Big Data Trends and Analytics: A Survey

Payal Saha
Department of CST
Guru Nanak Dev University, Amritsar, India

Mohit Mittal
Department of CSE
Model Institute of Engineering and Technology
Jammu, India

Shreya Gupta
Department of CST
Guru Nanak Dev University,
Amritsar, India

Marwa Sharawi
Faculty of Technology Arab Open University El-Sherouk City,Egypt

ABSTRACT
Big Data is nowadays one of the apex fields of research area. It is due to expansion in technological field at rapid rate. Expansion of storage area and data has been seen from past five year which is exponentially. It is envisioned that concept of Big Data will assure to reduce the huge chunks of data into manageable form. In this paper, we have discussed concept of Big Data, characteristics and challenges. Its main focus is over data generated in various sector, analytics and various tools to manage data.

Keywords
Big data, Hadoop, Mapreduce, Data analytics, Big data tools.

1. INTRODUCTION
Big data handle the dataset in a very traditional way by database system. For business purpose a very high technology or techniques is required to handle these data. In big data it includes unstructured data; semi structured data and structured data. The unstructured data is that kind of data which includes unformatted data for example multimedia and social media [1]. In structured data, these are in formatted so we can directly use it in management system. Semi structured data includes either formatted or unformatted data. Data can be generated from different source and can be stored in a very high rate. Industries like Google, Facebook, Amazon, Myntra were used for big data from very starting. Big data face many challenges like security, privacy, sensor design, storage problems, data analyze and many more. Big data [3], [24] is a data whose diversity, scale, volume and complexity requires a new technique, design or architecture, algorithm and analytic to manage the data and find some value or any kind of hidden information. Myraid hardware objects also provide big data which includes sensors and actuators which are embedded in physical objects which are known as things from internet. The techniques which are used to store data in big data include multiple clustered network attached storage. The different groups which have storage devices are attached with different network and then clustered together. We need a different kind of platform which solves all the complexity or problems of big data. The term is Hadoop which is used to process the semi structured and unstructured big data and make it useful for analytics. Hadoop uses the map-reduce paradigm to select the correct data, only the data directly answer the question or kind of query. For structured data there are several processes available like NoSQL and MongoDB [80]. Hadoop is open source software. It is planned to increase proportionally up from a single server to millions of machines, with a very high degree of error tolerance.

Data! It is a set of elements. It is really very difficult to measure huge set of data or total volume of data which are stored electronically. Data set have size, complexity and growth rate which make it difficult to analyze or capture from technology. 'Big data' is similar to 'small data' but the size is too big [7]. With the increment of data we need different approaches like techniques, architecture views and tools to solve the present, past and future problems. Traditional computing techniques cannot analyze the large amount of data where big data create value from storage and process. It is in very large quantities of digital information. For example: Face book handles 50 billion photos from its user.

There are several features of big data:

- Product based companies and some organizations moniter the social media like facebook, twitter, instagram get unstructured data which are filtered or managed by big data.
- Through big data recognization of sale and market opportunities increases.
- In healthcare centers bigdata is very useful. It analyze the medical data and patients record.
- Many advertising companies or insurance companies track social media using bigdata.

Figure 1 Big Data and its characteristics

1.1 5Vs of Big Data

Big data point out to large as well as massive volumes of datasets [29]. There are presence of several mode of explanations for big data. Here, 5Vs are typically used to characterize of Big Data as volume, velocity, variety, veracity and value [32].

Volume is denoted as specifically size of data; velocity means high speed of data; variety indicates different types of data; veracity emphasized on consistency of data; and value provides outputs for gains from large data sets.

1) **Volume: -** It directly refers to the amount of data. Day by day volume of data increases, the amount data which are stored in enterprise vault have grown from kilobytes, megabytes and gigabytes to zetabytes.
2) **Velocity: -** It tells the processing speed or motion of data. Data is created very fast and processed and analyzed.
3) **Variety: -** Here different kind of data from different types of source. These data are structured, unstructured and semi-structured. Structured data come from business lines or from applications. Unstructured or semi-structured data are come from social media, e-mail, audio and video etc.
4) **Veracity**: - It describes consistency of data which are coming from social media or another kind of source.
5) **Value: -** Value basically describes or provides outputs for gains from large data sets. It also describes statistical, events, correlations and hypothetical data sets.

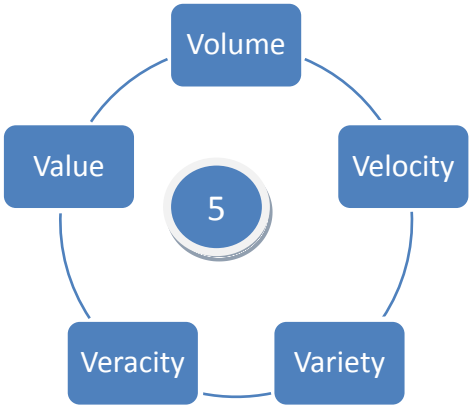

Figure 2 5Vs of big data

2. GOALS OF BIG DATA

a) To provide advantage to companies that makes effective business decisions.
b) In business field, increase sale per every year.
c) It provides a SQL interface to all data.
d) Big data use metaphor. Through this big data able to answering those question which are unpredictable.
e) The main goal of big data to achieve process optimization to give accurate outcomes to the user.

3. BIG DATA: ARCHITECTURE, ANALYTICS AND TECHNIQUES

Data management is the big challenging task to aggregate the data from different source. Data are in different form, like there is structured, unstructured and semi structured data which need to deal in big data. First we need to collect only relevant sources and collect the whole element and store it. Hadoop [26] is the only programming framework which provides distributed file system storage and fault tolerance [27]. Map reduce is the heart of Hadoop. Through map reduce we can take the benefit of Hadoop Distribution of File System. It makes the processing speed fast and quick as soon as possible.

3.1 Big Data: source

Big Data is concept of collection of massive amount of data that could be aggregated, stored, communicated, and analyzed. Nowadays, it is part of every sector and function of the global economy. Data can be categorized in many ways. In table 1, we have represented various data sources like banking, insurance, securities etc. Data can be captured that would be acoustic in nature i.e. audio, video, text and image files.

Table 1 Big data from different source

	Video	Image	Audio	Text/ numbers
Banking				
Insurance				
Securities and investment services				
Discrete manufacturing				
Process manufacturing				
Retail				
Wholesale				
Professional services				
Consumer and recreational services				
Health care				
Transportation				
Communications and media[2]				
Utilities				
Construction				
Resource industries				
Government				
Education				

Penetration
High
Medium
Low

3.2 Big Data: Architecture

For Big data simple counting is not a big complex problem. Large unstructured data from different sources create major complex problems. To resolve this problem, big data need some exceptional technologies [28]. Big data introduced different technologies and various techniques for analyzing data, manipulating and visualization the data. There are so many way to handle the big data, Hadoop [7] is very famous open source programming framework. It is used by researchers or by big data organizations. There are different techniques used to analyze the data. Hadoop software [33] is used for unstructured and semi structured data.

The Current Appache Hadoop environment consists of the Kernel, HDFS, Map Reduce and numbers of different types of components like Apache Hive, zookeeper and Base. There exists much software for analysis. NoSQL, that mean either 'no SQL' or 'not only SQL,' is determined by big data that is Available, Soft state, and eventually consistent (BASE), rather than the traditional database data characteristics of ACID. ACID refer to: Atomicity, Consistency, Isolation and Durability. Data analyzed by using NoSQL, therefore, is at times in a state of transition and may not be directly available; the data is in flux rather than set as in traditional database environments. MongoDb and ZettaStore are both NoSQL-related products that are used for "document-oriented applications" for storage and searching of whole invoices rather than the individual data fields from the invoice.

3.3 Big Data: Analytics

Big data analytics is the process in which we examine the huge amount of data to find some important information or some hidden data, unknown complementary relationships, market inclinations in particular directions, client preference, and other notable business information. Analytics [45] find the easy and more effective way of marketing, new and effective revenue opportunities and better client service, improve efficiency, advantage over organizations and other business benefits.

Figure 4 Big data analytics: Cost reduction, decision making and products and services

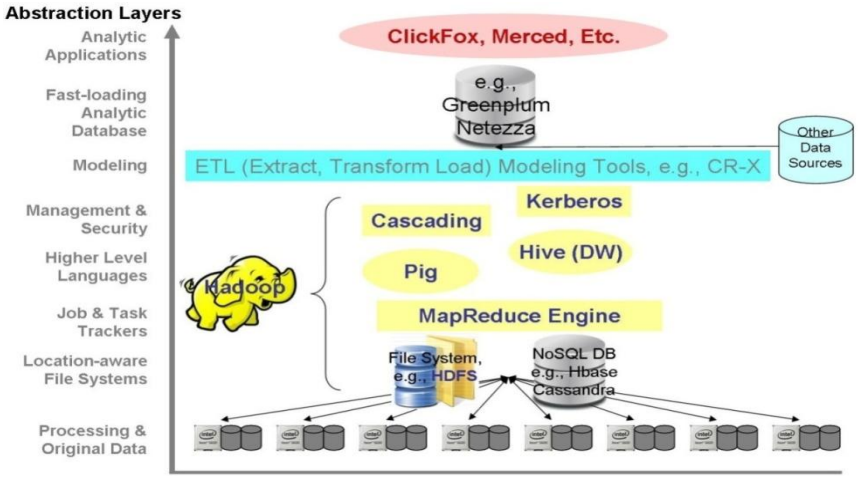

Figure 5 Various components of big data analytics

Table 2 Data analytics methods

Fields	References	Methods
Clustering	. (Zhang et. al, 1996) [77]	BIRCH
	(Ester et. al, 1996) [21]	DBSCAN
	. (Ester et. al, 1998) [21]	Incremental DBSCAN
	(Ordonez and Omiecinski, 2004) [50]	RKM
	(Elkan, 2003) [19]	TKM
Classification	(Mehta et. al, 1996) [36]	SLIQ
	(Micó L, et al.1996) [39]	TLAESA
	(Djouadi and Bouktache, 1997) [18]	FastNN
	(Ververidis and Kotropoulos, 2008) [64]	SFFS
	(Catanzaro et. al, 2008) [9]	GPU-based SVM
Association rules	(Pei et. al, 2000) [53]	CLOSET
	(Han et. al, 2000) [25]	FP-tree
	(Zaki and Hsiao, 2005) [76]	CHARM
	(Burdick et. al, 2001) [8]	MAFIA
	(Chen et. al, 2002) [10]	FAST
Sequential patterns	(Zaki, 2001) [75]	SPADE
	(Yan et. al, 2003) [73]	CloSpan
	(Pei et. al, 2001) [54]	PrefixSpan
	(Ayres et. al, 2002) [2]	SPAM
	(Masseglia et. al, 2003) [34]	ISE

3.3.1 Types of big data analytics

The legal contract of "analytics" is not especially exact, so before fixing to analytics into the correct technology [31], it helps to have some exactness.

a) Research and Development

The invention of conception and growth of algorithms for all forms of measurements functions deserves the research and development. This is the reservation of an expert on mathematics, statisticians and other clean quantitative scientists. The invention and improvement of computer-based algorithms for canopy clustering, k-means and Naïve Bayes type concepts is mostly the country of academia and different research institutions.

b) Data Scientists

In data scientist it includes advertising software companies, vertical software process, and even the great weighted "quants" who work in company, who apply this process specifically to the work they do, since they operate in much the same way as commercial software companies, but for just one customer (though they often start their own software companies, too).

c) Operational Analytics

It is most interesting type of big data analytics. For example, in the second type of analytics the intelligent employee may create or to change a specific direction to scoring model for their company. In the third type of analytics all parameters are selected by third type of analytics and input are given into the model, the scores which are generated are calculated by data scientist and lodged into an operational system that produce proposal for credit card. Models are developed by data scientist and lodged a number a ways. The applications of second type of data analytics application into real work is country of third type of analytics. Decision-making systems which are dependent on quantitative process that are not accurately comprises by the operator can lead to difficult situation. They must be carefully created to eschew overly load the recipients of unhelpful or not applicable knowledge.

d) Business Intelligence and Discovery

The third type of analytics isn't of great valuable if their application in real business conditions cannot be evaluated for their effectiveness. This is the analytical work we are almost acquainted with via reports, OLAP, dashboards and visualizations.

3.3.2 Big Data: Techniques for analyzing

1) Text analytics

Text mining is the process in which we analyze the unstructured data to get some useful information. For example text data from blog, mails, online forums, corporate document and bill centre are. The decomposition and elicitation process in text analytics make use of modes and techniques involved in statistics analysis, computation linguistic and machine learning.

2) Question answering (QA) techniques

They are designed to give response to an interrogative asked in natural language like Apple's Siri and IBM; S. Watson is commercially used QA systems. QA techniques can be divided into 3 categories: information accessing data (IR) based approach, knowledge-based approach, and the hybrid approach.

3) Audio analytics

Audio analytic is those techniques used to resolve and extract some information from unstructured audio data. The current application areas of audio analytic are customer call centers and healthcare institutions. Using these techniques, call centers analyze the million hours of the recorded calls in order to gain insight into customer's behavior and identify the issues regarding the product or service to improve customer experience, for the evaluation of the performances of the agents and enhance sales turnover rates.

4) Video analytics

It has several of techniques which program, analyze, and abstract sense knowledge from video streams. The video analysis demand is growing day by day as the use of closed circuit television (CCTV) cameras is increasing and websites for sharing videos are also getting popular. Though various techniques are developed for video analytics but they are still not much efficient due to the challenge imposed because of the diaphanous size of video. The first application of videos analytics is in automated security and surveillance system. Video analytics [65], [66] can be used throughout and effectively to perform an individual functions like detection of rift of restricted areas, identification of articles removed or left not attended and tampering in cameras, recognition of suspicious activities, and upon detection of threat the system can apprise the personals in real-time or can trigger automatic alerts (e.g. lock doors, sound alarm etc) or the video data from CCTV cameras can be analyzed to keep the count of the number of customers visiting the store, to measure how much time they are staying and spending in the different areas of the store, detecting their movement patterns and monitor queues in real-time.

5) Social media analytics

Social media analytics is the analysis of organize and unorganized data composed from manifold social media channels. The data across the internet, from a number of different online platforms which facilitates users to create and exchange data is gathered together for gaining valuable insights. These online social media platforms are social networks (e.g. Facebook and LinkedIn), blogs (example Blogger and WordPress), micro blogs (e.g. twitter and tumbler), social news (e.g. Digg and Reddit), social bookmarking (e.g. Delicious), media sharing (e.g. Instagram and YouTube), wikis (e.g. Wikipedia), question-answer sites (Yahoo! Answers and Ask.com) and review sites (e.g. TripAdvisor). The structure of a social network is in form of two types of graphs: social graphs which represent the existence of the link (e.g. friendship) and activity graphs which represent the actual interactions between the entities. Agility graphs are preferable to social graphs, because an agility relationship is more applicable to analysis than marsh connections.

6) Predictive analytics

It comprises of several practical for predicting future outcomes based on the past or historical and current data. The most famous predictive modeling techniques used is NNs, SVMs, decision trees, linear and symbolic regression, clustering, affiliation rules, and scorecards. We can apply predictive analysis to all the disciplines from the prediction of breakdown of jet engines based on data streams coming from thousands of detectors, to predicting clients next postures based on what they are buying, when

they are buying and what they are saying on social media sites.

3.3.3 Various characteristics and its description

Table 3 Characteristics and descriptions

S. No.	Characteristics	Description
1.	Hadoop	The most cited technology in connection with big data. Hadoop is an open source framework. It allows distribute processing of large data-sets across clusters of in corporative computers. Often used to analyze customer behavior at large retailers, it could also be used to analyze security events or process large image sets to look for meaningful changes over time. Hadoop is at center of a cluster of open source projects, which include tools for management data collection storage and machine learning.
2.	Map Reduce	Usually mentioned the same breath as Hadoop, map reduce is a programming technique for processing parallelizable problems across very large datasets. Many aspects of public safety systems would fit into this framework as large network of sensors could be individually analyzed in parallel in one another to extract information and patterns that could be summarized at a higher level.
3.	NoSQL databases	NoSQL is class of database system that can be used when the underlying data relationship are not usefully understood through the traditional relation database model. The emphasis of these databases is on storing and retrieving very large data sets for real time or statistical analysis. An example might be the large stream of motion detection and other sensors events that come from an enterprise or municipal surveillance system.
4.	Columnar databases	A type of database is which storage access is optimized around columns rather than rows, which is more efficient for certain type of computations particularity those based on aggregation of columnar data sets. This technology can reduce storage costs and speed or data access for certain classes of problems, possibly edging proposed solution closer to feasibility or AOI.
5.	In-memory databases	In-memory databases are those fit large data sets into main RAM memory as opposed to disk file storage. This approach can speed query time try orders of magnitude - a make-or-break difference. In-memory datasets have become practical, as RAM prices have fallen dramatically over the past decade. They are useful when response time in critical, which again includes many life safety and loss prevention summaries.
6.	Advantage Storage	None of these technologies would have gained a foothold were if not for the availability of very fast, inexpensive storage technology. While disk drive price have continued to drop solid state drives have further reduced success access times for primary storage and both can be joined searnlessly with the virtual storage management software tools that are needed for very large distributed data sets. Most large enterprise is already using these storage technologies for their own corporate data, and would recognize the value in reducing loss or minimizing risk through large scale security event processing.

7.	Talent	Talent has been added to the list because without the right type of data scientists in an organization, none of the technologies will yield any useful results. This is a big shift for most organization and the physical security world in particular as deep analytics has not been a major focus of our products outside of national security applications classified or otherwise. In this regard Mckinsey predicts that by 2018 there will be a shortage of 140,000 to 190,000 people of these needed skills.

3.3.4 Big data tools and its description

Table 4 Big data tools and its description

Field	Tools	Description	References
Data Storage & Management	Hadoop	It is a framework for storing large sets of data by distributing it on computer clusters.	(Bhosale and Gadekar, 2014) [7]
	Cloudera	CDH builds data hub for business enterprises and helps organization with better access to the data.	(Pol, 2014) [55]
	MongoDB	MongoDB is a data storage and management tool for structured and semi-structured data that changes frequently.	(Zhao, 2013) [80]
	MDM	MDM combine real-time data of products and customers across the businesses and integrate it into one form.	(Myers, 2016) [47]
Data Cleaning	Open Refine	OpenRefine is a open source data cleaning tool that helps in organizing unstructured data to a structured form.	(Larsson, 2013) [33]
	Data Cleaner	DataCleaner cleans the semi-structured data and transform into clean readable data that can be utilized further for analysis.	(Choudhary, 2014) [17]
Data Mining	Rapid Miner	Rapid Miner is open source data science platform to streamline your predictive analytics process and deliver projects faster.	(Rangra, and Bansal, 2014) [57]
	Oracle Data Mining	Oracle Data Mining helps in discovering the future insights, making predictions and also provides access to Oracle Data.	(Berger, 2012) [5]
	Teradata	It is RDBMS for large scale data warehousing	(Xiao, and Cheng, 2015) [71]

	Kaggle	Kraggle is a platform for predictive modeling and has analytics competitions on which companies and researchers post their data and statisticians and data miners from all over the world	(Masurel et. al, 2013) [35]
Data Analysis	Qubole Data Service	It provides users with management of Hadoop infrastructure and helps them to analyze the data in the cloud.	(Kumar and Selvan, 2015) [31]
	BigML	BigML collects the data, create data sets and turn the data into predictive models. Further, it combines the models with existing data to get predictions.	(Xing et. al, 2016) [72]
Data Visualization	Tableau	It is a business intelligence tool to analyze the data visually.	(Pradhananga et. al, 2015) [56]
	Platfora	Platfora converts raw big data in Hadoop into interactive data processing engine.	(Wang et. al, 2015a) [65]
	Datawrapper	Datawrapper is an open source tool helping everyone to create simple, correct and embeddable charts in minutes.	(Babu and Babu, 2016) [3]
Data Languages	R	R is a programming language and software environment for statistical computing and graphics	(Goyal and Singh, 2015) [24]
	Python	Python is an open source language for manipulation and analysis of data.	(Babu and Babu, 2016) [3]

4. DATA IN FINANCIAL SECTOR

Financial institutions deals with raising capital trading in securities and managing corporate mergers and acquisitions and retail banking are of key importance in big data technology. Word "STAC" technology used to identify to create at the same time to use uses that pose big data challenges. Our important objective was to examination cases that were to banking by investing staff with direct knowledge of cases, to characterize workloads and understand the problems that arise with traditional technologies and in this we will able to summarizes advantages and challenges of new approaches. The purpose was to lay the ground work for technology benchmark standards that applied to big data problems. The STAC benchmark council develops benchmark specifications for technologies used in strategic business functions by which user firms and vendors can understand the capabilities of competing solution stacks. The specifications cover performance, scaling, resource efficiency, resilience, security and entitlements which will rise lend of awareness in the industry and attract more people to this project. Big data is defined as work load that is too difficult or expensive to handle using traditional technologies largely due to data scale or complexity. This study was qualitative, not quantitative which characterize important workload.

5. BIG DATA: FUTURE TRENDS

Big data has marked its footsteps in many sectors. It has already influenced the world with its advanced technologies and efficient results. Further, Big data is an evolving technology that will always continue to contribute to the giant companies and individuals with its new technologies and tools. With better technology and tools, big data will help the organizations to take more accurate decisions and advance their business to another level. Microsoft predicted the

importance of big data in future by stating "*Big data absolutely has the potential to change the way governments, organizations, and academic institutions conduct business and make discoveries, and its likely to change how everyone lives their day-to-day lives*".

Many sectors that need to make risk driven decisions in the advent of both advancing technologies and the changing nature of data; they are starting to dig deeper into the big data and its benefits. Even Indian government has also acknowledged the importance of big data and has recently announced that it will use Big Data analytics to understand Indian citizen's sentiments and ideas through crowd sourcing platform www.mygov.in and social media to get a picture of common people's thought and opinion on government actions. The IDC has given its prediction about the future trends of big data in its Future Scope for Big Data and Analytics. We will discuss below some predictions that represent expected trends with greatest potential impact on Big Data and analytics initiatives.

- It will be expecting 3x money will spent on cloud big data in coming 5 years span than on-premise upfront cost.
- Due to technology being in infancy, there will be a great need of efficient workforce skilled in big data analytics to complete supply-demand chain. There will be a need of more than one hundred and fifty thousand roles in US alone.
- Applications incorporating predictive analysis and machine learning are bound to facilitate user experience and hence giving edge to the apps without the BDA application.
- Analyzing live feeds of data using Internet of Things (IoT) [52] is increasing at a rapid pace. IoT is one of the emerging trends which engulf various technologies and provide a better platform to manage it.
- Due to increase in of demand of 5G communication, multimedia analytic will expand parallel to database analytics.

6. CONCLUSION

Big data analytics is evolved as transformation in every sector of information. It has capabilities to reduced huge chunks of information into meaning sets. There exists many analytical methods and are found to be effective one but still required more efficient method. This paper has focused on various trends in big data, categories of type of information, analytical methods and study of financial sector data. We have mange a table on tools and analytics of big data that have been created but still there is required more research as due to exponentially increasing volume of data day by day.

7. REFERENCES

[1] Abbass , H. A., Leu, G., and Merrick, K. 2016. 'A Review of Theoretical and Practical Challenges of Trusted Autonomy in Big Data', IEEE Access, Vol. 4, pp. 2808 – 2830.

[2] Ayres J, Flannick J, Gehrke J and Yiu T. 2002. 'Sequential Pattern Mining using a bitmap representation', Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 429–435.

[3] Babu, T. G., and Babu, G. A. 2016. 'A Survey on Data Science Technologies & Big Data Analytics', International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 6, Iss. 2, pp. 322-327.

[4] Balshetwar, S.V., and Tugnayat, R.M. 2015. 'Techniques for analyzing framed data', global journal of engineering science and researches, vol.2, iss. 8, pp. 80-83.

[5] Berger, C. 2012. 'Oracle Data Mining 11g Release 2 Competing on In-Database Analytics', Oracle Corporation, pp. 1-25.

[6] Bhatnagar, V. 2013. 'Data mining-based big data analytics: parameters and layered framework', Int. J. of Computational Systems Engineering, Vol.1, No.4, pp.265 – 276.

[7] Bhosale, H. S. and Gadekar, D. P 2014. 'A Review Paper on Big Data and Hadoop', International Journal of Scientific and Research Publications, Vol. 4, Iss. 10, pp. 1-7.

[8] Burdick D, Calimlim M and Gehrke J. 2001. 'MAFIA: a maximal frequent itemset algorithm for transactional databases', Proceedings of the International Conference on Data Engineering, pp 443–452.

[9] Catanzaro B, Sundaram N and Keutzer K. 2008. 'Fast support vector machine training and classification on graphics processors', Proceedings of the International Conference on Machine Learning, pp 104–111.

[10] Chen B, Haas P and Scheuermann P. 2002. 'A new two-phase sampling based algorithm for discovering association rules', Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 462–468.

[11] Chen, H., Shi, Q., Tan, R., Poor, H. V., Sezaki, K. 2010. 'Mobile element assisted cooperative localization for wireless sensor networks with obstacles', IEEE Transactions Wireless Communications, Vol. 9, Issue: 3, pp. 956-963.

[12] Chen, H., Wang, G., Wang, Z., So, H. C., Poor, H. V. 2012. 'Non-Line-of-Sight Node Localization Based on Semi-Definite Programming in Wireless Sensor Networks', IEEE Transaction Wireless Communications, Vol. 11, Issue: 1, pp. 108-116.

[13] Chen, H., Gao, F., Martins, M., Huang, P., and Liang, J. 2013. 'Accurate and Efficient Node Localization for Mobile Sensor Networks', *ACM/Springer Journal on Mobile Networks and Applications (MONET)*, Vol. 18, pp. 141-147.

[14] Chen, H. M., Kazman, R., and Haziyev, S. 2016. 'Agile Big Data Analytics for Web-Based Systems: An Architecture-Centric Approach', IEEE Transactions on Big Data, Vol. 2, Iss. 3, pp. 234 – 248.

[15] Chi, M., Plaza, A., and Benediktsson, J. A. 2016 'Big Data for Remote Sensing: Challenges and Opportunities', Proceedings of the IEEE, Vol.104, Iss. 11, pp. 2207 – 2219.

[16] Cho, J. and Rajagopalan, S. 2002 'A fast regular expression indexing engine', ICDE, pp. 1-12.

[17] Choudhary, N. 2014 'A Study over Problems and Approaches of Data Cleansing/Cleaning', International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 4, Iss. 2, pp. 774-779.

[18] Djouadi A and Bouktache E. 1997 'A fast algorithm for the nearest-neighbor classifier', IEEE Trans Pattern Anal Mach Intel, vol. 19, pp. 277–282.

[19] Elkan C. 2003 'Using the triangle inequality to accelerate k-means', Proceedings of the International Conference on Machine Learning, pp. 147–153.

[20] Engle, R. s 2001. 'GARCH 101: An Introduction to The Use of ARCH/GARCH models in applied econometrics'. Journal of economics perspectives, Vol. 15, No. 4, pp. 157-168.

[21] Ester M, Kriegel HP, Sander J and Xu X 1996. 'A density-based algorithm for discovering clusters in large spatial databases with noise', Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, pp. 226–231.

[22] Ester M, Kriegel HP, Sander J, Wimmer M and Xu X 1998 'Incremental clustering for mining in a data warehousing environment', Proceedings of the International Conference on Very Large Data Bases, pp 323–333.

[23] Fan, P. 2016. 'Coping with the big data: Convergence of communications, computing and storage', China Communications, Vol. 13, Iss. 9, pp. 203 – 207.

[24] Goyal, H., and Singh, S. 2015. 'Big Data Analysis Using R (Big Data Analysis Applications, Challenges, Techniques)', International Journal of Advanced Research in Computer Science and Software Engineering Vol. 5, Iss. 9, pp. 818-823.

[25] Han J, Pei J and Yin Y. 2000. 'Mining frequent patterns without candidate generation' Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 1–12.

[26] Hernandez, A. F. R., and Garcia, N. Y. G. 2016. 'Distributed processing using cosine similarity for mapping Big Data in Hadoop', IEEE Latin America Transactions , Vol.14, Iss. 6, pp. 2857 – 2861

[27] Hu, J., and Vasilakos, A. V., 2016. 'Energy Big Data Analytics and Security: Challenges and Opportunities', IEEE Transactions on Smart Grid, Vol.7, Iss.5, pp. 2423 – 2436

[28] Jiang, H., Wang, K., Wang, Y., Gao, M., and Zhang, Y. 2016 'Energy big data: A survey', IEEE Access, Vol. 4, pp. 3844 – 3861

[29] Kanoun, K., Tekin, C. and Atienza, D. (2016) 'Big-Data Streaming Applications Scheduling Based on Staged Multi-Armed Bandits, IEEE Transactions on Computers, Vol. 65, Iss. 12, pp. 3591 – 3605.

[30] Kong, L., Zhang, D., He.,Z., Xiang, Q., Wan, J., and Tao, M. 2016. 'Embracing big data with compressive sensing: a green approach in industrial wireless networks', IEEE Communications Magazine, Vol. 54, Iss. 10, pp. 53 – 59.

[31] Kumar P. S., and Selvan, T. T. 2015. 'A Survey of Qubole Data Service on Big Data Analytics and Cloud Computing', International Journal of Advanced Research in Computer and Communication Engineering, Vol. 4, Iss. 6, pp. 46-48

[32] Laney, D. 2001. '3D data management: controlling data volume, velocity, and variety', META Group.

[33] Larsson, P. 2013. 'Evaluation of Open Source Data Cleaning Tools: Open Refine and Data Wrangler', http://courses.cs.washington.edu/courses/cse544/13sp/final-projects/p12-plarsson.pdf

[34] Masseglia F, Poncelet P and Teisseire M. 2003. 'Incremental mining of sequential patterns in large databases', Data Knowl Eng., vol. 46, pp. 97–121.

[35] Masurel, P., Bourguignat, C., Hasegawa, K. L., and Scordia, M. 2013. 'Dataiku's Solution to Yandex's Personalized Web Search Challenge', WSCD workshop, pp. 1-9.

[36] Mehta M, Agrawal R and Rissanen J. 1996. 'SLIQ: a fast scalable classifier for data mining', Proceedings of the 5th International Conference on Extending Database Technology Advances in Database Technology, pp 18–32.

[37] Mertz, L. 2016a. 'What Can Big Data Tell Us About Health? Finding Gold Through Data Mining, IEEE Pulse, Vol. 7, Iss. 5, pp. 40 – 44.

[38] Mertz, L. 2016b. 'The Case for Big Data: New York City's Kalvi HUMAN Project Aims to Use Big Data in Resolving Big Health Questions', IEEE Pulse, Vol. 7, Iss. 5, pp. 45 – 47.

[39] Micó L, Oncina J and Carrasco R C. 1996. 'A fast branch and bound nearest neighbour classifier in metric spaces', Pattern Recogn Lett, vol.17, issue 7, pp.731–739.

[40] Mittal, M. and Kumar, K. 2015a. 'Delay Prediction in Wireless Sensor Network Routing Using ART1 Neural Network', IEEE, African Journal of Computing & ICT, vol 8. no. 3, pp. 175-180.

[41] Mittal, M. and Kumar, K. 2015b. 'Energy Efficient Homogeneous Wireless Sensor Network Using Self-Organizing Map (SOM) Neural Networks', IEEE, African Journal of Computing & ICT Vol 8. No. 1, pp. 179-184.

[42] Mittal, M. and Kumar, K. 2015c. 'Quality of Services Provisioning in Wireless Sensor Networks using Artificial Neural Network: A Survey', International Journal of Computer Application (IJCA), pp. 28-40.

[43] Mittal, M. and Kumar, K. 2016. 'Data Clustering In Wireless Sensor Network Implemented On Self Organization Feature Map (SOFM) Neural Network' IEEE international conference on Computing Communication and Automation(ICCCA).

[44] Mittal, M. and Bhadoria, R. S. 2017, 'Aspect of ESB with Wireless Sensor Network"', Exploring Enterprise Service Bus in the Service-Oriented Architecture Paradigm", IGI-global publications, pages 319.

[45] Miyoshi, T., Lien, G. Y., Satoh, S., Ushio, T., Bessho, K., Tomita, H., Nishizawa, S., Yoshida, R., Adachi, S. A., Liao, J. Gerofi, B.,Ishikawa, Y.,Kunii, M., Ruiz, J., Maejima, Y., Otsuka, M. Okamoto, K., Seko, H. 2016. 'Big Data Assimilation Toward Post-Petascale Severe Weather Prediction: An Overview and Progress', Proceedings of the IEEE, Vol. 104, Iss. 11, pp. 2155 – 2179

[46] Moyne, J., Samantaray, J., and Armacost, M. 2016. 'Big Data Capabilities Applied to Semiconductor Manufacturing Advanced Process Control', IEEE

Transactions on Semiconductor Manufacturing, Vol. 29, Iss. 4, pp. 283 – 291

[47] Myers, J. 2016. 'Master Data Management for Data Driven Organizations', an Enterprise Management Associates, pp. 1-12.

[48] O'Leary, D. E. 2016. 'Ethics for Big Data and Analytics, IEEE Intelligent Systems', Vol. 31, Iss. 4, pp. 81 – 84.

[49] Oneto, L., Bisio, F., Cambria, E., and Anguita, D. 2016. 'Statistical Learning Theory and ELM for Big Social Data Analysis', IEEE Computational Intelligence Magazine, Vol. 11, Iss. 3, pp. 45 – 55.

[50] Ordonez C and Omiecinski E. 2004. 'Efficient disk-based k-means clustering for relational databases', IEEE Trans. Knowl. DataEng, vol. 16, issue 8, pp. 909–921.

[51] Pan, E., Wang, D., and Han, Z. 2016. 'Analyzing Big Smart Metering Data Towards Differentiated User Services: A Sublinear Approach', IEEE Transactions on Big Data, Vol. 2, Iss. 3, pp. 249 – 261

[52] Paul, A., Ahmad, A., and Rathore M. M. 2016. 'Smartbuddy: defining human behaviors using big data analytics in social internet of things', IEEE Wireless Communications, Vol. 23, Iss. 5, pp. 68 – 74.

[53] Pei J, Han J and Mao R. 2000. 'CLOSET: an efficient algorithm for mining frequent closed itemsets', Proceedings of the ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, pp 21–30.

[54] Pei J, Han J, Asl MB, Pinto H, Chen Q, Dayal U and Hsu MC. 2001. 'PrefixSpan mining sequential patterns efficiently by prefix projected pattern growth', Proceedings of the International Conference on Data Engineering, pp. 215–226.

[55] Pol, U. R. 2014. 'Big Data and Hadoop Technology Solutions with Cloudera Manager', International Journal of Advanced Research in Computer Science and Software Engineering, Vol.4, Iss.11, pp. 1028-1034.

[56] Pradhananga, Y., Karande, S., and Karande, C. 2015).'CBA: Cloud-Based Bigdata Analytics', International Conference on Computing Communication Control and Automation (ICCUBEA), pp. 47-51.

[57] Rangra, K., and Bansal, K. L 2014. 'Comparative Study of Data Mining Tools', International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 6,pp. 216-223.

[58] Ranjan, R., Ranjan, J. and Bhatnaga, V. 2013, 'Critical success factor for implementing data mining in higher education: Indian perspective' Int. J. of Computational Systems Engineering, Vol.1, No.3, pp.151 – 161.

[59] Russom, P. 2013. 'Integrating Hadoop Into Business Intelligence and Data Warehousing', Tdwi Best Practices Report, pp. 1-38.

[60] Shao, W., Salim, F. D., and Song, A. 2016. 'Clustering Big Spatiotemporal-Interval Data, IEEE Transactions on Big Data, Vol. 2, Iss. 3, pp. 190 – 203.

[61] Singh, S., Firdaus, T., and Sharma, A. K. 2015. 'Survey On Big Data Using Data Mining', International Journal of Engineering Research and Development, Vol. 4, Iss. 4, pp. 135-143.

[62] Tawalbeh, L. A., Mehmood, R., Benkhlifa, E., and Song, H. 2016. 'Mobile Cloud Computing Model and Big Data Analysis for Healthcare Applications', IEEE Access, Vol. 4, pp. 6171 – 6180

[63] Ulfarsson, M. O., Palsson, F., Sigurdsson, J., and Sveinsson, J. R. 2016. 'Classification of Big Data With Application to Imaging Genetics', Proceedings of the IEEE, Vol.104, Iss. 11, pp. 2137 – 2154

[64] Ververidis D and Kotropoulos C. 2008 'Fast and accurate sequential floating forward feature selection with the bayes classifier applied to speech emotion recognition', Signal Process, vol. 88, pp. 2956–2970.

[65] Wang, L., Wang, G., and Alexander, C. A. 2015a. 'Big Data and Visualization: Methods, Challenges and Technology Progress' Digital Technologies, Vol. 1, pp. 33-38.

[66] Wang, L, Wang, G. and Alexander, C. A. 2015b. 'Natural language processing systems and Big Data analytics', Int. J. of Computational Systems Engineering, Vol.2, No.2, pp.76 – 84

[67] Wang, J., Wu, Y. , Yen, N., Guo, S., and Cheng, Z. 2016. 'Big Data Analytics for Emergency Communication Networks: A Survey', IEEE Communications Surveys & Tutorials, Vol. 18, Iss.3, pp. 1758 – 1778

[68] Wang, X., and He, Y. 2016. 'Learning from Uncertainty for Big Data: Future Analytical Challenges and Strategies', IEEE Systems, Man, and Cybernetics Magazine, Vol. 2, Iss. 2, pp. 26 – 31

[69] Wang, Y., Chen, Q., Kang, C., and Xia, Q. 2016. 'Clustering of Electricity Consumption Behavior Dynamics Toward Big Data Applications, IEEE Transactions on Smart Grid, Vol. 7, Iss. 5, pp. 2437 – 2447

[70] Wu, C., Chen, Y., and Li, F. 2016. 'Decision model of knowledge transfer in big data environment', China Communications, Vol. 13, Iss. 7, pp. 100 – 107

[71] Xiao, B., and Cheng, G. 2015. 'The Research of Teradata Data Warehouse Technology', International Conference on Computational Intelligence and Communication Networks (CICN), pp. 982-984.

[72] Xing, E. P., Ho, Q., Xie, P., and Dai, W. 2016. 'Strategies and Principles of Distributed Machine Learning on Big Data', Engineering, published by Elsevier, pp. 179-195.

[73] Yan X, Han J and Afshar R. 2003. 'CloSpan: mining closed sequential patterns in large datasets', Proceedings of the SIAM International Conference on Data Mining, pp. 166–177.

[74] Yu, S. 2016. 'Big Privacy: Challenges and Opportunities of Privacy Study in the Age of Big Data', IEEE Access, Vol. 4, pp. 2751 – 2763.

[75] Zaki M J.2001. 'SPADE: an efficient algorithm for mining frequent sequences', Mach Learn, vol. 42, pp. 31–60.

[76] Zaki M J and Hsiao C-J 2005. 'Efficient algorithms for mining closed itemsets and their lattice structure', IEEE Trans Knowl Data Eng., vol. 17, pp. 462–478.

[77] Zhang T, Ramakrishnan R and Livny M. 1996. 'BIRCH: an efficient data clustering method for very large databases', Proceedings of the ACM SIGMOD International Conference on Management of Data, pp 103–114.

[78] Zhang, X., Yi, Z., Yan, Z. Min, G., Wang, W., Elmokashfi, A., Maharjan, S., Zhang, Y. 2016. 'Social Computing for Mobile Big Data Computer', Vol. 49, Iss. 9, pp. 86 – 90

[79] Zhang, Y., Cao, T., Li, S., Tian, X., Yuan, L., Jia, H., and Vasilakos, A. V. 2016. 'Parallel Processing Systems for Big Data: A Survey', Proceedings of the IEEE, Vol. 104, Iss. 11, pp. 2114 – 2136.

[80] Zhao, Y. 2013. 'Research on MongoDB Design and Query Optimization in Vehicle Management Information System', Applied Mechanics and Materials, Vols. 246-247,pp.418-422.