in cloud environments are prominent concerns. The integration of techniques like differential privacy, contextual multi-armed bandit modeling, and secure role re-encryption systems aims to address these issues, fostering both user trust and data protection. At the same time, network management solutions such as SDN-based traffic optimization and hybrid storage approaches in cloud computing work to enhance efficiency and manageability of data-driven applications, but further research is necessary to refine these systems, especially as data volume and diversity continue to expand.

As these technologies evolve, a focus on sustainable computing and machine learning methods is crucial to manage the environmental impact of large-scale data processing. The development of efficient data architectures, alongside energy-saving IoT systems, promises not only operational improvements but also significant contributions to environmental stewardship. Future work in the field should prioritize building resilient, adaptable infrastructures that ensure data quality and support seamless integration across platforms. By addressing these challenges and embracing emerging technologies, industries can better harness the full potential of big data, AI, and cloud computing to drive growth, improve quality of life, and support long-term sustainability.

REFERENCES

[1] Raghunath V. *AN ERP Based Artificial Intelligence on the Basis of SAP High-Performance Analytic Appliances (HANA)*. Volume 1. 2024; p. 2.

[2] Raghunath V. *Generative AI-Driven Interface Device for SAP HANA Analytics*. Volume 3. 2024; p. 1.

[3] Yang C, et al. "Big Data and cloud computing: innovation opportunities and challenges." *International Journal of Digital Earth*. 2017;10(1):13–53.

[4] El-Seoud SA, et al. "Big Data and Cloud Computing: Trends and Challenges." *International Journal of Interactive Mobile Technologies*. 2017;11(2).

[5] Raghunath V. *New Rains Research Excellence Award 2024*. New Rains, Report No. 1. 2024.

[6] Ji C, et al. "Big data processing in cloud computing environments." *2012 12th International Symposium on Pervasive Systems, Algorithms and Networks*. IEEE. 2012.

[7] Sandhu AK. "Big data with cloud computing: Discussions and challenges." *Big Data Mining and Analytics*. 2021;5(1):32–40.

[8] Okorie GN, et al. "Leveraging big data for personalized marketing campaigns: a review." *International Journal of Management & Entrepreneurship Research*. 2024;6(1):216–242.

[9] Kunkulagunta M. "Enhancing Big Data Analytics with Artificial Intelligence: Innovative Techniques and Applications in Various Sectors." *IEEE Conference*. 2024;3(2):1–6.

[10] Bao G, Guo P. "Federated learning in cloud-edge collaborative architecture: key technologies, applications and challenges." *Journal of Cloud Computing*. 2022;11(1):94.

[11] Kaya M, Yildirim E. "Strategic Optimization of High-Volume Data Management: Advanced Techniques for Enhancing Scalability, Efficiency, and Reliability in Large-Scale Distributed Systems." *Journal of Intelligent Connectivity and Emerging Technologies*. 2024;9(9):16–44.

[12] Hashem IAT, et al. "The rise of 'big data' on cloud computing: Review and open research issues." *Information Systems*. 2015;47:98–115.

[13] Ji C, et al. "Big data processing: Big challenges and opportunities." *Journal of Interconnection Networks*. 2012;13(03n04):1250009.

[14] Yang C, et al. "Utilizing cloud computing to address big geospatial data challenges." *Computers, Environment and Urban Systems*. 2017;61:120–128.

[15] Kunkulagunta M. "Leveraging IoT and AI Technologies for Real-Time Remote Patient Monitoring: Innovations in Healthcare Delivery and Outcomes." *IEEE Conference*. 2024;1(2):101–107.

[16] Ji C, et al. "Big data processing: Big challenges and opportunities." *Journal of Interconnection Networks*. 2012;13(03n04):1250009.

[17] Yang C, et al. "Utilizing cloud computing to address big geospatial data challenges." *Computers, Environment and Urban Systems*. 2017;61:120–128.

[18] Fernández A, et al. "Big Data with Cloud Computing: an insight on the computing environment, MapReduce, and programming frameworks." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. 2014;4(5):380–409.

[19] George J. "Optimizing hybrid and multi-cloud architectures for real-time data streaming and analytics: Strategies for scalability and integration." *World Journal of Advanced Engineering Technology and Sciences*. 2022;7(1):10–30574.

[20] Raghunath V. "Investigation on Cloud Security Frameworks, Problems and Proposed Solutions." *European Journal of Advances in Engineering and Technology*. 2024;11(9):95–102.

[21] Raghunath V. "Security Issues Analysis Based on Big Data in Cloud Computing." *World Journal of Advanced Research and Reviews*. 2024;23(3):2549–2557.

[22] Zhong RY, et al. "Big Data for supply chain management in the service and manufacturing sectors: Challenges, opportunities, and future perspectives." *Computers & Industrial Engineering*. 2016;101:572–591.

[23] Deng X, et al. "Geospatial big data: New paradigm of remote sensing applications." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*. 2019;12(10):3841–3851.

[24] Zanoon N, Al-Haj A, Khwaldeh SM. "Cloud computing and big data: is there a relation between the two? A study." *International Journal of Applied Engineering Research*. 2017;12(17):6970–6982.

[25] Zanoon N, Al-Haj A, Khwaldeh SM. "Cloud computing and big data: Is there a relation between the two? A study." *International Journal of Applied Engineering Research*. 2017;12(17):6970–6982.

[26] Raghunath V. "Analysis on Addressing the Threats to Cloud Computing on the Basis of Security Safeguards for SAP Cloud Services." *World Journal of Advanced Research and Reviews*. 2024;23(3):2539–2548.

[27] Zanoon N, Al-Haj A, Khwaldeh SM. "Cloud computing and big data: Is there a relation between the two? A study." *International Journal of Applied Engineering Research*. 2017;12(17):6970–6982.

[28] Ali M, Essien A. "How can big data analytics improve outbound logistics in the UK retail sector? A qualitative study." *Journal of Enterprise Information Management*. 2023.

[29] Raghunath V. "SAP S/4HANA Applications on Data Security and Protections for SAP Cloud Services." *World Journal of Advanced Research and Reviews*. 2024;23(3):2530–2538.

[30] Coandă P, Avram M, Constantin V. "A state of the art of predictive maintenance techniques." *IOP Conference Series: Materials Science and Engineering*. 2020;997(1).

[31] Nanga S, et al. "Review of dimension reduction methods." *Journal of Data Analysis and Information Processing*. 2021;9(3):189–231.

[32] Raghunath V. "Investigating the Adaptive Supply Chain Module for the Integration of Google Cloud and SAP HANA Technologies." *International Journal of System Design and Information Processing (IJSDIP)*. 2024;12(2):115–134.

[33] Hashem IAT, Yaqoob I, Anuar NB, Mokhtar S, Gani A, Khan SU. "The rise of 'big data' on cloud computing: Review and open research issues." *Information Systems*. 2015;47:98–115.

[34] Yu JH, Zhou ZM. "Components and development in Big Data system: A survey." *Journal of Electronic Science and Technology*. 2019;17(1):51–72.

[35] Liu Y, Li N, Zhu X, Qi Y. "How wide is the application of genetic big data in biomedicine." *Biomedicine & Pharmacotherapy*. 2021;133:111074.

[36] Rabhi L, Falih N, Afraites A, Bouikhalene B. "Big data approach and its applications in various fields." *Procedia Computer Science*. 2019;155:599–605.

[37] —

[38] Kumar S, Mohbey KK. "A review on big data-based parallel and distributed approaches of pattern mining." *Journal of King Saud University-Computer and Information Sciences*. 2022;34(5):1639–1662.

[39] Batty M. "Big data, smart cities, and city planning." *Dialogues in Human Geography*. 2013;3(3):274–279.

[40] Kumar S, Singh M. "A novel clustering technique for efficient clustering of big data in the Hadoop Ecosystem." *Big Data Mining and Analytics*. 2019;2(4):240–247.

[41] Mahmud MS, Huang JZ, Salloum S, Emara TZ, Sadatdiynov K. "A survey of data partitioning and sampling methods to support big data analysis." *Big Data Mining and*

Analytics. 2020;3(2):85–101.

[42] Oussous A, Benjelloun FZ, Lahcen AA, Belfkih S. "Big Data technologies: A survey." *Journal of King Saud University-Computer and Information Sciences*. 2018;30(4):431–448.

[43] Misra R, Panda B, Tiwary M. "Big data and ICT applications: A study." *Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies*. 2016; pp. 1–6.

[44] Saraladevi B, Pazhaniraja N, Paul PV, Basha MS, Dhavachelvan P. "Big data and Hadoop—a study in security perspective." *Procedia Computer Science*. 2015;50:596–601.

[45] Maitrey S, Jha CK. "MapReduce: Simplified data analysis of big data." *Procedia Computer Science*. 2015;57:563–571.

[46] Gai K, Qiu M, Zhao H. "Privacy-preserving data encryption strategy for big data in mobile cloud computing." *IEEE Transactions on Big Data*. 2017;7(4):678–688.

[47] Chaudhary R, Aujla GS, Kumar N, Rodrigues JJ. "Optimized big data management across multi-cloud data centers: Software-defined-network-based analysis." *IEEE Communications Magazine*. 2018;56(2):118–126.

[48] Lv Z, Song H, Basanta-Val P, Steed A, Jo M. "Next-generation big data analytics: State of the art, challenges, and future research topics." *IEEE Transactions on Industrial Informatics*. 2017;13(4):1891–1899.

[49] Ranjbarzadeh R, et al. "Brain tumor segmentation of MRI images: A comprehensive review on the application of artificial intelligence tools." *Computers in Biology and Medicine*. 2023;152:106405.

[50] Bao G, Guo P. "Federated learning in cloud-edge collaborative architecture: Key technologies, applications, and challenges." *Journal of Cloud Computing*. 2022;11(1):94.

[51] Havens TC, Bezdek JC, Leckie C, Hall LO, Palaniswami M. "Fuzzy c-means algorithms for very large data." *IEEE Transactions on Fuzzy Systems*. 2012;20(6):1130–1146.

[52] Shafiq DA, Jhanjhi NZ, Abdullah A. "Load balancing techniques in cloud computing environment: A review." *Journal of King Saud University-Computer and Information Sciences*. 2022;34(7):3910–3933.

[53] Amamou S, Trifa Z, Khmakhem M. "Data protection in cloud computing: A Survey of the State-of-the-Art." —

[54] Sun P. "Security and privacy protection in cloud computing: Discussions and challenges." *Journal of Network and Computer Applications*. 2020;160:102642.

[55] Karimian-Aliabadi S, et al. "Analytical composite performance models for big data applications." *Journal of Network and Computer Applications*. 2019;142:63–75.

[56] O'Driscoll A, Daugelaite J, Sleator RD. "'Big data,' Hadoop, and cloud computing in genomics." *Journal of Biomedical Informatics*. 2013;46(5):774–781.

[57] D. Chatzievangelou et al., "Integrating diel vertical migrations of bioluminescent deep scattering layers into monitoring programs," *Frontiers in Marine Science*, vol. 8, 2021, Art. no. 661809.

[58] J. Aguzzi et al., "Developing technological synergies between deep-sea and space research," *Elem Sci Anth.*, vol. 10, no. 1, 2022, Art. no. 00064.

[59] J. L. Schnase et al., "MERRA analytic services: Meeting the big data challenges of climate science through cloud-enabled climate analytics-as-a-service," *Computers, Environment and Urban Systems*, vol. 61, pp. 198–211, 2017.

[60] C. K. Leung et al., "Big data science on COVID-19 data," in *Proc. IEEE 14th Int. Conf. Big Data Sci. Eng. (BigDataSE)*, 2020, pp. 14–21.

[61] C. P. Chen and C. Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data," *Information Sciences*, vol. 275, pp. 314–347, 2014.

[62] T. Yang and Y. Zhao, "Application of cloud computing in biomedicine big data analysis cloud computing in big data," in *Proc. 2017 Int. Conf. Algorithms, Methodology, Models Appl. Emerging Technol. (ICAMMAET)*, 2017, pp. 1–3.

[63] G. S. Bhathal and A. Singh, "Big Data: Hadoop framework vulnerabilities, security issues and attacks," *Array*, vol. 1, Art. no. 100002, 2019.

[64] A. Katal, M. Wazid, and R. H. Goudar, "Big data: Issues, challenges, tools and good practices," in *Proc. 2013 Sixth Int. Conf. Contemporary Comput. (IC3)*, 2013, pp. 404–409.

[65] V. Kotsiou, G. Z. Papadopoulos, P. Chatzimisios, and F. Theoleyre, "LDSF: Low-latency distributed scheduling function for industrial Internet of Things," *IEEE Internet Things J.*, vol. 7, no. 9, pp. 8688–8699, 2020.

[66] A. P. Plageras, K. E. Psannis, C. Stergiou, H. Wang, and B. B. Gupta, "Efficient IoT-based sensor BIG Data collection–processing and analysis in smart buildings," *Future Generation Computer Systems*, vol. 82, pp. 349–357, 2018.

[67] C. L. Stergiou, K. E. Psannis, and B. B. Gupta, "IoT-based big data secure management in the fog over a 6G wireless network," *IEEE Internet Things J.*, vol. 8, no. 7, pp. 5164–5171, 2020.

[68] P. Zhou et al., "Privacy-preserving and residential context-aware online learning for IoT-enabled energy saving with big data support in smart home environment," *IEEE Internet Things J.*, vol. 6, no. 5, pp. 7450–7468, 2019.

[69] R. Manikyam, J. T. McDonald, W. R. Mahoney, T. R. Andel, and S. H. Russ, "Comparing the effectiveness of commercial obfuscators against MATE attacks," in *Proc. 6th Workshop Software Security, Protection, Reverse Eng.*, 2016, pp. 1–11.

[70] R. K. Manikyam, "Program protection using software-based hardware abstraction," Ph.D. dissertation, Univ. South Alabama, 2019.

[71] V. Subramaniyaswamy, V. Vijayakumar, R. Logesh, and V. Indragandhi, "Unstructured data analysis on big data using map reduce," *Procedia Computer Science*, vol. 50, pp. 456–465, 2015.

[72] A. Mohajer, M. Barari, and H. Zarrabi, "Big data based self-optimization networking: A novel approach beyond cognition," *Intell. Autom. Soft Comput.*, pp. 1–7, 2017.

[73] R. Nachiappan, B. Javadi, R. N. Calheiros, and K. M. Matawie, "Cloud storage reliability for big data applications:

A state-of-the-art survey," *J. Network Comput. Appl.*, vol. 97, pp. 35–47, 2017.

[74] A. Mehmood, I. Natgunanathan, Y. Xiang, G. Hua, and S. Guo, "Protection of big data privacy," *IEEE Access*, vol. 4, pp. 1821–1834, 2016.

[75] S. Kumar and M. Singh, "Big data analytics for healthcare industry: Impact, applications, and tools," *Big Data Mining Anal.*, vol. 2, no. 1, pp. 48–57, 2018.

[76] D. A. Shafiq, N. Z. Jhanjhi, and A. Abdullah, "Load balancing techniques in cloud computing environment: A review," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 34, no. 7, pp. 3910–3933, 2022.

[77] A. Fahad et al., "A survey of clustering algorithms for big data: Taxonomy and empirical analysis," *IEEE Trans. Emerg. Topics Comput.*, vol. 2, no. 3, pp. 267–279, 2014.

[78] K. S. Jadon, R. S. Bhadoria, and G. S. Tomar, "A review on costing issues in big data analytics," in *Proc. 2015 Int. Conf. Comput. Intell. Commun. Netw. (CICN)*, 2015, pp. 727–730.

[79] O. Y. Al-Jarrah, P. D. Yoo, S. Muhaidat, G. K. Karagiannidis, and K. Taha, "Efficient machine learning for big data: A review," *Big Data Res.*, vol. 2, no. 3, pp. 87–93, 2015.

[80] S. Salloum, J. Z. Huang, and Y. He, "Random sample partition: A distributed data model for big data analysis," *IEEE Trans. Ind. Inf.*, vol. 15, no. 11, pp. 5846–5854, 2019.

[81] L. Kong, Z. Liu, and J. Wu, "A systematic review of big data-based urban sustainability research: State-of-the-science and future directions," *J. Cleaner Prod.*, vol. 273, Art. no. 123142, 2020.

[82] P. Pääkkönen and D. Pakkala, "Reference architecture and classification of technologies, products and services for big data systems," *Big Data Res.*, vol. 2, no. 4, pp. 166–186, 2015.

[83] M. Wook et al., "Exploring big data traits and data quality dimensions for big data analytics application using partial least squares structural equation modelling," *J. Big Data*, vol. 8, Art. no. 1, 2021.

[84] S. Saif and S. Wazir, "Performance analysis of big data and cloud computing techniques: A survey," *Procedia Computer Science*, vol. 132, pp. 118–127, 2018.

[85] S. M. Shamsuddin and S. Hasan, "Data science vs big data@ UTM big data centre," in *Proc. 2015 Int. Conf. Sci. Inf. Technol. (ICSITech)*, 2015, pp. 1–4.

[86] J. Xiong, Y. Zhang, S. Tang, X. Liu, and Z. Yao, "Secure encrypted data with authorized deduplication in cloud," *IEEE Access*, vol. 7, pp. 75090–75104, 2019.

[87] J. Aguzzi et al., "Research trends and future perspectives in marine biomimicking robotics," *Sensors*, vol. 21, no. 11, Art. no. 3778, 2021.

[88] G. P. Buddha, S. P. Kumar, and C. M. R. Reddy, "U.S. Patent Application No. 17/203,879," 2022.

[89] G. S. Nadella, "Validating the Overall Impact of IS on Educators in US High Schools Using IS-Impact Model–A Quantitative PLS-SEM Approach," Ph.D. dissertation, University of the Cumberlands, 2023.

[90] P. Sharma and H. Gonaygunta, "Role of AI in Product Management Automation and Effectiveness," *SSRN*, 2021, Art. no. 4637857.

[91] S. C. Yarlagadda, "The Use of Artificial Intelligence and Machine Learning in Creating a Roadmap Towards a Circular Economy for Plastics," *Int. J. Recent Innov. Trends Comput. Commun.*, vol. 11, no. 9s, pp. 829–836, 2023.

[92] A. Kulkarni, "Amazon Athena: Serverless Architecture and Troubleshooting," *Int. J. Comput. Trends Technol.*, vol. 71, no. 5, pp. 57–61, 2023.

[93] B. K. Nagaraj, A. Kalaivani, S. Begum, S. Akila, and H. K. Sachdev, "The emerging role of artificial intelligence in stem higher education: A critical review," *Int. Res. J. Multidisciplinary Technovation*, vol. 5, no. 5, pp. 1–19, 2023.

[94] S. K. Mishra, D. Puthal, B. Sahoo, S. K. Jena, and M. S. Obaidat, "An adaptive task allocation technique for green cloud computing," *J. Supercomput.*, vol. 74, pp. 370–385, 2018.

[95] M. Muniswamaiah, T. Agerwala, and C. Tappert, "Big data in cloud computing review and opportunities," *arXiv preprint arXiv:1912.10821*, 2019.

[96] S. Aslam and M. A. Shah, "Load balancing algorithms in cloud computing: A survey of modern techniques," in *Proc. 2015 National Software Engineering Conf. (NSEC)*, 2015, pp. 30–35.

[97] C. W. Tsai, C. F. Lai, H. C. Chao, and A. V. Vasilakos, "Big data analytics: A survey," *J. Big Data*, vol. 2, pp. 1–32, 2015.

[98] S. K. Mishra, D. Puthal, B. Sahoo, S. K. Jena, and M. S. Obaidat, "An adaptive task allocation technique for green cloud computing," *J. Supercomput.*, vol. 74, pp. 370–385, 2018.

[99] H. Gonaygunta, "Machine learning algorithms for detection of cyber threats using logistic regression," Ph.D. dissertation, Dept. of Information Technology, University of the Cumberlands, 2023.

[100] D. Sivabalaselvamani et al., "Healthcare Monitoring and Analysis Using ThingSpeak IoT Platform: Capturing and Analyzing Sensor Data for Enhanced Patient Care," in *Adv. Appl. Osmotic Comput.*, IGI Global, 2024, pp. 126–150.

[101] M. Kunkulagunta, "ERP Based on Machine Learning for SAP HANA Analytics," 2024.

[102] M. Kunkulagunta, "Automated Device for Data Anomaly Detection in SAP HANA," 2024.

[103] M. Kunkulagunta, "Study on Connection Between ERP and Artificial Intelligence through the Application of ML Techniques in ERP Industrial Processes," *Processing*, vol. 3, no. 2, pp. 91–101, 2024.

[104] M. Kunkulagunta, "Evolution of Integrated Management Information Systems on the ERP Process System," *Processing*, vol. 2, no. 1, pp. 29–39, 2024.

[105] M. Kunkulagunta, "Study of Internet of Things Applications Supporting Technologies that are Used in the IT Sector," *Processing*, vol. 1, no. 2, pp. 1–15, 2024.

[106] M. Kunkulagunta, "Cloud Computing Applications for ERP Implementation," *Int. J. Comput. Eng. Technol. (IJCET)*, vol. 15, no. 2, 2024.

[107] M. Kunkulagunta, "Role of Machine Learning, Data Mining, and Analytics," *Int. J. Comput. Eng. Technol. (IJCET)*, vol. 15, no. 2, 2024.

[108] M. Kunkulagunta, "Studying of Exploring Based on Impacts of Artificial Intelligence & Machine Learning on Enterprise Resource Planning," *Int. J. Artif. Intell. & Machine Learn. (IJAIML)*, 2024.

[109] M. Kunkulagunta, "IoT-Based Enterprise Resource Planning: Challenges, Open Issue, Applications, and Architecture Analysis," *Int. J. Artif. Intell. & Machine Learn. (IJAIML)*, 2024.

[110] M. Kunkulagunta, "An Analysis of Adoption Factors and Attitudes Based on Cloud-Based ERP Systems," *Int. J. Data Anal. Res. Dev. (IJDARD)*, vol. 2, 2024.

[111] M. Kunkulagunta, "A Self-Organized Multi-Agent System with Industry 4.0 Coordination and Big Data-Based Feedback," *Int. J. Data Anal. Res. Dev. (IJDARD)*, vol. 2, 2024.

[112] M. Kunkulagunta, "New Rains Research Excellence Award 2024," *New Rains*, 2024.

[113] M. Kunkulagunta, "Analysis on Multi-Dimensional Model of Enterprise Resource Planning Critical Success Factors," *Int. J. System Design Inf. Process. (IJSDIP)*, vol. 12, no. 2, pp. 149–160, 2024.

[114] V. Raghunath, "Predictive Analytics on SAP Database (HANA) by Using Artificial Intelligence (AI) and Automated Machine Learning Capabilities," *Int. J. Comput. Eng. Technol. (IJCET)*, vol. 15, no. 3, pp. 65–78, 2024.

[115] V. Raghunath, "Investigating the Adaptive Supply Chain Module for the Integration of Google Cloud and SAP HANA Technologies," *Eur. J. Adv. Eng. Technol.*, vol. 11, no. 9, pp. 103–110, 2024.