Kurdish Studies
July 2022
Volume: 10, No: 2, pp 543-547
ISSN: 2051-4883 (Print) | ISSN 2051-4891 (Online)
www.KurdishStudies.net

DOI: 10.53555/ks.v10i2.3447

Big Data and Cloud Computing: An In-Depth Literature Analysis

Dr. Sandip Banerjee [1*], Anant Mittal[2], Hinia Jeram[3]

[1*]Assistant Programmer, Department of Election, Govt. of Arunachal Pradesh
[2]IPS (2015), Superintendent of Police, Special Investigation Cell (Vigilance), Govt of Arunachal Pradesh
[3]Assistant System Analyst, Department of Election, Govt. of Arunachal Pradesh

***Corresponding author: Dr. Sandip Banerjee**
*Assistant Programmer, Department of Election, Govt. of Arunachal Pradesh

1. OVERVIEW

Big data encompasses methods for analyzing, extracting information from, and managing datasets that are too massive or intricate for conventional data processing software. As its name implies, big data refers to an enormous volume of information. The typical characteristics of big data can be described using the 4V's. Cloud computing is essentially the on-demand availability of computer system resources, particularly data storage and processing power. It generally allows users to access, utilize, work on, and modify their projects while collaborating with colleagues. While cloud computing offers flexibility in work schedules, big data provides valuable insights and information. The analytics process involves characteristic analysis, storage management and cloud, big data processing, and ultimately deriving knowledge from the vast available data. In today's world, digital security is paramount. For big data, security is crucial, as the information often includes confidential data, secret keywords, and passwords, which could have severe consequences if compromised. Therefore, security is a critical consideration in big data and cloud computing. Various methods can be employed to achieve security, such as node authentication, encryption, access control, and honeypot nodes. Implementing this system may face challenges related to data storage, speed, security, processing, transmission, visualization, architecture, integration, and quality. The combination of cloud computing and big data has applications across numerous fields, including management and finance.

2. BIG DATA ANALYTICS

Cloud-based Big Data refers to massive datasets, often measuring dozens of terabytes or petabytes, which are challenging to manage using conventional local computer-based Database Management Systems. The cloud offers an ideal solution for the complex and expensive tasks of scaling storage, visualizing data, and handling data management and capture. Many global industry leaders now store their entire data repositories in the cloud. These organizations can leverage built-in cloud features or deploy custom functionality to explore vast amounts of detailed information, uncovering previously unknown insights. The cloud's near real-time capabilities provide significant advantages for businesses working with large datasets. Consequently, cloud systems require specialized data architectures, analytical methodologies, and tools to accommodate these needs effectively.

A. Attributes of Big Data –

Big data's key characteristics are categorized into four V's: Volume, Variety, Velocity, and Veracity. Volume, the most prominent and frequently considered aspect, denotes the magnitude or size of the data. Velocity indicates the speed at which information is gathered or evolves. Certain types of big data, such as stock market prices, are monitored and collected at extremely high rates, sometimes as frequently as every second. Variety, the third feature, describes the diverse sources from which data originates, including logs, social media platforms, and click streams. The final characteristic, Veracity, assesses the quality of the data. This is measured by examining patterns of inconsistency, incompleteness, approximation, deception, ambiguity, or latency within the dataset.

B. Storage Governance in Cloud Computing –

Various cloud-based software packages are available for cloud computing. Organizations can utilize enterprise data warehouses or, in cases involving unstructured data like extensive texts, NoSQL solutions. The most commonly used tools include Hadoop, Spark, MapReduce, and HBase. Hadoop, a widely accessible programming framework written in Java, excels at processing vast amounts of data. It enables the analysis of large datasets using server clusters, with thousands of nodes potentially running the application. Hadoop's framework mitigates the risk of system failure even when multiple nodes malfunction, offering a flexible and fault-tolerant computing solution. The Hadoop Distributed File System (HDFS) provides an efficient, high-tech file system that spans all nodes in a Hadoop cluster for data storage, linking local node file systems to significantly enhance reliability.

HBase, a NoSQL software within the Hadoop HDFS framework, is an open-source Java-based database modeled after Google's BigTable. It is widely employed for storing Big Data, particularly unstructured data. Spark, another open-source tool, serves as a unified analytics engine for large-scale data processing, offering an interface for programming entire clusters with implicit data parallelism and fault tolerance.

For cloud-based data storage, Azure HDInsight and Amazon EMR are the most popular services. These highly effective cloud-native services enable machine learning capabilities and allow for the creation of Hadoop, Spark, MapReduce, and HBase clusters. Cloud storage facilitates scaling workloads up and down to accommodate high-velocity data. These services enable the creation of data pipelines to optimize work, and by generating clusters on demand and paying only for what is used, costs can be significantly reduced.
Cloud platforms also provide various other open-source clusters such as Kafka, Apache Interactive Query, and Apache Storm to meet specific customer needs. They support multiple programming languages for data operations, including Java, .NET, Python, Go, Scala, and Clojure. Additionally, cloud services offer built-in powerful visualization tools, eliminating the need to purchase separate software. Azure HDInsight comes with Microsoft Power BI, while AWS is compatible with Tableau.

C. Big Data Operations –
Processing requires four essential components:
1. The capability to swiftly import data.
2. Rapid query execution.
3. Optimal use of storage resources.
4. Robust flexibility in handling dynamic workloads.

To effectively meet these requirements, cloud service providers offer Map Reduce Software, with both Azure HDInsight and Amazon EWS supplying Map Reduce frameworks. This framework significantly enhances processing through its parallel programming model. Instead of boosting a single server's storage or computational capacity, Map Reduce expands horizontally by adding more servers and computers, embracing a "scale out" rather than "scale up" approach.
Map Reduce divides tasks into stages for parallel execution, improving efficiency. Its operation is straightforward: the "Map" function assigns key-value pairs to smaller tasks. For instance, when dealing with unstructured text data, a word could serve as the key, with its frequency as the value. The "Reduce" function then aggregates and merges the "Map" output, combining values with identical keys to produce the final computational result.
This approach is particularly beneficial when combined with cloud architecture's speed, resulting in unparalleled performance compared to typical local computers. Such high processing speeds enable real-time data analysis and output generation. When implemented in the cloud, this system proves highly advantageous for handling Big Data with high velocity and volume, benefiting entities like companies and stock exchanges such as NASDAQ, BSE, and NSE. Cloud-based solutions offer more efficient storage, analytics, and processing at lower costs than traditional computing systems.

3. Key Challenges
Despite the numerous benefits of integrating cloud computing and big data, several challenges and risks must be considered when implementing big data in a cloud environment.

A. Data Storage Systems –
Technological advancements have led to an unprecedented increase in data generation. However, much of this data is discarded or erased due to insufficient storage capacity. Consequently, the primary obstacle for Big Data analysis is the development of adequate storage media and improved transmission rates. Current storage technologies lack the necessary capabilities to process Big Data effectively. Storing information on conventional physical storage systems is problematic, as hard disk drives are prone to failure, and traditional data protection methods are inadequate. Furthermore, Big Data's velocity requires storage systems that can rapidly scale up when needed, which is challenging to achieve with these conventional systems. The exponential growth of data has significantly increased data mining tasks, resulting in greater data diversity. There is a pressing need to focus on designing storage systems and developing efficient data analysis tools that can provide output guarantees, given that data is collected from various sources. Additionally, machine learning algorithms can be developed for data analysis, enhancing efficiency and scalability. Cloud storage services (such as Amazon S3 and Elastic Block Store) offer unlimited storage and high fault tolerance, providing solutions to Big Data storage challenges. However, hosting and transferring Big Data on the cloud is costly due to its enormous size.

B. Data Transfer –
Another challenge involves the transfer of massive amounts of big data (for example, hundreds of terabytes) to a public cloud within a short timeframe. This raises concerns about storage, reliability, privacy, and security. The transfer of enormous data volumes during different stages of the data life cycle presents challenges at each stage. As a result, it is necessary to develop intelligent pre-processing techniques and data compression algorithms to effectively reduce data size before transfer. To move data from local data centers to cloud platforms, efficient algorithms must be created that automatically recommend the most suitable cloud service (location) based on geo-temporal principles (as data can be situated in various locations) to maximize data transfer speed while minimizing costs.

C. Algorithmic Complexity –
Handling vast amounts of data necessitates substantial computing resources, typically addressed through enhanced storage, network, and CPU speeds. However, conventional computing systems often fall short in providing adequate processing power. Cloud computing offers a partial remedy with its seemingly limitless, on-demand processing capabilities. Yet, this shift introduces new challenges. Firstly, the restricted network bandwidth in cloud environments impacts the efficiency of large-scale data processing. Secondly, the geographical dispersion of data complicates its collection for preprocessing. Key cloud

computing features like virtualization, pooled data resources, and high computational power make it challenging to monitor and ensure data locality, hindering its ability to support data processing that involves intensive communication and information exchange.

D. Secure Data Management –
The convergence of Big Data and Cloud Computing gives rise to certain security vulnerabilities. Furthermore, existing data security policies and schemes, designed for structured data in traditional DBMS, prove ineffective for highly diverse and unstructured data. Consequently, there's a need to develop robust policies for data access control and safety management to accommodate new data management systems and storage structures. In the cloud era, safeguarding data confidentiality, integrity, and availability becomes crucial, given the limited control data owners have over their information and various resources. Heterogeneity stands out as a well-known security vulnerability in Big Data's cloud implementation. Often, Big Data deployment requires a new cloud platform, necessitating the development of novel security tools as existing practices may not suffice. These tools should encompass encryption, authentication, intrusion detection, access control, monitoring, and event logging. When integrating Big Data into cloud environments, it's essential to consider consolidation plans alongside security policies.

E. Personal Data Safeguarding –
Many researchers have observed that the cultural challenge of cloud computing and big data primarily revolves around privacy concerns. A significant portion of big data sources comprises documents, messages, images, audio and video posts, as well as highly sensitive information. This includes individual location data, behavioral patterns, transactions, and even companies tracking employee movements and productivity. These data points are often digitally recorded via social media, indicating that social platforms serve as a primary resource for big data. Consequently, accessing users' private information poses a major risk in this context.

F. Contrasting Conceptual Frameworks of these Fields –
Cloud Computing is founded on the principles of consolidation and resource sharing, while big data systems (e.g., Hadoop) are built on the concept of shared-nothing architecture, where each node operates independently. The integration of big data with cloud computing technologies offers promising future directions for businesses and educational institutions. Cloud computing's capacity to store vast amounts of diverse data and process it rapidly can yield insights that drive swift advancement in education and business sectors. However, significant concerns about security and privacy in cloud environments remain the primary obstacle preventing educational and commercial entities from adopting cloud solutions.

4. Technological Systems
Big Data analytics as a service in the cloud is offered through three main service models: infrastructure as a service (IaaS), platform as a service (PaaS), and software as a service (SaaS).

A. Infrastructure as a Service (IaaS) –
IaaS can be implemented on-premises or via cloud providers to enable organizations to allocate or purchase time on shared server resources (often virtualized) for managing Big Data analytics' computational and storage requirements. Cloud providers handle the management of high-performance networks, servers, and storage resources.

Organizations involved in Big Data need not maintain the hardware and software necessary for such performance. Hadoop, an open-source solution, utilizes distributed data storage and processing.

Technologies employed for IaaS purposes include the Hadoop framework and NoSQL databases such as MongoDB, Apache Cassandra, or Couchbase. IaaS solution providers include Amazon Web Services, Windows Azure, Citrix CloudPlatform, Microsoft System Centre, OpenStack software, and Rackspace.

B. Platform as a Service (PaaS) –
PaaS is utilized to provide advanced programming models and database systems. It offers tools and libraries for developing, testing, deploying, and running applications in cloud environments.

Amazon ElasticMapReduce provides a basic Hadoop framework PaaS environment. Windows Azure's HDInsight data service brings Hadoop to the cloud, coupled with Microsoft BI tools like Power Map and Power View. AWS offers common PaaS solutions, including DynamoDB for NoSQL database services and Redshift for data warehousing. Google also provides PaaS capabilities such as Bigtable and BigQuery.

C. Software as a Service (SaaS) –
SaaS is used to deliver applications via the internet. It includes offerings like jKool, which provides cloud-based business solutions and real-time analysis of time-sensitive information. Concur, a rapidly growing TE SaaS company, operates a single instance of its software containing preferences and history for millions of global business travelers, covering airlines, hotels, car rentals, and taxi services. Karmasphere offers a pay-as-you-go application that analyzes data stored in Amazon S3 using Amazon Elastic Map Reduce.

5. Safeguarding

A. Importance of security in Big Data:

Many businesses utilize big data without adequate security measures. Security breaches in big data can lead to severe consequences. Companies typically employ this technology to store zettabytes of corporate information, potentially resulting in critical data classification issues. To protect data, encryption, logging, or honeypot techniques are necessary. Big data analysis methods must be used to address the challenge of detecting attacks and intruders. When searching databases in big data, speed is crucial. However, the process can be cumbersome due to the inability to quickly traverse all relevant data in the entire database. As big data becomes more complex, its indices focus on simpler data types. Traditional sequential algorithms are ineffective for big data processing.

B. Security challenges in cloud computing:

Security issues in cloud computing environments can be divided into four categories:

1. Network level: These concerns involve network protocols and security, including distributed data, inter-node communication, and distributed nodes.
2. User authentication level: This category encompasses various encryption/decryption techniques, authentication methods for applications and nodes, logging, and administrative rights for nodes.
3. Data level: These challenges relate to data availability, protection, and distribution.
4. General types: This category includes traditional security tools and the use of various technologies.

C. Strategies for addressing security concerns:

1. Data encryption: Given that information in computers is typically stored in clusters, unauthorized access poses a significant risk to organizations. To mitigate this threat, implementing data encryption is crucial. Various encryption methods can be employed for different systems, with encryption keys securely stored behind firewalls. This approach ensures the protection of customer data.
2. Node verification: Whenever a node joins a cluster, it must undergo authentication. Any node identified as potentially malicious should be denied authentication.
3. Deceptive nodes: These nodes, known as honeypots, are designed to appear as regular nodes but function as traps. They automatically ensnare hackers, preventing harm to the system or its data.
4. Access regulation: Implementing diverse privacy and access control measures in distributed environments serves as an effective security measure. To prevent information leakage, the Linux operating system is often utilized. Linux features a mechanism that supports access control security policies through Linux Security modules in its kernels.

D. Methods for tackling security issues:

Cloud computing facilitates remote data storage, optimizing resource utilization. Consequently, safeguarding this data and restricting access to authorized individuals is paramount. This necessitates secure third-party data publication for both data outsourcing and external publications. In cloud computing, the machine acts as a third-party publisher, storing sensitive information in the cloud. The data requires protection, and the aforementioned techniques must be employed to ensure timely maintenance of authenticity and completeness.

6. Strengths:

The integration of cloud computing and big data offers numerous benefits. Big data requires multiple servers to handle its vast volume, high velocity, and variability. These servers operate concurrently to meet big data's demands. Cloud computing, which already utilizes multiple servers and provides resource allocation, serves as an ideal foundation for big data. Building big data on cloud multi-servers and leveraging the resource allocation capabilities of cloud environments enhances the efficiency of big data analysis. Utilizing cloud infrastructure as a storage system for big data improves the performance of both technologies.

Cloud services, built on remote multi-servers, can process enormous amounts of data simultaneously, enabling big data to manage large volumes using advanced analytics. The convergence of cloud computing and big data results in cost reductions. Instead of investing in new servers and volumes for big data, cloud computing systems can serve as the foundation, offering greater flexibility and scalability while eliminating substantial investments in big data hardware. Additionally, cloud computing provides faster provisioning for big data, as setting up servers in the cloud is straightforward and feasible. This allows the cloud environment to be scaled according to big data processing requirements, which is crucial given that data value diminishes rapidly over time.

Cloud computing complements big data by providing a simple, on-demand, and shared computing platform with minimal management effort and overhead. It enhances robustness, automation, and multi-tenancy in the environment. Big data enables end users to visualize data and helps companies explore new market opportunities. Data analytics, a significant advantage of big data, allows individuals to personalize information and interact with real-time websites. The convergence of cloud computing and big data makes big data resources more manageable, monitorable, and reportable. This integration also reduces complexity and increases efficiency. Due to these advantages, cloud-based approaches are considered the optimal models for deploying big data.

7. Operations

Large-scale distributed applications that handle extensive data sets are known as big data applications. As data exploration and analysis have become increasingly challenging across various industries, traditional data processing methods have proven

inadequate for managing the computational demands of large and complex data. This has led to the development of big data applications, with Google's map reduce framework and Apache Hadoop serving as key software systems. These applications generate substantial amounts of intermediate data and are primarily utilized in manufacturing and bioinformatics sectors.

In the manufacturing industry, big data provides a transparent infrastructure that addresses uncertainties such as inconsistency, component performance, and availability. Predictive manufacturing in big data applications begins with data acquisition, collecting various types of sensory information including pressure, vibration, acoustics, voltage, current, and controller data. Manufacturing big data is created by merging sensory data with historical information.

Big data systems employ parallel distributed computing, combining computer clusters and web interfaces in cloud computing. Software packages for big data offer a wide array of tools and options, enabling individuals to map the entire data landscape across an organization. This allows for the evaluation of internal risks and is considered a significant advantage, as big data enhances data security. It helps identify potentially sensitive information that may not be adequately protected and ensures compliance with regulatory requirements during processing.

The combination of big data and predictive analytics presents challenges for numerous industries, focusing on four key areas: risk calculation for large portfolios, financial fraud detection and prevention, improvement of delinquent collections, and execution of high-value marketing campaigns.

Organizations can leverage big data to create new products and services, enhance existing offerings, and even develop novel business models. Big data analytics can be applied to various fields, including consumer intelligence, supply chain intelligence, performance quality, risk management, and fraud detection. In risk management, sectors such as investment banking, retail banking, and insurance can benefit from big data analytics. It can assist in investment selection by analyzing the probability of gains versus losses, a crucial aspect of the financial services industry. Additionally, internal and external big data can be evaluated for comprehensive and dynamic assessment of risk exposures.

Fraud detection and prevention can be enhanced through the use of big data analytics, particularly in government agencies, banks, and insurance companies. While traditional analytics continue to play a significant role in automated fraud detection, there is a growing trend across various sectors and organizations to leverage big data for improving their systems. By utilizing big data, these entities can conduct faster analyses and cross-reference electronic information from diverse public and private sources. Several industries stand to gain from the implementation of big data analytics, including manufacturing, retail, central government, healthcare, telecommunications, and banking.

8. Final Thoughts

In today's digital landscape, big data and cloud computing have become increasingly significant. The integration of big data with cloud computing shows considerable promise for future advancements. When utilizing Software as a Service, big data is particularly crucial in providing insights for cloud-based applications. The combination of big data and cloud computing has diverse applications across various sectors. These include enhanced analytical capabilities due to the vast amount of data available, the development of cost-effective and efficient infrastructure over time, and improved security, integrity, and availability of cloud platforms. Furthermore, this integration enables businesses and platforms to expand through the utilization of big data resources.

REFERENCES

1. Alsghaier, Hiba & Al-Shawakfa, Emad. (2018). An empirical study of cloud computing and big data analytics. International Journal of Innovative Computing and Applications. 9. 180. 10.1504/IJICA.2018.10014870.
2. Yadav S., Sohal A. (2017) "Review Paper on Big Data Analytics in Cloud Computing" in International Journal of Computer Trends and Technology (IJCTT) V49(3):156-160, July 2017. ISSN:2231-2803.
3. Hariharan, U. & Kotteswaran, Rajkumar & Pathak, Nilotpal. (2020). The Convergence of IoT with Big Data and Cloud Computing. 10.1201/9781003054115-1.
4. Agrawal, Divyakant & Das, Sudipto & Abbadi, Amr. (2011). Big Data and Cloud Computing: Current State and Future Opportunities. ACM International Conference Proceeding Series. 530-533. 10.1145/1951365.1951432.