



A SURVEY ON BIG DATA ANALYTICS: CHALLENGES, OPEN RESEARCH ISSUES, TOOLS, REAL-WORLD APPLICATIONS, TECHNOLOGIES AND INTEGRATION OF AI AND BIG DATA

¹Dr. M. Hemalatha

¹Associate Professor

¹PG & Research Department of Computer Science

¹Sri Ramakrishna college of Arts and Science, Coimbatore.

ABSTRACT: An enormous volume of terabytes of data is produced daily through modern digital technologies such as the Internet of Things (IoT) and cloud computing. Deriving meaningful insights from this vast data requires multi-tiered efforts. This paper examines the challenges of big data, open research issues, tools, real-world applications, emerging technologies, and AI integration, serving as a foundation for future research.

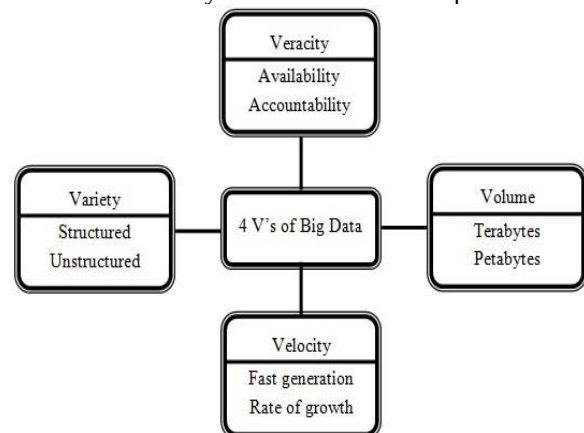
Keywords: [Big data analytics; Hadoop; Massive data; Structured data; Unstructured data;]

1. INTRODUCTION

In the digital age, rapid technological advancements have driven the growth of big data, which consists of vast and complex datasets that traditional processing tools struggle to handle. These datasets exist in structured, semi-structured, and unstructured formats. Big data is defined by the 3Vs—volume (data scale), velocity (data generation speed), and variety (data types), with veracity (data reliability) sometimes included. Its analysis aims to process large, fast, diverse, and accurate data using advanced techniques, enabling better decision-making, insights, and efficiency. However, defining big data remains problem-specific. Projected to reach 25 billion by 2015, big data drives IT industries through cloud computing, IoT, and social business. Data warehouses manage large datasets, but extracting valuable knowledge remains challenging due to gaps between database systems and analytical tools. Research focuses on defining big data's characteristics, addressing epistemological implications, and applying complexity theory to improve knowledge abstraction and computing models. Not all big data is useful, but academia and industry actively explore its potential. This paper discusses big data challenges, research gaps, tools, and applications, covering AI integration and technological advancements.

2. CHALLENGES IN BIG DATA ANALYTICS

Big data spans fields like healthcare, public administration, retail, and research. Web applications handle it in social computing and search indexing. While enabling knowledge processing, big data poses challenges in computation, security, and analytics. Key issues include storage, analysis, discovery, scalability, visualization, and security, requiring advanced methods beyond traditional techniques.



A. Data storage and analysis

The rapid growth of data from mobile devices, remote sensing, and RFID raises storage costs, often causing data deletion. Efficient high-speed storage remains a challenge despite SSDs and PCM. Data diversity complicates mining and feature selection, requiring better automation and scalable machine learning. Hadoop and MapReduce assist, but structured analysis is essential.

B. Knowledge Discovery and Computational Complexities

Big data knowledge discovery involves authentication, archiving, management, retrieval, and preservation. Techniques like fuzzy sets, rough sets, and PCA, along with hybrid methods, aid this process. Data warehouses and marts manage big data, but computational complexity and uncertainty persist. Machine learning research seeks cost

reduction, yet tools need further optimization for effective analysis.

C. Scalability and Visualization of Data

Big data analysis faces scalability and security challenges. Researchers enhance data processing with sampling, online, and multiresolution analysis, while parallel computing tackles growing data demands. Visualization tools like Tableau simplify complex data, aiding businesses, but lack efficiency. Integrating mathematical models with computer science is crucial for better scalability and analytics solutions.

D. Information Security

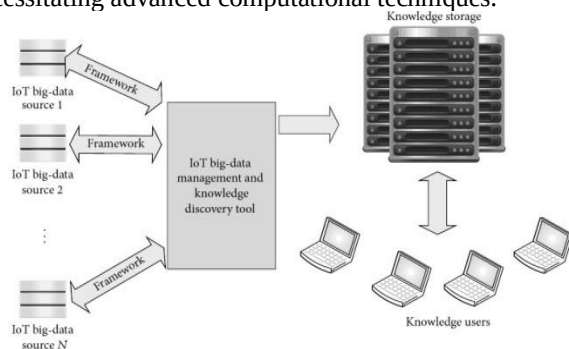
Big data analysis correlates, mines, and analyzes vast datasets, making data security essential. Organizations enforce policies to protect sensitive information, but risks persist. Authentication, authorization, and encryption enhance security, yet challenges like network scale, device diversity, and real-time monitoring remain. Advancing multi-level security models is crucial for privacy-preserving big data protection.

3. OPEN RESEARCH ISSUES IN BIG DATA ANALYTICS

Big data analytics and data science are key research areas in industry and academia, focusing on knowledge extraction and predictive analysis. Their applications include machine learning, statistical learning, and pattern recognition. Major research issues include IoT, cloud computing, bio-inspired computing, and quantum computing, with additional challenges in healthcare big data.

A. IoT for Data Analytics

The Internet has transformed global interactions, businesses, and cultural dynamics, now extending to autonomous devices through the Internet of Things (IoT). IoT enables devices to connect and communicate, presenting vast opportunities and challenges with economic and societal impacts. Advances in mobile devices, cloud computing, and data analytics drive its relevance. However, IoT data poses challenges due to its volume, velocity, and variety, necessitating advanced computational techniques.



IoT Big Data Knowledge Discovery

B. Cloud Computing for Big Data Analytics

Virtualization has made supercomputing accessible, allowing flexible computing resources. Cloud computing

leverages virtualization for scalable, cost-effective big data processing, enhancing availability and analytics. Platforms like Spark and R support interactive exploration. Users access Google, Amazon, or NetSuite services, but challenges remain, including data transfer costs and privacy concerns in public cloud storage.

C. Bio-inspired Computing for Big Data Analytics

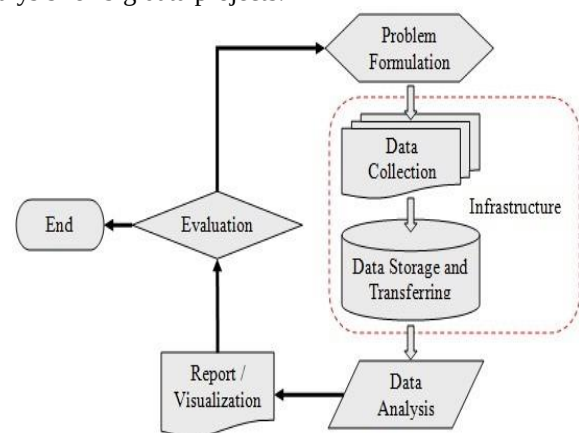
Bio-inspired computing mimics biological systems to solve complex problems through self-organization. It optimizes data management by reducing costs and efficiently processes data using DNA and proteins, making it ideal for big data. Balancing IoT, cloud, and bio-inspired computing enables cost-effective analytics, improving data mining, optimization, and intelligent interactions for future advancements.

D. Quantum Computing for Big Data Analysis

Quantum computers have exponentially larger memory and can process vast inputs simultaneously. Unlike classical bits, qubits leverage superposition and entanglement, enabling faster big data problem-solving. A 100-qubit system requires 2^{100} complex values, surpassing classical computing. Developing scalable quantum computers remains a challenge but holds immense potential for big data analytics.

4. TOOLS FOR BIG DATA PROCESSING

Various tools exist for big data processing, focusing on batch, stream, and interactive analysis. Key emerging tools include MapReduce, Apache Spark, and Storm. Hadoop-based tools like Mahout handle batch processing, while Storm and Splunk support real-time analytics. Interactive platforms like Dremel and Apache Drill enable user-driven analysis for big data projects.



Workflow of Big Data Project

A. Apache Hadoop and MapReduce

Apache Hadoop and MapReduce are leading platforms for big data analysis, comprising the Hadoop kernel, HDFS, and Apache Hive. MapReduce follows a divide-and-conquer approach with Map and Reduce steps. The master node assigns tasks to worker nodes, then aggregates results, enabling fault-tolerant storage and high-throughput data processing.

B. Apache Mahout

Apache Mahout provides scalable machine learning techniques for large-scale data analysis. Its core algorithms, including clustering, classification, pattern mining, and regression, run on Hadoop via MapReduce. Mahout fosters a collaborative community to address big data challenges. Companies like Google, IBM, Amazon, Yahoo, Twitter, and Facebook have implemented its algorithms.

C. Apache Spark

Apache Spark is an open-source big data framework designed for high-speed processing and advanced analytics. Developed at UC Berkeley's AMP Lab (2009) and open-sourced (2010), it supports Java, Scala, Python, and R. Running on Hadoop HDFS, Spark leverages RDDs for fault tolerance and efficiency, enabling applications to run 100x faster in memory.

D. Dryad

Dryad is a parallel and distributed computing model that processes data using a computational graph. It allows users to run applications across clusters without concurrent programming expertise. With features like job scheduling, failure handling, monitoring, and dynamic updates, Dryad efficiently scales across thousands of machines, optimizing task execution and resource management.

E. Storm

Storm is a distributed, fault-tolerant system for real-time stream processing, unlike Hadoop, which handles batch processing. It features scalability, reliability, and easy deployment. Storm clusters have Nimbus (master) and Supervisor (worker) nodes, managing task distribution. Unlike MapReduce jobs, Storm topologies run continuously until stopped, efficiently processing streaming data across multiple worker processes.

F. Apache Drill

Apache drill is another distributed system for interactive analysis of big data. It has more flexibility to support many types of query languages, data formats, and data sources. It is also specially designed to exploit nested data. Also, it has an objective to scale up on 10,000 servers or more and reaches the capability to process petabytes of data and trillions of records in seconds. Drill uses HDFS for storage and map reduce to perform batch analysis.

G. Jaspersoft

Jaspersoft is an open-source big data analytics tool that generates reports from databases. It supports fast data visualization on platforms like MongoDB, Cassandra, and Redis. A key feature is exploring big data without ETL, allowing users to create interactive HTML reports and dashboards and share them across organizations effortlessly.

H. Splunk

Splunk is a real-time analytics platform for processing machine-generated big data in industries. It integrates cloud and big data technologies to enable searching, monitoring,

and analysis via a web interface. Splunk uniquely indexes structured and unstructured data, offers real-time insights, dashboards, and reports, and supports IT diagnostics and business operations.

5. REAL-WORLD APPLICATIONS OF BIG DATA ANALYTICS

Big Data Analytics has transformed multiple industries by enabling data-driven decision-making, automation, and predictive insights. This section explores four key domains where big data is making a significant impact.

A. Healthcare: Revolutionizing Patient Care & Drug Discovery

Big data is playing a crucial role in predictive healthcare, personalized treatments, and operational efficiency. The massive amount of health data, including electronic health records (EHRs), medical imaging, genomic data, and real-time patient monitoring, has opened new frontiers in medical science.

Applications in Healthcare:

Disease Prediction & Early Diagnosis
Real-Time Patient Monitoring & IoT-Based Health Analytics
Drug Discovery & Genomics Research
Hospital Management & Predictive Analytics for Resource Optimization

B. Finance: Fraud Detection, Risk Management & Algorithmic Trading

The financial sector generates massive amounts of transactional data daily. Big data analytics helps in fraud detection, credit risk assessment, and high-frequency trading.

Applications in Finance:

Fraud Detection & Prevention
Risk Management & Credit Scoring
Algorithmic & High-Frequency Trading (HFT)
Customer Experience & Personalized Banking
C. Retail & E-commerce: Personalized Shopping & Customer Insights

Big data has transformed the retail industry by enhancing customer experience, optimizing supply chains, and driving targeted marketing campaigns.

Applications in Retail & E-commerce:

Personalized Recommendations & Customer Behaviour Analysis
Customer Sentiment Analysis & Chatbots
Supply Chain Optimization & Inventory Management
Dynamic Pricing & Promotions

D. Smart Cities: Data-Driven Urban Planning & Sustainability

Big data plays a critical role in urban planning, traffic management, energy efficiency, and public safety.

Applications in Smart Cities:

Traffic Management & Smart Transportation

Waste Management & Environmental Monitoring
Public Safety & Crime Prediction

6. Advanced Big Data Technologies & Frameworks

Big data processing has evolved significantly with the rise of new technologies and frameworks that enhance efficiency, security, and scalability. This section discusses Federated Learning, Edge Computing, and Graph Databases, which address key challenges in big data analytics.

A. Federated Learning: Privacy-Preserving Machine Learning for Big Data

Federated Learning (FL) is an advanced machine learning (ML) approach that allows multiple devices or organizations to train a shared AI model without sharing raw data. Instead of sending data to a central server, each device processes data locally and sends only model updates to a global server. This enhances data privacy and security, making it highly suitable for sensitive data applications like healthcare, finance, and mobile personalization.

B. Edge Computing: Real-Time Big Data Processing at the Source

Edge Computing is an advanced computing model where data processing happens closer to the source (e.g., IoT devices, sensors, smartphones) instead of a centralized cloud server. This reduces latency, bandwidth usage, and dependency on cloud resources, making it ideal for real-time big data applications.

C. Graph Databases: Efficient Relationship-Based Data Processing

A Graph Database is a NoSQL database designed to store and analyze complex relationships in data. Unlike traditional relational databases (SQL), which store data in tables, graph databases represent data as nodes (entities) and edges (connections).

7. INTEGRATION OF AI & BIG DATA

Artificial Intelligence (AI) and Big Data are increasingly interconnected, with AI algorithms being used to extract valuable insights from vast datasets. However, AI's adoption in big data analytics faces challenges such as lack of interpretability, scalability issues, and complexity in model selection.

This section explores three key advancements that enhance AI-driven big data analytics:

Explainable AI (XAI) – Making AI-based analytics transparent and interpretable.

Deep Learning for Big Data – Applying neural networks to analyze complex, large-scale data.

AutoML for Big Data – Automating machine learning model selection and optimization.

A. Explainable AI (XAI): Enhancing Transparency & Trust in AI Models

Explainable AI (XAI) includes techniques that enhance AI model transparency and interpretability. Traditional AI, especially deep learning, is often a “black box” due to its complexity. XAI connects AI's predictive abilities with human understanding, ensuring decisions in big data analytics remain trustworthy, fair, and accountable.

B. Deep Learning for Big Data: Scaling Neural Networks for Large-Scale Analytics

Deep Learning (DL) is a subset of machine learning that uses artificial neural networks (ANNs) to process and analyze large volumes of complex data. Unlike traditional machine learning, deep learning automatically extracts patterns and features from data without manual feature engineering.

C. AutoML for Big Data: Automating Machine Learning at Scale

AutoML (Automated Machine Learning) is an AI-driven approach that automates the process of selecting, tuning, and optimizing machine learning models for big data analytics. It allows businesses to deploy AI solutions without needing extensive data science expertise.

8. SUGGESTIONS FOR FUTURE WORK

Global data generation doubles every two years, necessitating advanced big data analysis techniques. Raw data lacks value without processing, requiring powerful computing for automation. Converting data into actionable insights is complex, involving parallel computing, AI models, and hybrid methods like fuzzy sets, neural networks, and rough sets. Feature reduction techniques filter crucial information, while handling missing values prevents data loss. Scalability, efficiency, and computational performance remain major challenges. Machine learning tools must better manage noise, imbalanced data, and inconsistencies. Research explores Quantum Machine Learning for speed, climate change predictions, and Human-Computer Interaction to enhance data visualization and decision-making.

CONCLUSION

The rapid growth of data presents challenges in analysis. This paper explores research issues, challenges, and tools for big data analytics. Platforms vary in focus, supporting batch or real-time processing. Techniques include statistical analysis, machine learning, cloud computing, and quantum computing. Future research will enhance these methods for efficient big data solutions.

REFERENCES

- [1]. M. K. Kakhani, S. Kakhani and S. R. Biradar, Research issues in big data analytics, International Journal of Application or Innovation in Engineering & Management, 2 (8) (2015), pp.228-232.

- [2]. A. Gandomi and M. Haider, Beyond the hype: Big data concepts, methods, and analytics, *International Journal of Information Management*, 35(2) (2015), pp.137-144.
- [3]. C. Lynch, Big data: How do your data grow, *Nature*, 455 (2008), pp.28-29.
- [4]. X. Jin, B. W. Wah, X. Cheng and Y. Wang, Significance and challenges of big data research, *Big Data Research*, 2(2) (2015), pp.59-64.
- [5]. R. Kitchin, Big Data, new epistemologies and paradigm shifts, *Big Data Society*, 1(1) (2014), pp.1-12.
- [6]. C. L. Philip, Q. Chen and C. Y. Zhang, Data-intensive applications, challenges, techniques and technologies: A survey on big data, *Information Sciences*, 275 (2014), pp.314-347.
- [7]. K. Kambatla, G. Kollias, V. Kumar and A. Gram, Trends in big data analytics, *Journal of Parallel and Distributed Computing*, 74(7) (2014), pp.2561-2573.
- [8]. S. Del. Rio, V. Lopez, J. M. Bentez and F. Herrera, On the use of mapreduce for imbalanced big data using random forest, *Information Sciences*, 285 (2014), pp.112-137.
- [9]. MH. Kuo, T. Sahama, A. W. Kushniruk, E. M. Borycki and D. K. Grunwell, Health big data analytics: current perspectives, challenges and potential solutions, *International Journal of Big Data Intelligence*, 1 (2014), pp.114-126.
- [10]. R. Nambiar, A. Sethi, R. Bhardwaj and R. Vargheese, A look at challenges and opportunities of big data analytics in healthcare, *IEEE International Conference on Big Data*, 2013, pp.17-22.
- [11]. Z. Huang, A fast clustering algorithm to cluster very large categorical data sets in data mining, *SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery*, 1997.
- [12]. T. K. Das and P. M. Kumar, Big data analytics: A framework for unstructured data analysis, *International Journal of Engineering and Technology*, 5(1) (2013), pp.153-156.
- [13]. T. K. Das, D. P. Acharjya and M. R. Patra, Opinion mining about a product by analyzing public tweets in twitter, *International Conference on Computer Communication and Informatics*, 2014.
- [14]. L. A. Zadeh, Fuzzy sets, *Information and Control*, 8 (1965), pp.338-353.
- [15]. Z. Pawlak, Rough sets, *International Journal of Computer Information Science*, 11 (1982), pp.341-356.
- [16]. D. Molodtsov, Soft set theory first results, *Computers and Mathematics with Applications*, 37(4/5) (1999), pp.19-31.
- [17]. J. F. Peters, Near sets. General theory about nearness of objects, *Applied Mathematical Sciences*, 1(53) (2007), pp.2609-2629.
- [18]. R. Wille, Formal concept analysis as mathematical theory of concept and concept hierarchies, *Lecture Notes in Artificial Intelligence*, 3626 (2005), pp.1-33.
- [19]. I. T. Jolliffe, *Principal Component Analysis*, Springer, New York, 2002.
- [20]. 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 Research*, 2(3) (2015), pp.87-93.
- [21]. Changwon. Y, Luis. Ramirez and Juan. Liuzzi, Big data analysis using modern statistical and machine learning methods in medicine, *International Neurology Journal*, 18 (2014), pp.50-57.
- [22]. P. Singh and B. Suri, Quality assessment of data using statistical and machine learning methods. L. C. Jain, H. S. Behera, J. K. Mandal and D. P. Mohapatra (eds.), *Computational Intelligence in Data Mining*, 2 (2014), pp. 89-97.
- [23]. A. Jacobs, The pathologies of big data, *Communications of the ACM*, 52(8) (2009), pp.36-44.
- [24]. H. Zhu, Z. Xu and Y. Huang, Research on the security technology of big data information, *International Conference on Information Technology and Management Innovation*, 2015, pp.1041-1044.
- [25]. Z. Hongjun, H. Wenning, H. Dengchao and M. Yuxing, Survey of research on information security in big data, *Congresso da sociedade Brasileira de Computacao*, 2014, pp.1-6.
- [26]. I. Merelli, H. Perez-sanchez, S. Gesing and D. D. Agostino, Managing, analysing, and integrating big data in medical bioinformatics: open problems and future perspectives, *BioMed Research International*, 2014, (2014), pp.1-13.
- [27]. N. Mishra, C. Lin and H. Chang, A cognitive adopted framework for iot big data management and knowledge discovery prospective, *International Journal of Distributed Sensor Networks*, 2015, (2015), pp. 1-13.
- [28]. X. Y. Chen and Z. G. Jin, Research on key technology and applications for internet of things, *Physics Procedia*, 33, (2012), pp. 561-566.
- [29]. M. D. Assuno, R. N. Calheiros, S. Bianchi, M. a. S. Netto and R. Buyya, Big data computing and clouds: Trends and future directions, *Journal of Parallel and Distributed Computing*, 79 (2015), pp.3-15.
- [30]. I. A. T. Hashem, I. Yaqoob, N. Badrul Anuar, S. Mokhtar, A. Gani and S. Ullah Khan, The rise of big data on cloud computing: Review and open research issues, *Information Systems*, 47 (2014), pp. 98-115.
- [31]. M. Herland, T. M. Khoshgoftaar and R. Wald, A review of data mining using big data in health informatics, *Journal of Big Data*, 1(2) (2014), pp. 1-35.
- [32]. T. Huang, L. Lan, X. Fang, P. An, J. Min and F. Wang, Promises and challenges of big data computing in health sciences, *Big Data Research*, 2(1) (2015), pp. 2-11.
- [33]. G. Ingersoll, introducing apache mahout: Scalable, commercial friendly machine learning for building intelligent applications, *White Paper, IBM Developer Works*, (2009), pp. 1-18.
- [34]. H. Li, G. Fox and J. Qiu, Performance model for parallel matrix multiplication with dryad: Dataflow graph runtime, *Second International Conference on Cloud and Green Computing*, 2012, pp.675-683.
- [35]. D. P. Acharjya, S. Dehuri and S. Sanyal, *Computational Intelligence for Big Data Analysis*, Springer International Publishing AG, Switzerland, USA, ISBN 978-3-319-16597-4, 2015.S