International Journal of Electrical, Electronics and Computer Engineering **6**(1): 92-99(2017)

ISSN No. (Online): 2277-2626

A Literature Review on Big Data Reduction Methods

Swati Yadav* and Ajay Phulre**

**M-Tech Scholar, Department of Computer Science and Engineering,*
SBITM, Betul, (Madhya Pradesh), INDIA
***Asst. Professor, Department of Computer Science and Engineering,*
SBITM, Betul, (Madhya Pradesh), INDIA

(Corresponding author: Swati Yadav)

(Received 27 February, 2017 Accepted 14 March, 2017)
(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: Research on big data analytics is the latest area where multiple gigabytes of data arrive in the big data systems every second. These big data systems collect different complex data streams due to the volume, velocity, value, variety, variability, and veracity in the acquired data and consequently give rise to the 6V's of big data and this is not limited to 6V's. The reduced and relevant data streams are considered to be more useful than collecting raw, redundant, inconsistent, and noisy data. Another perspective for big data reduction is that the million variables big datasets cause the curse of dimensionality which requires high computational resources to discover actionable knowledge patterns. This article presents a review of methods that are used for big data reduction. It also presents a detailed taxonomic discussion of big data reduction methods including the network theory, big data compression, dimension reduction, redundancy elimination, data mining, and machine learning methods. In addition, the open research issues pertinent to the big data reduction are also highlighted.

Keywords: Big data, Data compression Data reduction, Data complexity, Dimensionality reduction

I. INTRODUCTION

Big data is the combination of huge volume, high velocity and different varieties of data streams originated from heterogeneous and autonomous data sources [1]. The volume is one of the most important characteristic of big data that is represented by the acquisition of storage spaces in large-scale data centers and storage area networks. The massive size of the big data not only causes the data heterogeneity but also results in diverse dimensionalities in the datasets. Therefore, lot of efforts are required to reduce the volume to effectively analyze big data [2]. In addition, most of the time big data streams are needed to be processed online to avoid lateral resource consumption for storage and processing. The second key feature of big data is velocity. The velocity refers to the frequency of data streams, which is needed to be shorten in order to handle big data effectively. For example, Walmart generates 40 PB data per day and the analysis of such a fast big data is possible only after reduction or summarization [3]. On the other hand, big data inherits the 'curse of dimensionality.' In other words, millions of dimensions (variables, features, attributes) are required to be effectively reduced to uncover the maximum knowledge patterns [4, 5]. For example, behavior profiles of the Internet users that mainly comprise of searches, page-views, and click-stream data are sparse and high dimensional with millions of possible keywords and URLs [6]. Similarly, social network websites like facebook not only increases the volume and velocity of data but also adds to the high dimensionality of the data [7]. Therefore, it is imperative to reduce the high dimensions while retaining the most important and useful data.

Data reduction methods for big data vary from pure dimension reduction techniques to compression-based data reduction methods and algorithms for preprocessing, redundancy elimination, and implementation of network (graph) theory concepts. Dimension reduction techniques are useful to handle the heterogeneity and massiveness of big data by reducing million variable data into manageable size [8 9,10,11]. These techniques usually work at after the data collection phases. Similarly, cluster deduplication and redundancy elimination algorithms that remove duplicated data for efficient data processing and useful knowledge discovery are primarily post-data collection methods [12, 13, 14, 15]. Recently, the network theory concepts have also been used for big data reduction [16, 17, 18].

The above mentioned methods first extract the semantics and linked structures from the unstructured datasets and then apply graph theory for network optimization. Conversely, some methods to reduce big data during the data collection process are also proposed in the recent literature [19, 20, 21]. In this study, we presented a detailed discussion of these data reduction methods.

This article presents a literature review of methods for big data reduction. In this we also explain some of the precious study. However, these studies either present a generic discussion of big data reduction or discuss a specific group of relevant systems or methods. For example, the authors in [1] discussed the big data reduction to be the critical part of mining sparse, uncertain, and incomplete data. Similarly, the authors in [22, 23] argue big data reduction as the critical part of data analysis and data preprocessing. The authors in [4] discussed big data reduction issue specifically by focusing on dimension reduction, whereas the authors in [24] emphasized on the data compression. Therefore, we aim to present a detailed literature review that is specifically articulated to highlight the existing methods relevant to big data reduction. In addition, some open research issues are also presented to direct future researchers.

The main contributions of this article are:

-A detailed literature review and classification of big data reduction methods are presented.

-Recently proposed schemes for big data reduction are analyzed and synthesized.

-A detailed gap analysis for the articulation of limitations and future research challenges for data reduction in big data environments is presented.

The article is structured as follows: Sect. 2 discuses the complexity problem in big data and highlights the importance of big data reduction. The taxonomical discussion on big data reduction methods is presented in Sect. 3. The discussion on open issues and future research challenges is given in Sect. 4, and finally, the article is concluded in Sect. 5.

II. WHAT IS BIG DATA COMPLEXITY AND WHY WE NEED DATA REDUCTION

Big data systems include social media data combinations, industrial or wireless sensor networks, scientific experimental systems, aeroplane log file system, connected health, and several other application areas. The data collection from large-scale local and remote sensing devices and networks, Internet-enabled data streams, and/or devices, systems, and networks-logs brings massively heterogeneous, multi-source, multi-format, aggregated, and continuous big data streams. Effectively handling the big data stream to store, index, and query the data sources for lateral data processing is among the key challenges being addressed by researchers [25,26]. However, data scientists are facing data flood issue to uncover the maximum knowledge patterns at fine-grained level for effective and personalized utilization of big data systems [3, 27]. The data flood is due to 6Vs properties of big data, namely the volume, variety, value, velocity, veracity, and variability.

-Volume -The data size characterizes the volume of big data. However, there is no agreed upon definition of big data which specifies the amount of data to be considered as 'big' in order to meet the definition of big data. However, a common sense is developed in research community who consider any data size as big in terms of volume which is not easily processable by underlying computing systems. For example, a large distributed system such as computing clusters- or cloud-based data centers may offer to process multiple terabytes of data but a standalone computer or resource constrained mobile devices may not offer the computational power to process even a few gigabytes of data. Therefore, the volume property of big data varies according to underlying computing systems.

-Velocity -The velocity of big data is determined by the frequency of data streams which are entering in big data systems. The velocity is handled by big data systems in two ways. First, the whole data streams are collected in centralized systems, and then, further data processing is performed. In the second approach, the data streams are processed immediately after data collection before storing in big data systems. The second approach is more practical; however, it requires a lot of programming efforts and computational resources in order to reduce and filter the data streams before entering in big data systems.

-Variety-Big data systems collect data stream from multiple data sources which produce data streams in multiple formats. This heterogeneity in data sources and data types impacts the variety property-related characteristics. Therefore, big data systems must be able to process multiple types of data stream in order to effectively uncover hidden knowledge patterns.

-Veracity-The utility of big data systems increases when the data streams are collected from reliable and trustworthy sources. In addition, the data stream collection is performed with compromising the quality of data streams. The veracity property of big data relates to reliability and trustworthiness of big data systems.

-Variability-Since all data sources in big data systems do not generate the data streams with same speed and same quality. Therefore, variability property enables to handle the relevant issues. For example, the elastic resource provisioning as per the requirements of big data systems.

-Value-The value property of big data defines the utility, usability, and usefulness of big data systems. This property tends more toward the outcomes of data analytics and data processing processes and is directly proportional to other 5Vs in big data systems.

The well-designed big data systems must able to deal with all 6Vs effectively by creating a balance between data processing objectives and the cost of data processing (i.e., computational, financial, programming efforts) in big data systems. Moreover, the complexity in big data systems emerges in three forms: (1) data complexity, (2) computational complexity, and (3) system complexity [28]. The data complexity arises due to multiple formats and unstructured nature of big data, which elevate the issue of multiple dimensions and the complex inter-dimensional and intra-dimensional relationships. For example, the semantic relationship between different values of the same attribute, for example, noise level in the particular areas of the city, increases the inter-dimensional complexity. Likewise, the linked relationship among different attributes (for example, age, gender, and health records) raises the intra-dimensional complexity issue. In addition, the increasing level of data complexity in any big data system is directly proportional to the increase in computational complexity where only the sophisticated algorithms and methods can address the issue. Moreover, the system-level complexity is increased due to extensive computational requirements of big data systems to handle extremely large volume, complex (mostly unstructured and semi-structured), and sparse nature of the data. The extensive literature review exhibits that the big data reduction methods and systems have potential to deal with the big data complexity at both algorithms and systems level. In addition to data complexity, the big data reduction problem is studied in various other perspectives to articulate the effects and the need of data reduction for big data analysis, management, commercialization, and personalization.

Big data analysis also known as big data mining is a tedious task involving extraneous efforts to reduce data in a manageable size to uncover maximum knowledge patterns. To make it beneficial for data analysis, a number of preprocessing techniques for summarization, sketching, anomaly detection, dimension reduction, noise removal, and outliers detection are applied to reduce, refine, and clean big data [29]. The New York Times, a leading US newspaper, reports that data scientists spend 50-80% of the time on cleaning the big datasets [30]. The terms used in the industry for the aforementioned process are 'data munging,' 'data wrangling,' or 'data janitor work.' Another issue with the large-scale high-dimensional data analysis is the over-fitting of learning models that are generated from large numbers of attributes with a few examples. These learning models fit well within the training data, but their performance with testing data significantly degrades [31].

Data management is another important aspect to discuss the big data reduction problem. The effective big data management plays a pivotal role from data acquisition to analysis and visualization. Although data acquisition from multiple sources and aggregation of relevant datasets improve the efficiency of big data systems, it increases the in-network processing and data movement at clusters and data center levels. Similarly, the indexing techniques discussed in [26] enhance the big data management; however, the techniques come across data processing overheads. Although the conversion of unstructured data to semi-structured and structured formats is useful for effective query execution, the conversion in itself is a time- and resource-consuming activity. Moreover, big data is huge in volume that is distributed in different storage facilities. Therefore, the development of learning models and uncovering global knowledge from massively distributed big data is a tedious task. Efficient storage management of reduced and relevant data enhances both the local learning and global view of the whole big data [32, 33]. Currently, visual data mining technique of selecting subspace from the entire feature spaces and subsequently finding the relevant data patterns also require effective data management techniques. Therefore, the reduction in big data at the earliest enhances the data management and data quality and therefore improves the indexing, storage, analysis, and visualization operations of big data systems.

Recently, businesses particularly the enterprises are turning into big data systems. The collection of large data streams from Web users' personal data streams (click-streams, ambulation activities, geo-locations, and health records) and integration of those data streams with personalized services is a key challenge [34]. The collection of irrelevant data streams increases the computational burden that directly affects the operational cost of enterprises. Therefore, the collection of fine-grained, highly relevant, and reduced data streams from users is another challenge that requires serious attention while designing big data systems. Currently, user data collection by third parties without explicit consent and information about commercialization is raising the privacy issues. The participatory personal data where users collect and mine their own data and participate for further utilization and customization of services in ubiquitous environments can address the issue of fine-grained data availability for enterprises. Keeping in view the big data complexity, the need for big data reduction, and analyzing big data reduction problem in different perspective, we present a thorough literature review of the methods for big data education.

The core technological support for big data reduction methods is based on multilayer architecture (see Fig. 1). The data storage is enabled by large-scale data centres and networks of different computing clusters [35]. The storage infrastructures are managed by core networking services, embarrassingly parallel distributed computing frameworks, such as Hadoop map-reduce implementations and large-scale virtualization technologies [36]. In addition, cloud services for the provision of computing, networking, and storage are also enabled using different cloud-based operating systems. A recent phenomenon in cloud computing is enabling the edge-cloud services by the virtualization of core cloud services near the data sources. Recently, Cisco released a Fog cloud to enable the intercommunication between core cloud services and proximal networks of data sources [37, 38]. At the lowest layers of the big data architecture resides the multi-format data sources which include standalone mobile devices, Internet-enabled social media data streams, remotely deployed wireless sensor networks, and large-scale scientific data streams among many others.

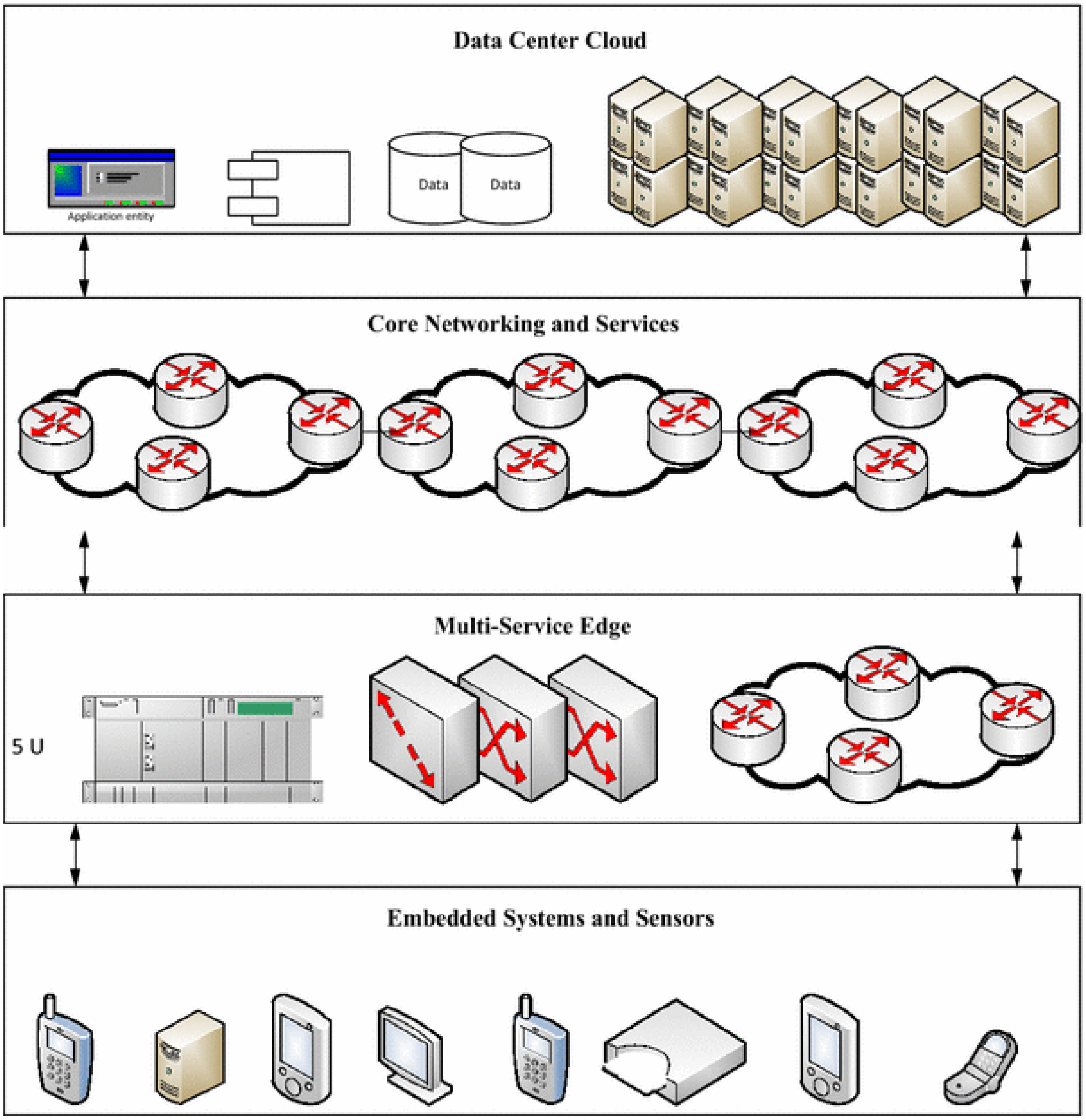

Fig. 1. Multilayer architecture for big data systems.

This layered architecture enables to process and manage big data at multiple levels using various computing systems with different form factors. Therefore, wide ranges of application models are designed and new systems have been developed for big data processing.

III. BIG DATA REDUCTION METHODS

This section presents the data reduction methods being applied in big data systems. The methods either optimize the storage or in-network movement of data or reduce data redundancy and duplication. In addition, some of the methods only reduce the volume by compressing the original data and some of the methods reduce the velocity of data streams at the earliest before entering in big data storage systems. Alternatively, some of the methods extract topological structures of unstructured data and reduce the overall big data using network theory approaches that are discussed as follows.

A. Network Theory

Network (also known as graph) theory is playing a primary role in reduction of high-dimensional unstructured big data into low-dimensional structured data [39]. However, the extraction of topological structures (networks) from big data is quite challenging due to the heterogeneity and complex data structures. The authors in [40] proposed network theory-based approach to extract the topological and dynamical network properties from big data. The topological networks are constructed by establishing and evaluating relationships (links) among different data points. The statistical node analysis of the networks is performed for optimization and big data reduction. The optimized networks are represented as small-world networks, free-scale networks, and random networks and are ranked on the basis of statistical parameters, namely mean, standard deviation and variance.

B. Compression

The reduced-size datasets are easy to handle in terms of processing and in-network data movement inside the big data storage systems. Compression-based methods are suitable candidates for data reduction in terms of size by preserving the whole data streams. Although computationally inefficient and involving decompression overhead, the methods allow to preserve the entire datasets in the original form. Numerous techniques for big data compression are proposed in the literature, including spatiotemporal compression, gzip, anamorphic stretch transform (AST), compressed sensing, parallel compression, sketching, and adaptive compression. a reduction in cloud environments is quite challenging due to multiple levels of virtualization and heterogeneity in the underlying cloud infrastructure. A spatiotemporal technique for data compression on big graph data in the cloud generates reduced datasets. The technique performs online clustering of streaming data by correlating similarities in their time series to share workload in the clusters. In addition, it performs temporal compression on each network node to reduce the overall data. The proposed technique effectively meets the data processing quality and acceptable fidelity loss of the most of the application requirements. On the other hand, wireless sensor networks (WSNs) are generating large data streams at massive scales. The spatiotemporal data compression algorithm ensures efficient communication, transmission, and storage of data in WSNs-based big data environment. The proposed approach not only reduces the size of transmitted data but also ensures prolonged network lifetime. The algorithm measures the correlation degree of sensed data, which determines the content of the data to be transmitted.

C. Data Duplication (Redundancy Elimination)

Data redundancy is the key issue for data analysis in big data environments. Three main reasons for data redundancy are: (1) addition of nodes, (2) expansion of datasets, and (3) data replication. The addition of a single virtual machine (VM) brings around 97% more redundancy, and the growth in large datasets comes with 47% redundant data points [13]. In addition, the storage mechanism for maximum data availability (also called data replication) brings 100% redundancy at the cluster level. Therefore, effective data deduplication and redundancy elimination methods can cope with the challenge of redundancy. The workload analysis shows that the 3× higher throughput improves performance about 45% but in some extreme cases the performance degrades up to 161%. The energy overhead of deduplication is 7%; however, the overall energy saved by processing deduplicated data is 43%. The performance is degraded to 5%, whereas energy overhead is 6% for pure solid state drive (SSD) environments. However, in hybrid environment the system's performance is improved up to 17%.

D. Dimension Reduction

Big data reduction is mainly considered to be the dimension reduction problem because the massive collection of big data streams introduces the 'curse of dimensionality' with millions of features (variables and dimensions) that increases the storage and computational complexity of big data systems [5]. A wide range of dimension reduction methods are proposed in the existing literature.

The methods are based on clustering, map-reduce implementations of existing dimension reduction methods, feature selection techniques, and fuzzy logic implementations.

IV. OPEN RESEARCH ISSUES

The discussion on the open research issues, limitations, and possible future research directions is presented in this section.

Network theory-The extraction of topological network and ranking of network nodes from big data is a complex process due to inherent big data complexity. In addition, the complex interactions among different nodes of the extracted networks increase the computational complexity of existing network theory-based methods. The scale-free networks and random networks can effectively reduce complex big datasets. However, the full network extraction from inconsistent and missing data is the key challenge [16,40]. Big data systems contain many small and manageable datasets, but finding the connections among these datasets is a crucial task. The similarity graph is generated from big data where vertices represent datasets and the weighted edges are defined on the basis of similarity measure. The graph is further reduced by merging similar datasets to reduce the number of nodes. The similarity-based big data reduction methods are good choice for network extraction and reduction. However, a range of new similarity measures are required to deal with the evolving complexity and to fully comply with 6Vs of big data [17].

Compression: Big data processing in cloud computing environments involves challenges relevant to inefficiency, parallel memory bottlenecks, and deadlocks. The spatiotemporal compression is a key solution for processing big graph data in the cloud environment. In spatiotemporal compression-based methods, the graph is partitioned and edges are mapped into different clusters where compression operations are performed for data reduction. The spatiotemporal compression is an effective approach for big data reduction. However, the research is required to find new parameters that are helpful in finding additional spatiotemporal correlations for maximum big data reduction.

Data deduplication (redundancy elimination): Cluster level data deduplication is a key requirement to comply with service-level agreements (SLAs) for privacy preserving in cloud environments. The main challenge is the establishment of trade-off between high deduplication ratio and scalable deduplication throughput. The similarity based deduplication scheme optimizes the elimination process by considering the locality and similarity of data points in both the intra-node and inter-node scenarios. The approach is effective for data reduction, but it requires to be implemented with very large-scale cluster data deduplication systems [12]. The I/O latency and extra computational overhead of cluster-level data deduplication are among the key challenges. The authors in [13] characterized the deduplication schemes in terms of energy impact and performance overhead. The authors outlined three sources of redundancy in cluster environment including: (1) the deployment of additional nodes in the cluster, (2) the expansion of big datasets, and (3) the usage of replication mechanisms. The outcomes of the analysis reveal that the local deduplication, at cluster level, can reduce the hashing overhead. However, local deduplication cannot achieve the maximum redundancy. In contrast, global deduplication can achieve maximum redundancy but compromises on the hashing overheads. In addition, fine-grained deduplication is not suitable for big datasets especially in streaming data environments [13].

Data pre-processing: The investigations of research problems relevant to pre-processing techniques of big data are still at the initial level. Most of the works are based on the adoption of existing pre-processing methods that were earlier proposed for historical large datasets and data streams. The forefront deployment of data pre-processing methods in the big data knowledge discovery process requires new, efficient, robust, scalable, and optimized pre-processing techniques for both historical and streaming big data. The application of appropriate and highly relevant pre-processing methods not only increases data quality but also improves the analytics on reduced datasets. The research on new methods for sketching, anomaly detection, noise removal, feature extraction, outliers detection, and pre-filtering of streaming data is required to reduce big data effectively. In addition, the deployment of adaptive learning models in conjunction with said methods can aid in dynamic pre-processing of big streaming data [21].

Dimension reduction -Big data reduction is traditionally considered to be a dimension reduction problem where multimillion features spaces are reduced to manageable feature spaces for effective data management and analytics. Unsupervised learning methods are the key consideration for dimensionality reduction problem. However, this literature review revealed several other statistical and machine learning methods to address this issue. The techniques to combine conventional dimension reduction methods with statistical analysis methods can increase the efficiency of big data systems [8]. This approach may aid in targeting highly dense and information oriented structures (feature sets) to achieve maximum and efficient big data reduction.

Alternately, tensor decomposition and approximation methods are useful to cope with the curse of dimensionality that arises due to high-dimensional complex and sparse feature spaces [10]. The main application of TD-based methods is witnessed in the scientific computing and quantum information theory domain.

V.CONCLUSIONS

Big data complexity is a key issue that is needed to be mitigated. The methods discussed in this article are an effort to address the issue. The presented literature review reveals that there is no existing method that can handle the issue of big data complexity single-handedly by considering the all 6Vs of big data. The studies discussed in this article mainly focused on data reduction in terms of volume (by reducing size) and variety (by reducing number of features or dimensions). However, further efforts are required to reduce the big data streams in terms of velocity and veracity. In addition, the new methods are required to reduce big data streams at the earliest immediately after data production and its entrance into the big data systems. In general, compression-based data reduction methods are convenient for reducing volume. However, the decompression overhead needs to be considered to improve efficiency.

Similarly, network theory-based methods are effective for extracting structures from unstructured data and to efficiently handle the variety in big data. The data deduplication methods are useful to improve the data consistency. Therefore, the aforementioned methods are a suitable alternative to manage the variability issues in big data. Likewise, data pre-processing, dimension reduction, data mining, and machine learning methods are useful for data reduction at different levels in big data systems. Keeping in view the outcomes of this survey, we conclude that big data reduction methods are emerging research area that needs attention by the researchers.

REFERENCES

[1]. Wu X et al (2014). Data mining with big data. *IEEE Trans Knowl Data Eng* **26**(1): 97-107.

[2]. Che D, Safran M, Peng Z (2013) From big data to big data mining: challenges, issues, and opportunities. In: Database systems for advanced applications

[3]. https://datafloq.com/read/big-data-walmart-big-numbers-40-petabytes/1175

[4]. Zhai Y, Ong Y-S, Tsang IW (2014). The emerging "big dimensionality". *Comput Intell Mag IEEE* **9**(3): 14-26.

[5]. Fan J, Han F, Liu H (2014). Challenges of big data analysis. *Nat Sci Rev* **1**(2):293-314

[6]. Chandramouli B, Goldstein J, Duan S (2012). Temporal analytics on big data for web advertising. In: 2012 IEEE 28th international conference on data engineering (ICDE)

[7]. Ward RM et al (2013). Big data challenges and opportunities in high-throughput sequencing. *Syst Biomed* **1**(1): 29-34.

[8]. Weinstein M et al (2013). Analyzing big data with dynamic quantum clustering. arXiv preprint arXiv:1310.2700.

[9]. Hsieh C-J et al (2013). BIG & QUIC: sparse inverse covariance estimation for a million variables. In: Advances in neural information processing systems.

[10]. Vervliet N et al (2014). Breaking the curse of dimensionality using decompositions of incomplete tensors: tensor-based scientific computing in big data analysis. *IEEE Signal Process Mag* **31**(5): 71-79.

[11]. Feldman D, Schmidt M, Sohler C (2013). Turning big data into tiny data: constant-size coresets for k-means, pca and projective clustering. In: Proceedings of the twenty-fourth annual ACM-SIAM symposium on discrete algorithms.

[12]. Fu Y, Jiang H, Xiao N (2012). A scalable inline cluster deduplication framework for big data protection. In: Middleware 2012. Springer, pp 354-373.

[13]. Zhou R, Liu M, Li T (2013) Characterizing the efficiency of data deduplication for big data storage management. In: 2013 IEEE international symposium on workload characterization (IISWC).

[14]. Dong W et al (2011). Tradeoffs in scalable data routing for deduplication clusters. In: FAST.

[15]. Xia W et al (2011). SiLo: a similarity-locality based near-exact deduplication scheme with low RAM overhead and high throughput. In: USENIX annual technical conference.

[16]. Trovati M, Asimakopoulou E, Bessis N (2014). An analytical tool to map big data to networks with reduced topologies. In: 2014 international conference on intelligent networking and collaborative systems (INCoS).

[17]. Fang X, Zhan J, Koceja N (2013). Towards network reduction on big data. In: 2013 international conference on social computing (SocialCom)

[18]. Wilkerson AC, Chintakunta H, Krim H (2014). Computing persistent features in big data: a distributed dimension reduction approach. In: 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)

[19]. Di Martino B et al (2014). Big data (lost) in the cloud. *Int J Big Data Intell* **1**(1-2): 3-17.

[20]. Brown CT (2012). BIGDATA: small: DA: DCM: low-memory streaming prefilters for biological sequencing data

[21]. Lin M-S et al (2013). Malicious URL filtering-a big data application. *In 2013 IEEE international conference on big data.*

[22]. Chen J *et al* (2013). Big data challenge: a data management perspective. *Front Comput Sci* **7**(2):157-164MathSciNet.

[23]. Chen X-W, Lin X (2014). Big data deep learning: challenges and perspectives. *IEEE Access* **2**: 514-525.

[24]. Chen Z et al (2015). A survey of bitmap index compression algorithms for big data. *Tsinghua Sci Technol* **20**(1):100-115MathSciNet.

[25]. Hashem IAT et al (2015). The rise of "big data" on cloud computing: review and open research issues. Inf Syst **47**:98-115.

[26]. Gani A et al (2015). A survey on indexing techniques for big data: taxonomy and performance evaluation. In: Knowledge and information systems, pp 1-44.

[27]. Kambatla K et al (2014). Trends in big data analytics. *J Parallel Distrib Comput* 74(7):2561-2573

[28]. Jin X et al (2015). Significance and challenges of big data research. *Big Data Res* **2**(2): 59-64.

[29]. Li F, Nath S (2014). Scalable data summarization on big data. *Distrib Parallel Databases* **32**(3): 313-314.

[30]. Lohr S (2014). For big-data scientists, 'janitor work' is key hurdle to insights. http://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html

[31]. Ma C, Zhang HH, Wang X (2014). Machine learning for big data analytics in plants. *Trends Plant Sci* **19**(12): 798-808

[32]. Ordonez C (2013). Can we analyze big data inside a DBMS? In: Proceedings of the sixteenth international workshop on data warehousing and OLAP.

[33]. Oliveira J, Osvaldo N et al (2014). Where chemical sensors may assist in clinical diagnosis exploring "big data". *Chem Lett* **43**(11): 1672-1679.

[34]. Shilton K (2012). Participatory personal data: an emerging research challenge for the information sciences. *J Am Soc Inform Sci Technol* **63**(10): 1905-1915.

[35]. Shuja J et al (2012). Energy-efficient data centers. *Computing* **94**(12): 973-994.

[36]. Ahmad RW *et al* (2015). A survey on virtual machine migration and server consolidation frameworks for cloud data centres. *J Netw Comput Appl* **52**: 11-25.

[37]. Bonomi F et al (2014). Fog computing: a platform for internet of things and analytics. In: Big data and internet of things: a roadmap for smart environments. Springer, pp 169-186.

[38]. Rehman MH, Liew CS, Wah TY (2014). UniMiner: towards a unified framework for data mining. In: 2014 fourth world congress on information and communication technologies (WICT).

[39]. Patty JW, Penn EM (2015). Analyzing big data: social choice and measurement. Polit Sci Polit 48(01):95-101.

[40]. Trovati M (2015). Reduced topologically real-world networks: a big-data approach. *Int J Distrib Syst Technol (IJDST)* **6**(2): 13-27.