A Methodological Review on Challenges in Big Data Environment

Manpreet Singh
Student
Department of Computer Science
College of Engineering & Management
Kapurthala
Email: mpanesar114@gmail.com

Rimmy Yadav
Assistant Professor
College of Engineering & Management
Kapurthala
Email: rimmyanmol05@gmail.com

Balwinder Ram
Student
Department of Computer Science
College of Engineering & Management
Kapurthala
Email:Balwinderbangar73@gmail.com

Abstract: Big data has shown the great potential in optimizing, making decisions, spotting business trends in various fields such as manufacturing, finance, Information Technology. This paper provides a multi- disciplinary overview of the research issues in big data and its techniques, tools and framework related to the privacy , data storage management, network and energy consumption, fault tolerance and data visualizations. Besides this, outcome challenges and opportunities available in this big data platform have made.

Keywords: Big data, privacy, data storage organization, network and energy consumption organization.

Introduction

The term of big data mainly used to describe massive, heterogeneous, and unstructured digital content that is difficult to process using traditional management tools and techniques. The emergence of big data significantly increasing the managerial, tactical and operational capabilities of an organization rapidly in the terms of data processing, information retrieval, privacy and security and most importantly decision making via advanced big data's tool and techniques. The continuous adoption of Big Data Systems (BDS) are rapidly providing various advantages such as (a) storing vast amount of data (in terms of Petabyte (10^{15}), Exabyte (10^{18}), and Zettabyte (10^{21}), (b) structure (text based data) and unstructured data (images, audio, video and text based data as well), (c) data can be store for a long period of time, (d) helps the decision makers to seek unforeseen data predictions throughout the datasets to make better judgment to grow in the market place [3]. Opposed to tremendous advantages by the big data, there are still a number of performance hindrance subjects (i.e. Personal information leakage, data storages' security issue, network latency and so on) in big data that forces the managerial experts to think about whether to adopt the big data platform or to continue to work with the existing databases. In this paper, a review of literature on privacy techniques, data processing and management approaches along with their shortcoming has been showed.

Problem Definition: The main objective of this paper is outline various disadvantages, challenges and opportunities of big data in terms of privacy, security, data visualization, data processing and management and network and energy management identified during the analysis of available literature on big data.

Literature Review

According to [1] generalize research methodology is adopted. We begin with the study of reputed academic journal databases (Springer, Taylor & Francis, Google Scholar, IEEE, Science

Direct) related to big data privacy approaches, frameworks and techniques for data storage and quality management, fault tolerance and visualization techniques. To reach at the specific end, collected research papers are filtered using abstract analysis. In the final phase, carefully chosen

Big data Privacy approaches

Sr. No	**Ref. No**	**Approach/ Method Used**	**Advantages**	**Limitations**
01.	[26]	• Privacy- Preserving aggregation (Based homomorphic encryption technique.) • De- identification privacy preserving technique.	• It can protect individual privacy in the phases of big data collecting and storing. • De- identification can make data analytics and data mining more effective and flexible.	• Since aggregation is purpose- specific, one- purpose aggregated data usually cannot be used for other purposes. • Its inflexibility prevents running complex data mining to exploit new knowledge. • Insufficient for big data analytics. • An attacker can possibly get more external information assistance for de- identification in the big data era. • Not sufficient for protecting big data privacy.
03.	[16]	A methodological framework which uses data fragmentation combined with a data encryption applied in big data.	• Overcome the threads and information leakage in big data environment, especially for NOSQL databases.	• Manually filtering the irrelevant data consumer more computational time.
04	[47]	Flexible, scalable, dynamical and cost effective privacy- preserving framework is proposed which is based on map reduce on cloud.	• The privacy- preserving framework handles the dynamical update of data sets as well as to maintain the privacy of requirements of such data sets. • Improves the privacy of the disparate data sets located at multiple storage locations.	• Proposed framework still needs extensive investigation to improve the privacy of data sets.
05	[41]	• Information dispersion algorithm (IDA) for data confidentiality in storage systems. • Column- access-	• They deal with the attacks that target to compromised data confidentiality without compromising integrity. • Denial- of- Service attacks. • Achieves the practical trade- off between the	• They did not deal with the internal attacks such as from malicious partners. • Network latency issues deteriorates the overall performance.

		via- proxy operations for data confidentiality in query accesses between clients and server. • B+- tree index for efficient query processing.	security and performance.	
06.	[33]	Frequency distribution block (FDB) and Quasi-identifier distribution block (QIDB) anonymization methods.	• Overcomes record linkage, probabilistic attacks and attribute linkage of patient sensitive or private data.	• Unable to save and support unstructured data.
07.	[50]	• Differential privacy protection scheme for big data in body sensor network (BSN) is proposed. • Dynamic noise threshold.	• Reduces the risk of privacy exposure while ensuring the data availability and accuracy. • Noisy and outlier data is easily detected and removed from the sample data sets.	• No theoretical analysis is provided to apply differential privacy for big sensitive data in body sensor networks (BSN).
08.	[19]	• *Epic*, an extensible and scalable system is proposed. • Map reduce and relational data processing models are implemented.	• Avoid I/O overheads. • Easily parallelize the large data sets on to working nodes and handles the multi-structured data.	• Both the epic and Hadoop experience 2X slowdown when node failure occur.
09.	[12]	A privacy preserving methodology with salted hashing techniques.	• Increase privacy of the electronic services and e-governance. • The proposed transformation can also be applied on to other information, data and attributes of the personal identifiers. • Scalable to large distributed environment and large data	• Unable to empower citizen's to retain control over their personal information while using the advanced governmental electronic services.

			sets.	
10	[51]	Secure and reliable framework called Rampart is developed.	• Eliminate the privacy risk in data preparation. • Reconstruction and modification approaches of rampart framework protect sensitive information from unwanted discovery by data mining algorithms.	• Data provenance method based on game theory requires more research to work reliably.
11	[49]	• Intrusion detection system, named IDnS is developed. • Reputation engine based on big data is built which include approximately 400 million DNS queries.	• Effectively and efficiently detect Advanced Persistence Threat (APT) malware infections based on malicious DNS and traffic analysis. • Reduce the volume of network traffic and improves the sustainability of the network.	• Sometimes the proposed model cannot judge whether a host is infected or not. • For a large and high sped network, it is hard to record all inbound and outbound traffic. • Not good at detecting malware infections that do not rely on domains, such as Trojan use IP address directly to locate the command and control server.
Big data management approaches				
12	[42]	• TAMP: (Transform, aggregation, merge and post preprocessing) model is developed.	• Multi-dimensional data updating problem is minimized. • The join free approach is scalable, efficient and stable despite of the large number of table involved.	• Network latency increased due to the multiple queries. • Computational and query processing time increase.
13	[34]	• A conceptual framework is proposed which includes capturing data, organizing data, analyzing data and values & decision modules.	• Reduces the problem of data processing of huge amount of data. • Proposed model provide reliable and efficient solutions through data capturing, organizing data, analyzing data and finally helps the decision makers to make better decisions.	NA
14	[32]	• Framework for scheduling big data applications over	• Reduces data extraction cost as well as provide query results timely.	• Higher resource utilization rate. • Unable to manage

		geographical distributed cloud data centers is projected.	•	workload among data centers and virtual clusters. • Not suitable for media stream applications.
15.	[29]	• Implement MapReduce model to parallelize the calculations for special type of multi-dimensional data analysis query, namely multiple group- by queries.	• Intermediate data transfer reduction, by mappers and combiners, decreases the communication cost.	• Size of intermediate data increases significantly. • A small part of cost is increased as number of worker node increases in the MapReduce model.
16.	[27]	• Model to improve the data quality in big data is proposed.	• Easily classify the data whether it is relevant to its intended user or not.	• Disadvantage of proposed model is that it partially clean or filter the data.
17.	[18]	• Model for OLAP process is developed.	• Minimize the Input/ Output resource contention in terms of data manipulation, read and write operations throughout the	• Varying number on nodes in a cluster and unbalanced workload deteriorates the overall performance of the model.
18.	[13]	• Newton Raphson-likelihood optimization as a new large- scale learning classifier is proposed for big biomedical data sets.	• Overcome the challenge of long- execution time while accessing the patients' health related data. • Have best correct data classification rate.	• Provide fewer predictions within multi-dimensional data.
19	[9]	• Framework for big data enterprise information processing network is developed.	• Provide valuable information for supply chain decision making. • Respond to customers queries in minimal and fast time. • Enhance the capabilities of mining, warehousing and extracting modules.	NA
20.	[9]	• Graph based data storage technique is implemented to	• Visualization tool helps the decision makers to make relevant/ irrelevant decision	• User's private information leaked by data processing module.

		store and process voluminous data.	throughout the decision process.	
21.	[3]	• Data classification based framework is developed.	• Helps the data analyst to categorize structured and unstructured data within the data sets. • Noise and irrelevant data is easily eliminated.	• Private information is exposed during data processing of data sets.
22	[4]	• Quality assessment based framework is proposed to improve the quality of data sets.	• Quality metrics such as efficiency, performance and flexibility helps to analyze the quality and reliability of data sets.	•
23.	[2]	• Data reduction technique is adopted to optimize the performance of big data.	• Assists in memory overhead reduction. • Reduces the overall computation taken by the mappers and reducers.	• Only suitable for medium scale geographically data sets.
24.	[25]	• Chi- FRBCS (Fuzzy Rule Based Classification System) for imbalanced data present in the big data sets.	• Structured and unstructured data is easily classified with minimal processing time.	• Unable to handle and classify imbalanced data when scale to large distributed environment.
25.	[48]	• K- Mean clustering algorithm.	• Helps to identify knowledge pertaining to research and academia.	• Performance is worst when sets scale to large data.
26.	[51]	• Develop a complete churn analysis model include; (a) Churn prediction model (b) Information prediction model (c) Negative inter-subscriber influence model.	• Telecom operators easily analyze churn behaviors of telecom subscribers. • The prediction model helps the telecom operators to retain existing and to attract new subscribers by offering the better service and quality.	• Distributed cloud infrastructure incurs high cost.
Network and energy efficient approaches				
27	Eleni, Stai	• Hyperbolic data analysis (HDA) for network/ networked	• Assume and identify the missing network links and social network analysis	• Did not perform well in decentralized big network data.

		data analytics. • Rigel and HyperMap data embedding technique.	metrics (network links). • Provides optimal advertisement strategy and targeting for marketing purposes.	
28	[24]	• Computation-efficient heuristic algorithm is implemented.	• Virtual machine placement policy manages the workload effectively within data centers. • Data is processed at a large scale.	• Higher intercommunication traffic and communication cost.
29.	[23]	• A big data traffic data processing framework is proposed using Hbase to analyze the data of intelligent monitoring and recording system (IMRS) is proposed.	• Solves the big data storage and analysis problem vehicle behavior based big data sets. • Increase the data query speed and computing efficiency. • Cluster size can be expanded within the data centers according the requirements so as to improve the concurrent capacity and processing speed.	• For meet these requirements still it is difficult for the programmer to design an analysis algorithm. • The extra consumption of storage space required by the proposed framework is very high.
30	[21]	• Path aware selection method is implemented.	• Maintain the topology of characteristic in data centers. • Effectively detect faulty links in the communication network. • Less number of path selections makes the network topology reliable.	• Computational time increases while selecting the suitable path for data transmission.
31.	[18]	• Hybrid electrical and optical network architecture for big data is advanced.	• Decreases the cost of cooling system from 49.575 to 27% of total cost. Scalability and fault tolerance features of developed model are very high.	• Initial setup/ installation cost for cooling procedure and optical interconnections are very high.
32.	[17]	• RS- pooling strategy which divides the overall storage devices into disjoint subsets or pools and then fault	• Approach is very beneficial for the organization which faces vendor- lock in problem in big data as well as cloud environment. •	• Renting storage resources increases the overall operational cost.

		tolerance is applied to these subsets. • Replication technique is employed in case of any storage device failure.		
33.	[15]	• Efficient Topological based model is proposed.	• The proposed unified model allows better utilization of the network and reduces the communication overhead within big data.	• Service disruption and increase in computational power hinders the performance of projected model.
34.	[13]	• Framework for distributed workload handling in big data and for cloud is implemented. • Vast data sources from cloud and web are stored in a distributed fault tolerant databases and processed through a programming model.	• Vast amount of data sets are easily managed.	• Information leakage is still unresolved. • Performance is very slow when scale to a large scale environment.
35.	[11]	• A unified benchmark approach aiming to assess all the performance parameters involved in cloud based big data application system is proposed.	• Existing parallel framework's performance can be easily obtained.	• Did not include security and workload measures.
36	[7]	• Meta-heuristic search technique is applied.	• Deal with communication failure in big data cluster grids. • Continue to work even in the presence of communication link failure.	• They did not employ automatic repair and recovery of faulty nodes.
37.	[17]	• GreenDi (Green Director): Network-	• Helps to select more energy efficient path to access,	• Huge amount of computation time is

		based routing algorithm acts as a bridge between intended users and most energy efficient paths.	retrieve and process the data. • Works even in the presence of node failures.	consumed to select the best optimal energy efficient path.
38.	[1]	• Multi- domain hierarchical scheduling process is created. • Shortest path algorithm handles more complex data intensive applications in big data platform.	• Improve link utilization within datacenters. • Application completion time is greatly reduced.	• The algorithm use in their research work suffer from longer scheduling time. • Unable to recover distributed resources in case of failure.
39.	[38]	• Energy efficient and critical path scheduling algorithm is applied.	• Maintains the tradeoff between increased energy efficiency and decrease response time effectively in big data environment.	•
40	[50]	Meta- MapReduce (MMR) algorithm	• Resolves the problem of iterations in Hadoop. • Error rates of MMR are smaller. • The speed performance of MMR proves that MapReduce improves the computation complexity substantially on big datasets. • MMR reduces computational complexity.	• The proposed algorithm did not perform accurately well in parallelized environment to increase efficiency.
41.	[2]	• Hit rate geographical location analysis algorithm (HIRAGLAA)	• Manage the geographically distributed big data infrastructure. • Minimize the energy consumption of storage devices. • Load balancing and fault tolerance features make the infrastructure more reliable.	•

literature has studied and depicted in a table form.

From table 1, further detailed analysis can be made to identify in the previous knowledge gap and to suggest some areas for future research work. For this purpose, studies has been carried out in above said directions and these are (big data's privacy, data management, network and energy management, data visualization and fault tolerance, challenges and opportunities) to analyze the current status big data.

The study of table 2 shows the presence of various challenges of big data approaches identified during review of relevant literature. According to the literature review security, privacy, dealing with huge amount of data sets, prediction analysis, scalability and finding root cause failure in big data are the main challenges in big data. Furthermore, this knowledge can be further exploited to design new methods or approaches, frameworks for the minimization of challenges and maximization of opportunities of big data platform.

Conclusion, open research issues, and future directions

In the past few decades, data has been produced in massive quantities by big data connected devices and applications. Continuous advancement in the computational power and the advancement in the recent technologies are the main reason for the data growth. In this paper, we have studied privacy, data management, network and energy consumption management, Fault tolerance and visualization in big data platform. A comprehensive survey of techniques/ frameworks/ methods along with (a) advantage, (b) disadvantages is explored in this paper. The knowledge provided in this paper can be further exploited to design and model new mechanisms or approaches in the cloud. The outlined research gaps will be helpful for the researchers who are motivated to work in the field of big data.

Issues in big data platform
• Security issues are the major concern in big data. One of the important security issues on the input part of the big data is to make sure that the sensors will not be compromised by attacks.
• Security problem on the communication between the big data and other external system is also major concern.
• Dealing with security issues on the analysis part of big data is also a challenging task.
• To identify the Advanced Persistent (APTs) security issues through the disparate systems is a crucial task.
• Leakage of private information by the data analytics techniques to other people after the big data analysis process is a challenging task.
• To deal with confirmation of a file (COF) and learning contents of files (LCF) security attacks in big data a foremost vital task.
• How to get the tradeoff between data privacy and reproducible research is the biggest challenge in big data.
• Different convergence speeds of the same data mining algorithms leading to the problem of synchronization.

• How to mitigate the impact of noise, outliers, incomplete and inconsistent data in big data storage is becoming an open issue. The integration of different data sources is a timely problem. Data integration is confronted with many challenges, such as different data patterns and a large amount of redundant data.
• Manually cleaning data in big data is considered as the main challenge in the arena of big data due to the increasing volume, velocity and variety of data.
• Challenge is to develop a filtering mechanism to keep the useful information in big data.
• Deciding the location where to store the big data is another challenging task.
• To find the root cause failure of distributed compute nodes, databases middleware is an extremely laborious process.
• Storing and comparing large volume of log based data for defect analysis, testing, detecting security breaches is raising issue.
• Collecting the log fie from remote location becomes a challenging due to network bandwidth caps and fragile networks.
• The growth of log repository in the big data systems is becoming a critical subject for the data analysts.
• Existing bug reproduction techniques may expose customer's sensitive information when failures occur in the big data systems.
• Operations over encrypted data are usually complex and time consuming, while big data is high volume and needs us to mine new knowledge in a reasonable timeframe, running operations over encrypted data is inefficient in big data analytics.
• Due to the variety of disparate data sources and the sheer volume, it is difficult to collect and integrate data with scalability from distributed locations. [hu2014]. HAN HU[1]
• Big data systems need to store and manage the gathered massive and heterogeneous data sets in terms of fast retrieval, scalability, and privacy protection. HAN HU[1]
• Hadoop must integrate with real time massive data collection & transmission and provide faster processing beyond the batch processing.
• Hadoop provides a concise user programming interface, while hiding the complex execution. In some senses, this simplicity causes poor performance. So more advanced interface similar to DBMS should be implemented from single angle.
• In big dimensionality, scalability poses as the key challenge to many existing state- of-art methods.
• Computational intelligence and data mining in big data cannot deal with huge size of datasets elegantly and often scored miserably.
• Achieving real- time analysis and prediction on big dimensionality in big data is a new challenge of computational intelligence on portable platforms.
• Big data transfer involves big data generation, acquisition, transformations in the spatial domain. Big data transfer usually incurs high costs, which is a bottleneck for big data computing. And improving the transfer efficiency of big data is a key factor to improve big data computing. [chen 2014]

Future Directions

- Different kind of big data approaches, framework has been developed to provide various services like workload management of big data sets, providing security and privacy measures to stored data, but rarity of work has explored or found in the area of fault tolerance in big data.
- Developing an approach for security and privacy of big data platform with other systems (cloud computing, grid computing, cluster computing) requires attention of researchers in future.
- Problems related to manually cleansing the data sets for data processing phase in big data poses a research issues. So there is a need to develop self-programmed data filtering technique to eliminate noisy, outlier and irrelevant data within the data sets.
- Scalability issue of the existing big data approaches in terms of data management, privacy and security, resource optimization and data extraction by different data mining algorithms is rapid growing challenge which needs to be tackle.
- There is a need to develop a machine learning systems and to define intuitive gestures in research for general and more specific (medical) purposes.

From the review of literature, authors identified few research gaps in context to different areas such as privacy, data management, network and energy management, data visualization, fault tolerance, challenges and opportunities. These research areas can be explored as future work. For this, more analytical studies are required for developing a suitable solution either in the form of framework or method to satisfy the need of enterprises and business organizations.

References

[1] Abouelela, Mohamed, and Mohamed El-Darieby. "Scheduling big data applications within advance reservation framework in optical grids." *Applied Soft Computing* 38 (2016): 1049-1059.

[2] Al-Absi, Ahmed Abdulhakim, Dae-Ki Kang, and Myong-Jong Kim. "Enhancing Dataset Processing in Hadoop YARN Performance for Big Data Applications." *Advanced Multimedia and Ubiquitous Engineering*. Springer Berlin Heidelberg, 2016. 9-15.

[3] Al-Najran, Noufa, and Ajantha Dahanayake. "A Requirements Specification Framework for Big Data Collection and Capture." *New Trends in Databases and Information Systems*. Springer International Publishing, 2015. 12-19.

[4] Anindita. "Evolutionary Algorithm Based Techniques to Handle Big Data." *Techniques and Environments for Big Data Analysis*. Springer International Publishing, 2016. 113-158.

[5] Anuradha, D., and S. Bhuvaneswari. "A Detailed Review on the Prominent Compression Methods Used for Reducing the Data Volume of Big Data." *Annals of Data Science* (2016): 1-16.

[6] Babiceanu, Radu F., and Remzi Seker. "Manufacturing Cyber-Physical Systems Enabled by Complex Event Processing and Big Data Environments: A Framework for Development."

Service Orientation in Holonic and Multi-agent Manufacturing. Springer International Publishing, 2015. 165-173.

[7] Baker, T., et al. "GreeDi: An energy efficient routing algorithm for big data on cloud." *Ad Hoc Networks* 35 (2015): 83-96.

[8] Balicki, Jerzy, Waldemar Korłub, and Maciej Tyszka. "Harmony Search to Self-Configuration of Fault-Tolerant Grids for Big Data." *Advanced and Intelligent Computations in Diagnosis and Control*. Springer International Publishing, 2016. 411-424.

[9] Bilal, Muhammad, et al. "Big data architecture for construction waste analytics (CWA): A conceptual framework." *Journal of Building Engineering* (2016).

[10] Chao, Chian-Hsueng. "The Framework of Information Processing Network for Supply Chain Innovation in Big Data Era." *The 3rd International Workshop on Intelligent Data Analysis and Management*. Springer Netherlands, 2013.

[11] Corbellini, Alejandro, et al. "An Evaluation of Distributed Processing Models for Random Walk-Based Link Prediction Algorithms Over Social Big Data." *New Advances in Information Systems and Technologies*. Springer International Publishing, 2016. 919-928.

[12] Drogkaris, Prokopios, and Aristomenis Gritzalis. "A Privacy Preserving Framework for Big Data in e-Government Environments." *Trust, Privacy and Security in Digital Business*. Springer International Publishing, 2015. 210-218.

[13] Elsebakhi, Emad, et al. "Large-scale machine learning based on functional networks for biomedical big data with high performance computing platforms." *Journal of Computational Science* 11 (2015): 69-81.

[14] Ferrarons, Jaume, et al. "Primeball: a parallel processing framework benchmark for big data applications in the cloud." *Performance Characterization and Benchmarking*. Springer International Publishing, 2013. 109-124.

[15] Fouad, Mohamed Mostafa, et al. "Big Data Pre-processing Techniques Within the Wireless Sensors Networks." *Proceedings of the Second International Afro-European Conference for Industrial Advancement AECIA 2015*. Springer International Publishing, 2016.

[16] Heni, Houyem, and Faiez Gargouri. "A Methodological Approach for Big Data Security: Application for NoSQL Data Stores." *Neural Information Processing*. Springer International Publishing, 2015.

[17] Jaafar, Nouf, Manal Al-Jadaan, and Reem Alnutaifi. "Framework for Social Media Big Data Quality Analysis." *New Trends in Database and Information Systems II*. Springer International Publishing, 2015. 301-314.

[18] Jemal, Dhouha, et al. "MapReduce-DBMS: An Integration Model for Big Data Management and Optimization." *Database and Expert Systems Applications*. Springer International Publishing, 2015.

[19] Jiang, Dawei, et al. "epiC: an extensible and scalable system for processing Big Data." *The VLDB Journal* 25.1 (2016): 3-26.

[20] Kchaou, Hamdi, Zied Kechaou, and Adel M. Alimi. "Towards an Offloading Framework based on Big Data Analytics in Mobile Cloud Computing Environments." *Procedia Computer Science* 53 (2015): 292-297.

[21] Kung, Sun-Yuan. "Discriminant component analysis for privacy protection and visualization of big data." *Multimedia Tools and Applications* (2015): 1-36.

[22] Li, Baobin, and Tingshao Zhu. "Visualization Analysis for Big Data in Computational CyberPsychology." *Human Centered Computing*. Springer International Publishing, 2014. 701-707.

[23] Lin, Chi, et al. "Protecting Privacy for Big Data in Body Sensor Networks: A Differential Privacy Approach." *Collaborative Computing: Networking, Applications, and Worksharing*. Springer International Publishing, 2015. 163-172.

[24] Liu, Xuan, et al. "Meta-mapreduce for scalable data mining." *Journal of Big Data* 2.1 (2015): 1-21.

[25] López, Victoria, et al. "Cost-sensitive linguistic fuzzy rule based classification systems under the MapReduce framework for imbalanced big data." *Fuzzy Sets and Systems* 258 (2015): 5-38.

[26] Lu, Rongxing, et al. "Toward efficient and privacy-preserving computing in big data era." *Network, IEEE* 28.4 (2014): 46-50.

[27] Merino, Jorge, et al. "A Data Quality in Use model for Big Data." *Future Generation Computer Systems* (2015).

[28] Palmieri, Francesco, et al. "Grasp-based resource re-optimization for effective big data access in federated clouds." *Future Generation Computer Systems* 54 (2016): 168-179.

[29] Pan, Jie, et al. "Parallelizing multiple group-by queries using MapReduce: optimization and cost estimation." *Telecommunication Systems* 52.2 (2013): 635-645.

[30] Quezada-Naquid, Moisés, et al. "RS-Pooling: an adaptive data distribution strategy for fault-tolerant and large-scale storage systems." *The Journal of Supercomputing* (2015): 1-21.

[31] Rahman, Mohammad Naimur, and Amir Esmailpour. "A Hybrid Data Center Architecture for Big Data." *Big Data Research* (2016).

[31] Sanchita, Ghosh, and Desarkar Liu, Xuan, et al. "Meta-mapreduce for scalable data mining." *Journal of Big Data* 2.1 (2015): 1-21.

[32] Sandhu, Rajinder, and Sandeep K. Sood. "Scheduling of big data applications on distributed cloud based on QoS parameters." *Cluster Computing* 18.2 (2015): 817-828.

[33] Sarjapur, Kashinath, et al. "Big Data Management System for Personal Privacy Using SW and SDF." *Information Systems Design and Intelligent Applications*. Springer India, 2016. 757-763.

[34] Sanyal, Manas Kumar, Sajal Kanti Bhadra, and Sudhangsu Das. "A Conceptual Framework for Big Data Implementation to Handle Large Volume of Complex Data." *Information Systems Design and Intelligent Applications*. Springer India, 2016. 455-465.

[35] Shi, Liang, et al. "Green and Fault-Tolerant Routing in Data Centers." *Big Data Computing and Communications*. Springer International Publishing, 2015. 465-478.

[36] Stai, Eleni, Vasileios Karyotis, and Symeon Papavassiliou. "A hyperbolic space analytics framework for big network data and their applications." *Network, IEEE* 30.1 (2016): 11-17.

[37] Subbiah, Sankari, et al. "Energy Efficient Big Data Infrastructure Management in Geo-Federated Cloud Data Centers." *Procedia Computer Science* 58 (2015): 151-157.

[38] Sun, Dawei, et al. "Re-Stream: Real-time and energy-efficient resource scheduling in big data stream computing environments." *Information Sciences* 319 (2015): 92-112.

[39] Szczerba, Monika, et al. "Scalable Cloud-Based Data Analysis Software Systems for Big Data from Next Generation Sequencing." *Big Data Analysis: New Algorithms for a New Society*. Springer International Publishing, 2016. 263-283.

[40] Wang, Cong, et al. "Privacy-preserving public auditing for data storage security in cloud computing." *INFOCOM, 2010 Proceedings IEEE*. Ieee, 2010.

[41] Wang, Shiyuan, Divyakant Agrawal, and Amr El Abbadi. "A comprehensive framework for secure query processing on relational data in the cloud." *Secure Data Management*. Springer Berlin Heidelberg, 2011. 52-69.

[42] Wang, Huiju, et al. "Efficient query processing framework for big data warehouse: an almost join-free approach." *Frontiers of Computer Science* 9.2 (2015): 224-236.

[43] Wang, Shiyuan, Divyakant Agrawal, and Amr El Abbadi. "A comprehensive framework for secure query processing on relational data in the cloud." *Secure Data Management*. Springer Berlin Heidelberg, 2011. 52-69.

[44] Xia, Yingjie, et al. "Big Traffic Data Processing Framework for Intelligent Monitoring and Recording Systems." *Neurocomputing* (2015).

[45] Xu, Lei, et al. "A Framework for Categorizing and Applying Privacy-Preservation Techniques in Big Data Mining." *Computer* 49.2 (2016): 54-62.

[46] Zeng, Deze, Lin Gu, and Song Guo. "A General Communication Cost Optimization Framework for Big Data Stream Processing in Geo-Distributed Data Centers." *Cloud Networking for Big Data*. Springer International Publishing, 2015. 79-100.

[47] Zhang, Xuyun, et al. "Privacy preservation over big data in cloud systems." *Security, Privacy and Trust in Cloud Systems*. Springer Berlin Heidelberg, 2014. 239-257.

[48] Zhang, Yi, et al. "Topic analysis and forecasting for science, technology and innovation: Methodology with a case study focusing on big data research." *Technological Forecasting and Social Change* (2016).

[49] Zhao, G. U. O. D. O. N. G., et al. "Detecting APT Malware Infections Based on Malicious DNS and Traffic Analysis." *Access, IEEE* 3 (2015): 1132-1142.

[50] Zhong, Ray Y., et al. "Visualization of RFID-enabled shop floor logistics Big Data in Cloud Manufacturing." *The International Journal of Advanced Manufacturing Technology* (2015): 1-12.

[51] Zhu, Yuqing, et al. "Bigop: Generating comprehensive big data workloads as a benchmarking framework." *Database Systems for Advanced Applications*. Springer International Publishing, 2014.