Indonesian Journal of Computer Science
ISSN 2549-7286 (*online*)
Jln. Khatib Sulaiman Dalam No. 1, Padang, Indonesia
Website: ijcs.stmikindonesia.ac.id | E-mail: ijcs@stmikindonesia.ac.id

Distributed Architectures for Big Data Analytics in Cloud Computing: A Review of Data-Intensive Computing Paradigm

Chiai Al-Atroshi[1], Subhi R.M. Zeebaree[2]
chiai.mohammed@auas.edu.krd, subhi.rafeeq@dpu.edu.krd
[1]Information Technology Department, Technical College of Informatics-Akre, Akre University for Applied Sciences, Duhok, Iraq
[2]Energy Eng. Department, Technical College of Engineering, Duhok Polytechnic university, Duhok, Iraq

Article Information

Submitted : 6 Mar 2024
Reviewed: 13 Mar 2024
Accepted : 8 Apr 2024

Keywords

Big Data, Cloud computing, Data Intensive Clouds

Abstract

Big Data challenges are prevalent in various fields, including economics, business, public administration, national security, and scientific research. While it offers opportunities for productivity and scientific breakthroughs, it also presents challenges in data capture, storage, analysis, and visualization. This paper provides a comprehensive overview of Big Data applications, opportunities, challenges, and current techniques and technologies to address these issues. This study presents a system for managing big data resources using cloud for the development of data-intensive applications. It addresses even the challenges related to technologies that combine cloud computing with other allied technologies and devices. In addition, the increasing volume, velocity, and variety of data in the era of Big Data necessitate advanced methods for data processing and management. This study delves into the intricacies of data scalability, real-time processing, and the integration of diverse data types. Furthermore, it explores the role of machine learning algorithms and artificial intelligence in extracting meaningful insights from massive datasets.

A. Introduction

The increasing significance of data-intensive applications has led to various approaches for data distribution, management, and processing, characterized by tools, frameworks, and architectures using common abstractions. The Internet's widespread popularity has significantly increased data generation and computation, presenting immense potential for data utilization and analysis across various users and applications. However, data-related challenges have emerged, such as incorporating user-clicks and geographical location for improved relevance in search engines and user-related search results. Data-intensive computing [1][2] is revolutionizing informatics and researchers' methods, from hardware and algorithms to knowledge presentation. Applications in various disciplines shift focus from large datasets to real-time processing of massive data streams, requiring efficient data-handling capacity.

The era of Big Data is entering, affecting various fields like healthcare, public sector administration, retail, global manufacturing, and personal location data, according to a McKinsey report.

Distributed architectures for big data analytics in cloud computing include Hadoop, Spark, Apache Flink, Kafka, NoSQL databases like HBase, Cassandra, and MongoDB, software containers like Docker, Kubernetes container orchestration technology, serverless computing, cloud storage like Azure Blob Storage or S3, and managed analytics services like Databricks for big data processing and AWS EMR. These platforms offer scalable and fault-tolerant distributed storage, resource management, task scheduling, streaming, MLlib, RDDs, and Spark SQL. These technologies help handle massive amounts of data and provide efficient data processing and storage solutions [3][4].

In the big data era, application execution and simulation become resource-intensive [1][5][6] due to increased data creation and the challenges posed by IoT, Big Data, and Industry 4.0 disciplines. Cluster computing, grid computing, cloud computing, fog computing, and dew computing are the distributed computing concepts that have defined and sparked the computing revolution (Fig. 1). These contemporary computing paradigms have the capacity to handle increasing data volumes and offer processing and storage resources.for applications that require a lot of data.

The methods used to draw conclusions from enormous, intricate databases are referred to as big data analytics. Distributed architectures are necessary for efficient storage and processing of big data due to its volume, velocity, and variety. Cloud computing offers on-demand access to servers, networking, and storage—all of which are perfect for big data analytics [7].

Cloud infrastructure can be used to deploy popular distributed big data frameworks like Spark, Hadoop, and Flink. Hadoop uses HDFS and MapReduce programming to enable distributed processing of huge files across clusters. In-memory cluster computing distributed is also possible with Spark. Stream processing is Flink's primary focus.[8]

Scalable and resilient storage across commodity servers is offered by cloud-native NoSQL databases such as Cassandra and HBase. Large datasets can be stored using cloud object stores like S3. Docker and other containerisation platforms make it easier to deploy distributed applications on cloud servers.

The process of building big data clusters is automated by managed analytics cloud services such as AWS EMR, Azure HD Insight, and Google Dataproc. Analysis programmes can be done in a completely managed manner using serverless platforms like as AWS Lambda.

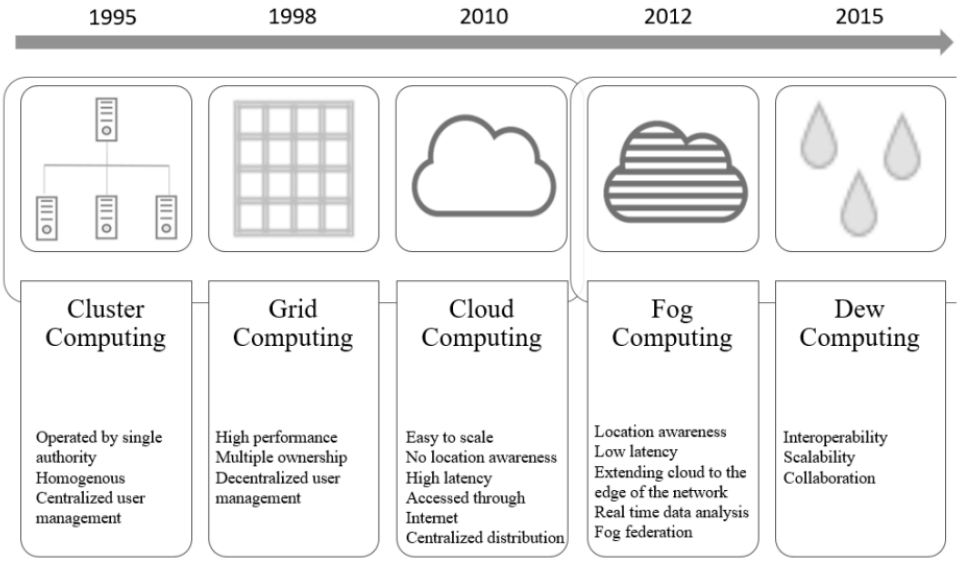

Figure 1. Timeline of paradigms in Distributed Computing

B. Related Work

2.1 Big Data Analytics

Big data refers to very large, complex datasets that traditional data processing tools cannot easily handle. The data is often unstructured or semi-structured. Big data analytics refers to the processes and techniques used to analyze and extract insights from big data. The goal is to uncover patterns, trends; correlations, preferences, and other useful business information that can help organizations make more informed decisions. Common techniques include data mining, machine learning, predictive analytics, text mining, social media analytics, network analysis, natural language processing, and spatial analysis. Big data analytics typically involves using specialized tools and frameworks like Hadoop, Spark, NoSQL databases, stream processing platforms, and advanced analytics software. The ability to analyze big data allows businesses to optimize operations, improve customer service, develop new products, identify market trends, gain competitive insights, and more. Challenges include data quality [9][10][11] issues, integrating disparate data sources, lack of data science skills, ensuring adequate data security and governance, and determining return on investment. Adoption continues to grow as big data platforms and analytical techniques mature, and more organizations realize the business value of big data analytics for data-driven decision making. In summary, big data analytics extracts valuable business insights from huge volumes of heterogeneous data using specialized techniques and technologies [12][13]. It brings data-driven innovation but requires careful data management and governance.

Big data analytics involves analyzing large, complex datasets to uncover insights. This requires substantial computing resources. Cloud computing provides convenient, on-demand access to computing resources via the internet. This makes it well-suited for big data analytics. Cloud computing provides convenient, on-demand access to computing resources via the internet. This makes it well-suited for big data analytics. Distributed architectures divide data processing across multiple nodes/servers. This allows for parallel processing, improving speed and scalability for big data workloads. Distributed architectures allow big data analytics to leverage the on-demand scalability of cloud computing for faster, more advanced analytics on ever-growing data volumes. Popular distributed data processing frameworks like Hadoop MapReduce, Spark, and Flink are commonly used for big data analytics in the cloud. These frameworks provide capabilities like distributed storage, distributed processing, fault tolerance, and more to handle big data workloads. Cloud services like Amazon Elastic MapReduce allow these frameworks to be deployed on cloud infrastructure in a managed way. Key challenges include handling data transfer bottlenecks, latency, and coordination between processing nodes. In summary, distributed architectures enable big data analytics to scale in the cloud, providing faster insights from large datasets but requiring careful design to overcome challenges like latency. [14][15]

2.2 Cloud computing

Cloud computing is a crucial IT paradigm that distributes resources and services on computers and devices, accessible through the internet infrastructure, without end-user ownership. Cloud computing provides convenient, on-demand access to computing resources - storage, networking, servers, analytics tools etc. This makes it well-suited for big data analytics. The cloud enables distributing data processing across virtual clusters for parallel computing on large datasets.

2.3 Problems in Big Data for Data Intensive Applications

Data-intensive computing [5][16][17][18] involves parallel computing applications processing large volumes of data, typically terabytes or petabytes in size. These applications are referred to as compute-intensive, focusing on computational requirements, while data-intensive applications handle large volumes of data in multistep analytical pipelines, reducing data analysis cycles and developing new algorithms for efficient data processing [19][20].

Data-intensive computing platforms use a parallel computing approach, combining multiple processors and discs in large commodity computing clusters connected by high-speed communications networks.[19][20][21] This method allows data to be divided among available resources and processed independently, resulting in performance and scalability based on data volume. This method is suitable for data-intensive computing and "embarrassingly parallel" problems, adapting to clusters, data grids, and cloud computing.

2.4. Existing Solutions to Data Intensive Application Development

Large datasets are processed on commodity machinery by means of distributed frameworks such as Hadoop, Spark, and Flink. Managed services like object storage, data warehousing, batch and stream processing, and data pipelines are provided by cloud providers including AWS, GCP, and Azure. Cassandra, MongoDB, and Redis are examples of contemporary distributed databases that have been optimized for high write throughput and scalability. Data workflows can be created, tracked, and managed with the use of programme's like Apache Airflow, Prefect, and Dagster[22][23][24].

Big Query and Snowflake are powerful tools for managing data warehouses, providing petabyte-scale analytical data storage and near-real-time querying. These tools enable the storage of large amounts of unstructured data, streaming data, and deploying big data applications across clusters. They also offer managed containers, analytics services, machine learning platforms, and data marketplaces for monetizing data and enabling organizations to share or sell it.[25]

2.5 A Review on Data Intensive Clouds

Social networks and geospatial processing are crucial tools in various industries, including ride hailing apps, e-commerce sites, digital advertising, fraud detection, IoT applications, search engines, cloud-native apps, and genomics [20][21][22][26]. These platforms use data streams, geospatial processing, and predictive modeling to match riders with nearby drivers in real-time. These technologies enable businesses to process orders, handle recommendations, and manage supply chain and logistics, ensuring efficient operations and reducing fraud.

2.5.1 Requirements and Expectations of Data Intensive Clouds

Data-intensive clouds offer high throughput and scalability to handle spikes in data volumes without downtime. They provide flexibility for analyzing various types of data, low latency, and support high velocity data flows. These clouds are durable, secure, and easy to use. They offer broad APIs, portability, cost optimization, compliance support, monitoring, metadata management, and developer productivity tools. They also support multi-cloud and hybrid environments, ensuring data privacy and compliance [22][23][24].

2.5.2 Security Aspects

Encrypt data is a crucial aspect of data-intensive applications, ensuring its security and integrity. Encryption, both at rest and in transit, is essential to protect data from unauthorized access. Network security, including VLANs, access control lists, and firewalls, is crucial for restricting network access to data systems. Identity and access management, including strong user authentication and session management, is essential for managing data access. Logging and monitoring are crucial for real-time monitoring and alerts for malicious activities. Regular vulnerability scanning and patching of security bugs are essential for maintaining data security. Hardening systems by removing unnecessary software and closing unused ports/services on servers is also essential. Data leak prevention (DLP)

tools like firewall-based data leak prevention can prevent accidental data leaks. Compliance controls, such as masking, are necessary to meet HIPAA, PCI DSS, and GDPR regulations. Incident response plans are essential for detecting and containing security incidents. Data validation ensures data integrity, and backups are maintained to prevent ransomware loss. Code security involves security reviews, static analysis, and dependency checking for apps handling sensitive data [27].

2.5.3 Classification of Technologies

Data-intensive applications utilize various technologies, including relational databases, NoSQL databases, data warehouses, data lakes, distributed file systems, and object storage. Data processing and analysis involve batch processing, stream processing, query engines, and data pipeline orchestration. Infrastructure and deployment include cloud computing, container orchestration, and infrastructure as code. Data integration and movement involve ETL tools, change data capture, message queues, and data virtualization. Analytics and machine learning involve business intelligence, machine learning libraries, and MLOps. Monitoring and metadata include data catalogs, data quality, logging and metrics, and tracing. The choice of technology depends on specific use cases, data types, infrastructure, and architecture patterns. Most real-world systems use a combination of these technologies. The choice of technology depends on the specific use cases, data types, infrastructure, and architecture patterns.[28]

C. Literature Review

As opined by Shamsi et al. [1] Cloud computing is increasingly essential for data-intensive applications like scientific computing, machine learning, and big data analytics. It offers pay-as-you-go pricing, quick elasticity, and hardware management outsourcing.

The article by Graetsch et al. [5] discusses the challenges of constructing and implementing data-intensive software systems due to the growing volume and complexity of data. The article also discusses the impact of data on software development, infrastructure costs, talent needs, security, scalability, and performance. Raymond et al. [16] proposes a machine learning-based automated method for data-intensive applications, addressing the challenges of requirements engineering. It uses natural language processing to extract features from text and uses a Random Forest classifier to sort needs into functional and non-functional categories, including analytics, reporting, and storage. This method automatically tags and retrieves relevant requirements during analysis, facilitating effect analysis. Tests show the method performs well, has higher recall than keyword-based search, lowers errors, and improves reuse of requirements. Loncar et al. [17] explores four-dimensional large data analysis computational paradigms, including batch processing frameworks like Spark, Dryad, and MapReduce, online processing paradigms like Storm, S4, and Samza, and paradigms like machine learning, graph processing, SQL-like analysis, and hybrid systems. The works by Jha et al [18] discusses probabilistic databases, statistical methodologies, and trust analysis for

data quality and validity. Kouzes et al [31] emphasizes the progression of computing paradigms from centralized to distributed to decentralized, identifying open problems in real-time analytics, effective joins, data quality, and user-friendliness. Wu et al. [22] covers various data challenges, including storage, real-time processing, noise, and rules. It recommends using approaches like data pipelines, NoSQL databases, horizontal scaling, data virtualization, and governance frameworks. It emphasizes the importance of early data issues in requirements, design, and architecture phases. Fernandez et al, [23] in the paper discusses the security challenges in data-intensive computing, including authentication, access control, data integrity, confidentiality, privacy, and availability. It explores risks such as denial-of-service attacks, malicious attacks, and data leaks. The article discusses security measures in big data platforms like Hadoop, Spark, and NoSQL, including firewalls, VPNs, TLS/SSL, honeypots, auditing, encryption, and authorization. It also discusses new directions like homomorphic encryption, trustworthy computing, and hardware security modules. The article recommends a defense-in-depth approach at the infrastructure, platform, and application layers. The study in [24] by Al-Jumaili et al, explores the use of cloud computing and big data analytics in power system planning and management. It highlights the benefits of big data analytics in risk analysis, asset management, forecasting, and intelligent control. Rao et al [29] thoroughly incestigated about the challenges include lack of standards, real-time performance, cloud security issues, and data quality. Future directions include investing in fog/edge computing solutions, open architectures, and deep learning, blockchain, and quantum computing. Dave et al. [19] discusses the tradeoffs of distributed vs centralized analysis, batch vs real-time, complexity vs performance, and key challenges in robust big data analysis. Klepmann et al [20] propounds achieving cloud computing's promises presents technological and policy obstacles. Factors such as data storage, privacy, security, heterogeneity, latency limitations, and resource management complexity pose significant challenges. The article by Abdalla et al. [21] disserts about the technical architecture of big data pipelines, including data sources, storage, processing, analysis, and visualization. Key technologies studied include Hadoop, NoSQL databases, MapReduce, Storm, and Spark. Machine learning methods are explored, and practical issues like scalability, timeliness, privacy, and skill shortage are examined. However, possible solutions include compression, deduplication, data localization, containerization, decentralised computing, automated resource optimization, and different service models. Ortega et al [22] discusses the challenges and opportunities in developing analytics techniques for data-intensive Internet of Things and sensor applications. It highlights the need for advanced methods like data cleaning, compression, and feature extraction, as well as the use of windowing techniques and online learning methodologies for real-time stream management. It also highlights the need for research in location-aware analytics, predictive maintenance, benchmarking, and data quality. The cloud allows for easy sharing of analytics across organizations and combining data from multiple platforms [30]. Combining public, private, and hybrid cloud deployments optimizes cost, size, control, and performance. Self-service access simplifies implementation for businesses without specialized infrastructure. However, cloud-

based big data pipelines must consider variables like availability, multi-tenancy, vendor lock-in, and data transportation costs.

Qualitative research is a systematic investigation into how individuals and groups behave[31], how organisations function, and how interactions shape relationships in a natural setting [32]. Unlike quantitative research, qualitative research uses textual data. Sometimes quantitative methods like surveys cannot answer the research question. Numerical data collection can measure patterns like consumer shopping behaviour, according to Busetto et al. [33]. However, qualitative methods are better for finding hidden patterns.

A qualitative study is a research method that offers flexibility, openness, and responsiveness to the context, allowing researchers to explore unexplored research topics. Through an in-depth review of existing literature, the researcher can identify relevant issues and conduct a thorough exploration of the research topic.

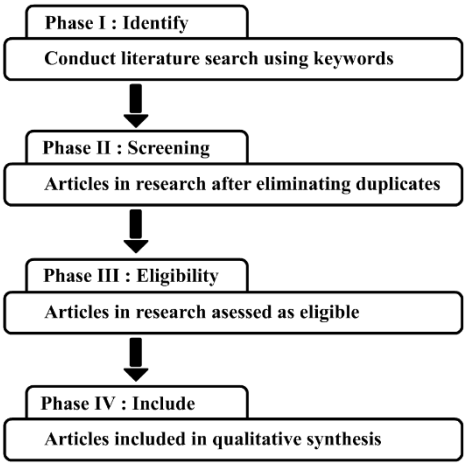

Figure 2. PRISMA flow diagram for qualitative systematic literature review

The outcomes are presented based on criteria such as published literatures in various databases, progress in publications over time beginning in the year 2015. In terms of publications that prevail with the concepts of cloud, big data and data intensive applications, articles chosen from various publications include; Hindawi 16, MDPI 12, IEEE 26, Springer 32, Elsevier 18 and other assorted categories as 36, putting a total of 140.

Table 1. Criteria for selecting publications in the review articles

Parameters	Inclusion	Exclusion
Date Range of (Jan 2015 to Sep 2023)	Publications of Research Consensus within this period	Publications of Research Consensus not this period
Topic of Study	Big Data, Data Intensive Applications	other than the topics of inclusion
Keywords	"cloud","big data","data intensive", "fog","real-time","design"	all other synonymous and generic search words that lead to bias
Redundancy	represent unique titles of publications	scrutinized the title, author names, publications

Qualified	all publications relevant to big data and data intensive applications	all publications that does not reflect ideas of data intensive applications and big data
Language	english	non-english
Criteria	comprehended thoroughly full text and importance is notified for deriving suitable keywords	publications with exclusive presence of keywords

Journal articles that have undergone peer review are included in this comprehensive evaluation of the literature. Making thoughtful, strategic judgments that align with a study's research goals is crucial. For transparency and reliability, all knowledge, insights, and conclusions are thoroughly documented.

Almost since 2015, there has been a discernible rise in the number of articles published on “big data”, “data intensive applications” and “cloud”. The following figure illustrates the distribution of the select research consensus amongst the publication databases of repute between 2015 and 2023.

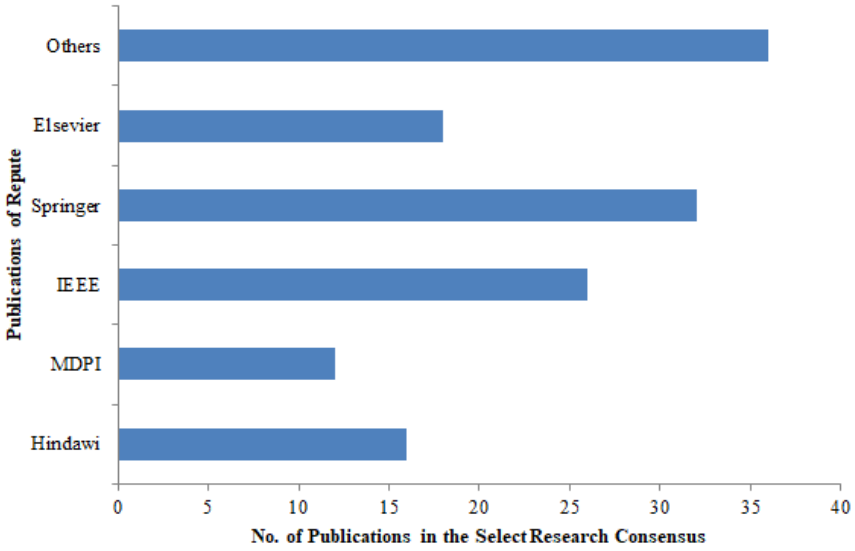

Figure 3: Number of Articles in the Select Research Consensus during 2015-2023

The collected articles are combined with the assorted publications of repute and the thematic distribution of select research consensus is illustrated as follows:

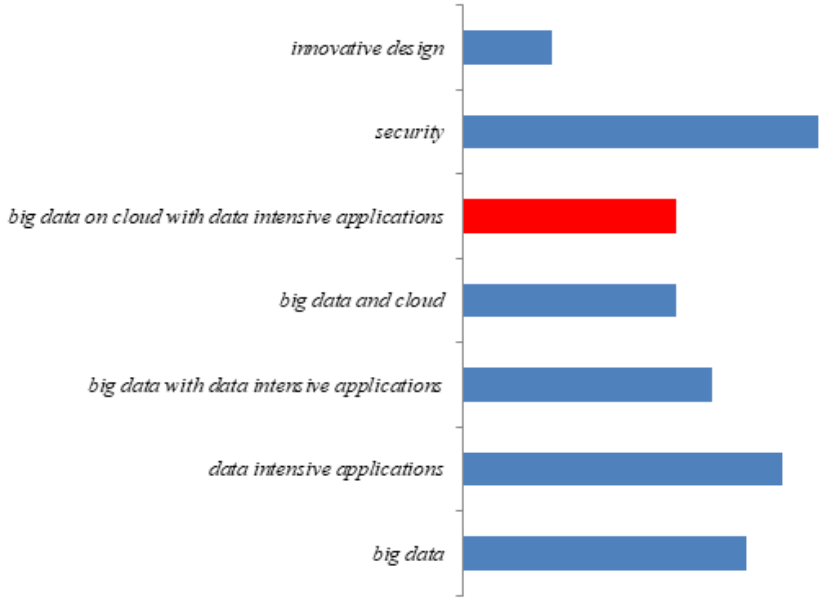

Figure 4: Thematic – distributions of select research consensus

The target of the study is on identifying the relevant articles matching with the themes and keywords and develops a classification model, discriminating the publications with context-relevance. Thematic analysis of data from the 140 articles identified with the themes exhaustively determined as big data, data intensive applications, big data with data intensive applications, big data and cloud, big data on cloud with data intensive applications, security, innovative design. The assessment of the content in each article is iteratively analyzed towards the objective of contributions to the development of an ideal model data intensive application using big data.

D. Discussion and Comparison

As mentioned in the sections of related work, the development of ideal model of application-oriented data intensive application on big data, comprises of smart systems, privacy, security, data collectors and overheads of cloud systems. Large volumes of data were allowed to gather in typical scientific research pertaining to various fields using high-performance data collecting devices. Without modelling and simulation, processing experimental data, or observational apparatus, scientific advances are hard to imagine. Infrastructure is needed for top-level science in order to process data quickly and efficiently, store and transfer massive volumes of data, and enable remote access to resources and systems for the collaborative work of numerous geographically dispersed researchers. Infrastructure must be easily accessible, scalable, and adaptable.

Applications with large data face unique challenges related to data access and storage. In order to lessen these challenges, this current study guides in defining the method for combining the potent cloud computing technology with edge computing concepts to produce low latency response to high data-intensive applications. The transmission rates and waiting time calculations are utilized by the contingent middle and cloud layer controllers to ensure a low latency response. Using the controllers and the priority-based scheduling method also

results in efficient data allocation. Data and storage specific allocation and management of low latency response is the need of the hour. Future efforts will focus on minimising energy and cost for data-intensive applications using genuine cloud and fog setups.

Sensors are crucial for the Information Society's development, [34][36] capturing information for decision-making tools in various sectors like healthcare, retail, and energy [30][35]. Analytics, a practice of data science, uses statistical tools and techniques to analyze data. The overabundance of data generated by sensors has led to the rise of data-intensive scientific discovery. Sensors provide an unparalleled data source for real-time applications, as they provide timestamped information with enough detail to characterize observed phenomena [1][5][19][20].

Table 1. Various works on developing DIA distinguished with Methodologies, Strategies and their Salience.

Ref.	Topics Elicited	Methodology Analyzed	Algorithm / Strategy	Salience / Advantages
[1]	Data-Intensive Computing on Heterogenous Data Sets	Scalability, Elasticity and Heterogenity	A Heterogenity Aware Iterative Strategy	Conservative Data Space, Reduce Latency in Data Access
[2]	Challenges Faces in multi-disciplinary data-intensive software teams	Theory Development using Socio Technical Grounded Theory	A Multi-variate factor analysis	Evaluating Theory, arriving into a mature theory for modeling Data Intensive Applications and Quality Issues
[3]	DIA for Business Domains using Classification Theory	A Framework for Elicitation, Pre-Processing, Extration, Discovery and Classification	An Iterative method applied on dependent variables to elucidate the interactive requirements for the development of DIA	Fast Elucidation using k-NN and Random Forest Methods
[4]	Novel methods of DIA Modeling and Development	A study on Grid Computing and Cloud Computing	Cooperative Paradigms that support modeling DIA	Directions considering Migration, Privacy and Security, Adaptability of various data standards
[5]	Tools that support High Performance Computing for modeling DIA	Comparative study on SQL and NoSQL databases	Resource Management for High Performance model of DIA using Apache Hadoop tools/suite	Scalable agreements between Apache Hadoop and HPC
[6]	DIA with Data Warehouses	Relational Database Technology	Concepts and Tools underlying implementation of Web-enabled, Data Warehouses for modeling a DIA	Characteristics and Requirements of DIA Models

[7]	Various types of DIA	Programming Models for Storage Coupled Arch.	Inverse Computing with Analogy	Suggestions/Requirements for modeling DIA and as well as to solve Computationally Intensive Scientific Applications
[8]	Security, Privacy and Attacks on DIAs	Applying Hadoop and various other Data Modeling environments	Role Based Secure Systems controlling access to the storage systems	Suggestions/Recommendations for Developing SLA to host DIA
[9]	Power Management Aspects while modeling DIA	Monitoring Power System Controller	Blended approach to orchestrate with the problems with Parallel Computing and Cloud computing	Recommendations of Seamless API for Parallel, Cloud Computng in Hadoop
[10]	Semantics Aware Performance Optimization parametres	Understanding Paradigm Shifts Programming and Processing Technologies	A study to list challenges and opportunities in modeling DIA	Recommendations of Effective Strategies for Systems Performance, Semantics Aware Technology with Performance Optimization parameters
[11]	Integrations of Hadoop, SkyTree and Mahout	Consolidation of various tools operating on Big Data	Operations of various tools	Recommendations for Integration of Data Clusters and Distributable In-Memory storages
[12]	Basic ACID rules for modeling DIA	Features of Apache Kafka	Monitoring Event Logs for Adapting Database for DIA models	Event log : Views on Services, Easy Debugging, Easy to recognize failures of transactions

E. Extracted Statistics

The models described in the previous studies have been assessed with their performance metrics. The best DIA design shall contain features of data integration, data adaptation. Data migration is another important aspect for the DIA while integrating heterogenous sources. DIA shall be interactive preserving privacy, security ensuring the confidentiality of personal data. Therefore, the properties insist on the overall throughput time and the latency of responses to the queries received from the user-end. DIA are also responsible for interactive queries and instant visualization of data. From the minimum implementation to the large scale implementation, the DIA shall follow basic relational database rules with sophisticated data rules on cloud systems. The following chart describes the models developed from minimum configuration to the sophisticated configurations ranging from 1 to 20. The model identified with number 1 is said to have minimum configuration of DIA with the basic principles of relational databases with interactivity and the model identified with higher numbers consist of sophisticated configurations dealing with the security, privacy, heterogeneity, interactive and supporting representations of visualizations, which also influential

in the organization of databases. Scalability, Resource utilization, reliability and availability, usability, fault tolerance and compliance of data representation with industry standards are secondary features of the DIA.

Therefore, the statistics represent the features of performance, complexity and seamless operability of DIA. The throughput represents the number of transactions, internal queries, data units processed per unit of time. The time that consumes to process a request from the end user from the initial trigger of operation to the end result to complete the fulfillments of the end user.

Performance measures are collected from the sources and the gross performance observed from the literature, that the Data Intensive Models developed in Cloud platform has considerable progress in development, which is indicated by the key performance measures throughput and latency.

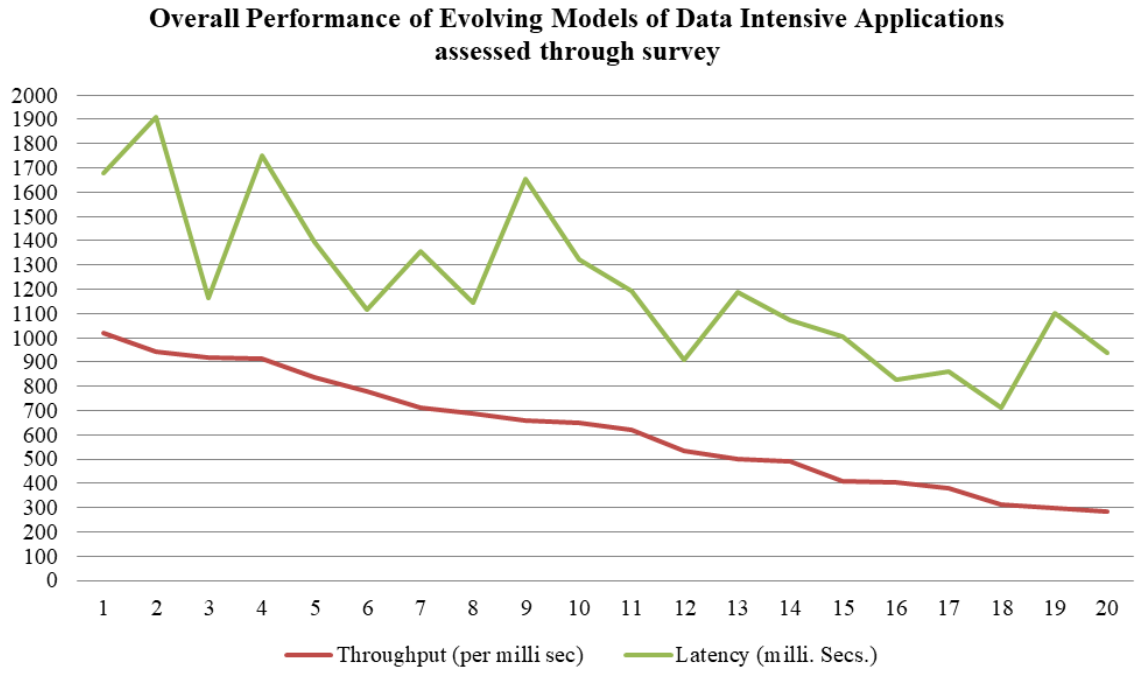

Figure 5: The chart representing the decreasing throughput time, latency proportionate the progressive sophistication in DIA models.

F. Recommendations

Architecture, tools, scalability, and performance are just a few of the factors that must be carefully taken into account when developing data-intensive applications using Big Data on the Cloud. For modelling and creating such apps, the following are the best recommendations and suggestions:

Select the Appropriate Cloud Platform: Select the cloud platform that best meets your needs after evaluating your options. Google Cloud Platform (GCP), Microsoft Azure, Amazon Web Services (AWS), and other options are popular choices.

Scalability in Architecture: Create a scalable infrastructure that is capable of accommodating different workloads. To dynamically modify resources in response

to demand, make use of cloud services like auto-scaling, load balancing, and serverless computing.

Data Administration and Storage: Based on the requirements of your application, choose the right data storage options. Cloud databases (like Amazon Aurora, Azure Cosmos DB), data lakes (like Amazon S3, Azure Data Lake Storage), and distributed file systems (like Hadoop Distributed File System) are some of the available options.

Data Processing Frameworks: For processing massive amounts of data, make use of distributed data processing frameworks like Apache Spark, Apache Flink, or Apache Hadoop. Large datasets can be handled well by these frameworks, which allow for parallel processing.

Instantaneous and Interactive Processing: Incorporate real-time data processing capabilities by utilising tools such as Apache Storm or Apache Flink for real-time analytics, and Apache Kafka for streaming data.

Integrating Artificial Intelligence: If machine learning is a component of the DIA model, frameworks such as TensorFlow, PyTorch, or Scikit-Learn are mandatory components for integration. Machine learning services are inherent and are frequently provided by cloud providers, making model scaling and deployment easier.

Optimization: Selecting appropriate pricing models, using spot instances or reserved instances, and keeping a close eye on resource utilisation, can minimise expenses. Assured minimum expenses during times of low traffic, use auto-scaling to modify resources according to demand.

Microservices and APIs: Though not discussed thoroughly in this article, The DIA model should host a collection of microservices with clear APIs that encourages modularity, maintainability, and the capacity to scalability.

Compliance and Data Governance: Additionally, the DIA model shall consists of ideal data governance policies to guarantee data integrity, quality, and legal compliance. Track data movement and modifications by implementing data lineage and auditing features.

Disaster Recovery and Backup: Put strong disaster recovery and backup plans into action. Make sure you have a plan in place for a prompt recovery in the case of data loss or system failure and take advantage of cloud services for automated backups.

Testing for performance: Perform comprehensive performance testing to locate bottlenecks and enhance the efficiency of your application. Take into consideration using programes like Gatling or Apache JMeter to simulate various system loads.

Update with Cloud Services: The DIA model shall adapt and update to the new features and services that are frequently added to cloud platforms. Following the most recent developments will make the DIA model a niche, reflect high quality and improve the functionality, scalability, and efficiency.

G. Conclusion

Data-intensive base study involves analyzing great volumes of data accessible from portable devices. Scaling data presents challenges for tools and technologies in data management. Grid, Fog, Edge and cloud computing are major paradigms for

scientific applications, with cloud infrastructures playing a significant role in processing and analyzing multidisciplinary data. Big data poses challenges for scientists and IT experts, driving market growth and creating new technologies. To overcome computing challenges, cooperation of data paradigms and new approaches to data handling are needed. Safe and adequate data storage is crucial, and adoption of standards is needed to reduce energy and resource loss. The next step is compatibility of these paradigms in data storage and security.

A Data-Intensive Application (DIA) model is crucial for providing fault tolerance, privacy protection, heterogeneity, scalability, and resilience. It involves using exception management, error handling, and resilient designs to address vulnerabilities. A scalable architecture using cloud services, microservices architecture, and scalable data storage options is essential for managing growing workloads. Heterogeneity of data is handled by supporting various data sources, formats, and processing needs. Compatible technologies and standards are used for easy communication and system integration. Privacy is protected by robust encryption techniques, identity management, access controls, and anonymization. To tolerate faults, redundancy and failover techniques are incorporated into the application design. Data replication and backup procedures are implemented to ensure data availability in case of loss or corruption. Regular fault tolerance testing is conducted to ensure the system can recover and continue working. In conclusion, careful testing, redundancy, and error handling are essential for a robust and scalable DIA model.

H. References

[1] Shamsi, Jawwad, Muhammad Ali Khojaye, and Mohammad Ali Qasmi. “Data-intensive cloud computing: requirements, expectations, challenges, and solutions.” Journal of grid computing 11, no. 2 (2013): 281-310.

[2] Zebari, Rizgar R., S. R. Zeebaree, Karwan Jacksi, and Hanan M. Shukur. "E-business requirements for flexibility and implementation enterprise system: A review." International Journal of Scientific & Technology Research 8, no. 11 (2019): 655-660.

[3] Abdullah, Pavel Y., S. R. Zeebaree, Karwan Jacksi, and Rizgar R. Zeabri. "An hrm system for small and medium enterprises (sme) s based on cloud computing technology." International Journal of Research-GRANTHAALAYAH 8, no. 8 (2020): 56-64.

[4] Rashid, Zryan N., Karzan Hussein Sharif, and S. Zeebaree. "Client/Servers clustering effects on CPU execution-time, CPU usage and CPU Idle depending on activities of Parallel-Processing-Technique operations." Int. J. Sci. Technol. Res 7, no. 8 (2018): 106-111.

[5] Graetsch, Ulrike M., Hourieh Khalajzadeh, Mojtaba Shahin, Rashina Hoda, and John Grundy. “Dealing with data challenges when delivering data-intensive software solutions.” IEEE Transactions on Software Engineering (2023).

[6] Malallah, HayfaaSubhi, Subhi RM Zeebaree, Rizgar R. Zebari, Mohammed AM Sadeeq, Zainab Salih Ageed, Ibrahim Mahmood Ibrahim, Hajar Maseeh Yasin, and Karwan Jameel Merceedi. "A comprehensive study of kernel (issues and concepts) in different operating systems." Asian Journal of Research in Computer Science 8, no. 3 (2021): 16-31.

[7] Abdullah, Pavel Y., S. R. Zeebaree, Hanan M. Shukur, and Karwan Jacksi. "HRM system using cloud computing for Small and Medium Enterprises (SMEs)." Technology Reports of Kansai University 62, no. 04 (2020): 04.

[8] Zeebaree, Subhi RM, Amira B. Sallow, Bzar Khidir Hussan, and Sundos Mohammad Ali. "Design and simulation of high-speed parallel/sequential simplified DES code breaking based on FPGA." In 2019 International Conference on Advanced Science and Engineering (ICOASE), pp. 76-81. IEEE, 2019.

[9] Al-Jumaili, Ahmed Hadi Ali, Ravie Chandren Muniyandi, Mohammad Kamrul Hasan, Johnny Koh Siaw Paw, and Mandeep Jit Singh. "Big Data Analytics Using Cloud Computing Based Frameworks for Power Management Systems: Status, Constraints, and Future Recommendations." Sensors 23, no. 6 (2023): 2952.

[10] Rao, Bingbing, and Liqang Wang. "A survey of semantics-aware performance optimization for data-intensive computing." In 2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/ DataCom/CyberSciTech), pp. 81-88. IEEE, 2017.

[11] Jacksi, Karwan, Subhi RM Zeebaree, and Nazife Dimililer. "Lod explorer: Presenting the web of data." Int. J. Adv. Comput. Sci. Appl. IJACSA 9, no. 1 (2018): 1-7.

[12] Zebari, S. R., and Numan O. Yaseen. "Effects of parallel processing implementation on balanced load-division depending on distributed memory systems." J. Univ. Anbar Pure Sci 5, no. 3 (2011): 50-56.

[13] Jacksi, Karwan, Nazife Dimililer, and S. R. Zeebaree. "State of the art exploration systems for linked data: a review." Int. J. Adv. Comput. Sci. Appl. IJACSA 7, no. 11 (2016): 155-164.

[14] Zeebaree, Subhi RM, Hanan M. Shukur, Lailan M. Haji, Rizgar R. Zebari, Karwan Jacksi, and Shakir M. Abas. "Characteristics and analysis of hadoop distributed systems." Technology Reports of Kansai University 62, no. 4 (2020): 1555-1564.

[15] Shukur, Hanan, Subhi Zeebaree, Rizgar Zebari, Omar Ahmed, Lailan Haji, and Dildar Abdulqader. "Cache coherence protocols in distributed systems." Journal of Applied Science and Technology Trends 1, no. 3 (2020): 92-97.

[16] Raymond, Renita, and Margret Anouncia Savarimuthu. "Retrieval of Interactive requirements for Data Intensive Applications using Random Forest Classifier." Informatica 47, no. 9 (2023).

[17] Loncar, Petra. "Data-Intensive Computing Paradigms for Big Data." Annals of DAAAM & Proceedings 29 (2018).

[18] Jha, Shantenu, Judy Qiu, Andre Luckow, Pradeep Mantha, and Geoffrey C. Fox. "A tale of two data-intensive paradigms: Applications, abstractions, and architectures." In 2014 IEEE International Congress on Big Data, pp. 645-652. IEEE, 2014.

[19] Dave, Meenu, and Hemant Kumar Gianey. "Analysis of big data for data-intensive applications." In 2016 International Conference on Recent Advances and Innovations in Engineering (ICRAIE), pp. 1-6. IEEE, 2016

[20] Kleppmann, Martin, Alastair R. Beresford, and Boerge Svingen. “Online event processing.” Communications of the ACM 62, no. 5 (2019): 43-49.
[21] Abdalla, Hemn Barzan. “A brief survey on big data: technologies, terminologies and data-intensive applications.” Journal of Big Data 9, no. 1 (2022): 1-36.
[22] Wu, Yanhui, Guoqing Li, Lizhe Wang, Yan Ma, J. Kolodziej, and Samee U. Khan. “A review of data intensive computing.” In 12th International Conference on Scalable Computing and Communications (ScalCom), Changzhou, China. 2012.
[23] Fernandez, Eduardo B. “Security in data intensive computing systems.” In Handbook of Data Intensive Computing, pp. 447-466. New York, NY: Springer New York, 2011.
[24] Al-Jumaili, Ahmed Hadi Ali, Ravie Chandren Muniyandi, Mohammad Kamrul Hasan, Johnny Koh Siaw Paw, and Mandeep Jit Singh. “Big Data Analytics Using Cloud Computing Based Frameworks for Power Management Systems: Status, Constraints, and Future Recommendations.” Sensors 23, no. 6 (2023): 2952.
[25] Khalid, Zhwan M., and Subhi RM Zeebaree. "Big data analysis for data visualization: A review." International Journal of Science and Business 5, no. 2 (2021): 64-75.
[26] Jghef, Yousif Sufyan, Mohammed Jasim Mohammed Jasim, Hayder MA Ghanimi, Abeer D. Algarni, Naglaa F. Soliman, Walid El-Shafai, Subhi RM Zeebaree et al. "Bio-Inspired Dynamic Trust and Congestion-Aware Zone-Based Secured Internet of Drone Things (SIoDT)." Drones 6, no. 11 (2022): 337.
[27] Zeebaree, S. R., Rizgar R. Zebari, Karwan Jacksi, and Dathar Abas Hasan. "Security approaches for integrated enterprise systems performance: A Review." Int. J. Sci. Technol. Res 8, no. 12 (2019): 2485-2489.
[28] Zeebaree, S. R., and Karwan Jacksi. "Effects of processes forcing on CPU and total execution-time using multiprocessor shared memory system." Int. J. Comput. Eng. Res. Trends 2, no. 4 (2015): 275-279.
[29] Rao, Bingbing, and Liqang Wang. “A survey of semantics-aware performance optimization for data-intensive computing.” In 2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/ DataCom/CyberSciTech), pp. 81-88. IEEE, 2017.
[30] Assuno, Marcos D., Rodrigo N. Calheiros, Silvia Bianchi, M. A. Netto, and Rajkumar Buyya. “Big data computing and clouds: Trends and future directions.” Journal of Parallel and Distributed Computing 79, no. Supplement C (2015): 3-15.
[31] Conoscenti, Marco, Antonio Vetro, and Juan Carlos De Martin. “Blockchain for the Internet of Things: A systematic literature review.” In 2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA), pp. 1-6. IEEE, 2016.
[32] Kwon, Ohbyung, Namyeon Lee, and Bongsik Shin. “Data quality management, data usage experience and acquisition intention of big data

analytics." International journal of information management 34, no. 3 (2014): 387-394.

[33] Busetto, Loraine, Katrien Ger Luijkx, Arianne Mathilda Josephus Elissen, and Hubertus Johannes Maria Vrijhoef. "Intervention types and outcomes of integrated care for diabetes mellitus type 2: a systematic review." Journal of evaluation in clinical practice 22, no. 3 (2016): 299-310.

[34] Ma, Yan, Haiping Wu, Lizhe Wang, Bormin Huang, Rajiv Ranjan, Albert Zomaya, and Wei Jie. "Remote sensing big data computing: Challenges and opportunities." Future Generation Computer Systems 51 (2015): 47-60.

[35] Ortega, Felipe, and Emilio L. Cano. "Sensor data analytics: challenges and methods for data-intensive applications." Entropy 24, no. 7 (2022): 850.

[36] Haji, Lailan M., Subhi RM Zeebaree, Karwan Jacksi, and Diyar Q. Zeebaree. "A State of Art Survey for OS Performance Improvement." Science Journal of University of Zakho 6, no. 3 (2018): 118-123.