Big Data Challenges and Solutions in Healthcare: A Survey

Prabha Susy Mathew and Anitha S. Pillai

Abstract The digitization of medical data, field of genomics and use of wearable sensors to monitor patient health are some of the factors that have dramatically increased the growth of Big Data in Health Care/Biomedicine. Big data in healthcare actually refers to electronic health data sets which are large and complex that is very difficult to manage with traditional/conventional data management tools and techniques. Big data analytics in healthcare is cumbersome not just because of its volume but also because of the diversity of data types and the speed at which it is generated and must be managed/analyzed. Rapid progress is to be made for analyzing this data and for gleaning new insights for making better informed decisions. There are unprecedented opportunities to use big data. The Health Care Industry should find methods to properly analyze this Big HealthCare Data generated and stored around the world each seconds in order to discover associations, understand the patterns and trends which will provide significant opportunities for real-time tracking of diseases, predicting disease outbreaks, to improve care, save lives and lower costs. Extraction, integration and analysis of heterogeneous, enormous and complex HealthCare data captured from various Electronic Health Care sources are a major challenge. New methods, applications and tools that are used by Healthcare industries, practitioners and researchers to tackle the big data challenges are discussed in this paper.

Keywords Big data · Big data analytics · Biomedicine · Electronic Medical Record (EMR) · Healthcare · Genomics

P.S. Mathew (✉) · A.S. Pillai (✉)
MCA, School of Computing Sciences, Hindustan University, Chennai, India
e-mail: prabhasm@hindustanuniv.ac.in

A.S. Pillai
e-mail: anithasp@hindustanuniv.ac.in

V. Snášel et al. (eds.), *Innovations in Bio-Inspired Computing and Applications*,
Advances in Intelligent Systems and Computing 424,
DOI 10.1007/978-3-319-28031-8_48

1 Background

In this 'Digital Age' there is tremendous growth in data in terms of volume, variety and velocity. According to Gartner, Big Data is "high-volume, high-velocity, and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making" [1].

Recent years have seen a remarkable change in how data has been handled across industries. Healthcare industry in specific generates a vast set of data. As the Healthcare industry is information intensive [2] it is even more important to get the right inference from the data to make effective decision [3, 4]. The Healthcare data these days are originated from multiple sources including mobile devices, sensors attached to patient bed, wearable sensors, social media, Internet of Things (IoT), Electronic Medical Records (EMR), claim data, medical images, Radio Frequency Identification Device (RFID) monitoring/tracking devices etc [5, 6]. Such 'Data Explosions' have led to challenges in managing and analyzing these data. To do a timely analysis of this massive patient dataset in real time is a challenge. If the big data is synthesized and analyzed properly to identify associations, patterns and trends then healthcare providers and other stakeholders in the healthcare system can get better insightful diagnoses and treatments, which in turn would result in higher quality care at lower costs. For Example the increase in the search term in Google such as 'Flu symptoms' or 'Flu Treatment' or availability of public data sets available in Canada such as FluWatch can be used to predict the outbreak of flu in a particular region thereby keeping the medical facilities ready for treating the patients [7].

In this paper systematic review of literature related to healthcare, the Big Data problems faced, as well as solutions adopted by healthcare industry to combat these issues are discussed.

2 Related Work

The traditional approaches used to store and analyze the data no longer provide effective solution for handling Big Data. There are several solutions like Hadoop, High Performance Computing, In-Memory Computing, and Cloud Computing that can effectively handle the massive datasets. The Open-source software has proven to be quite instrumental in handling the big data challenges. Hadoop and MapReduce play a significant role in processing large clinical data [7].

According to experimental studies conducted [8] the authors have identified that the conventional relational database does not support all operations required for interactive visualization and analytics. NoSQL as a data management solution provides more flexibility to model different complex EMR data and is scalable for handling large data as it can be used with distributed computing architecture.

Panahiazar et al. [9] investigates, Hortonworks Data Platform (HDP) in Mayo's health care systems that uses Hadoop-MapReduce framework to predict survival score of each heart failure patients using Pig to translate queries to a sequence of MapReduce jobs. A test ran to compare Pig with other tools like SQL revealed that SQL took more time while Pig based alternatives took considerably less time to run a query.

Raghupathi and Raghupathi have described the use of big data analytics and applications in healthcare to gain valuable insights from the clinical data, but mercurial advances in big data platforms and tools can accelerate their maturing process [4].

Bioinformatics study increasingly relies on high-performance computation and large-scale data storage. The datasets are often complex, heterogeneous and incomplete. Considering these aspect, visual techniques plays a vital role in bioinformatics. There are many powerful scientific toolsets available ranging from software libraries such as SciPy, Chimera, Taverna, Galaxy, and the Visualization Toolkit (VTK). Most of them are designed for small, local datasets and cannot handle recent advances in data generation and acquisition.

To resolve these big data challenges, Steven et al. developed a visual analytics software framework named DIVE—Data Intensive Visualization Engine. It is a data-agnostic, ontologically-expressive software framework capable of streaming large datasets at interactive speeds. The platform provides parallelized operations, high-throughput and structured data streaming [10].

3 Big Data Challenges in Healthcare

The Healthcare industry leverages Big Data and Analytics to increase effectiveness in terms of better clinical care, reduced operational cost and providing personalized patient care. However there are growing complexities due to the presence of huge amount of diverse Healthcare data [11, 12].

3.1 V's of Big Data Analytics

The V's of Big Data is predominant even in healthcare mainly Volume, Variety and Velocity and Veracity. Huge volume of healthcare data comes from Electronic medical records, clinical images, Diagnosis data and health claim data, etc. in structured, semi-structured and unstructured form in real time. Capturing and analyzing streaming data is a challenge. Sensors attached to the patient's bedside to continually track patient vitals produce huge chunks of data that the traditional systems were incapable of storing and analyzing effectively [13]. Changes in

pattern in these vital signs are alerted to a team of doctors and assistants. All this was successfully achieved using Hadoop ecosystem components [4]. Real time processing of monitoring data can mean the difference between life and death for a patient. Mitsui Knowledge Industry MKI uses SAP HANA and Hadoop for providing personalized cancer treatment based on analysis of one's DNA. The solution uses Hadoop to align the patient's DNA sequence with the normal sequence, as the data is in a semi-structured and parallelization across multiple machines is possible. Identifying the mutations and predicting the best treatment requires a lot of highly iterative analysis which is achieved by SAP HANA. As a result the analysis which took them 2–3 days earlier has reduced to 20 min now [6]. Advancement in data management, Virtualization, cloud computing and Big Data Ecosystem is facilitating easier and effective means to capture, store and manipulate large volumes of data. Today, enterprises are adopting NoSQL (Not Only SQL) technology as it overcomes the limitations of the traditional Relational Database technology. NoSQL provides a more flexible schema less data model which can efficiently store unstructured and semi structured data. It provides an easier and cost effective approach to database scaling [14]. Several Big Data challenges are as represented in the Fig. 1. In this paper some of the major big data challenges and approaches will be discussed.

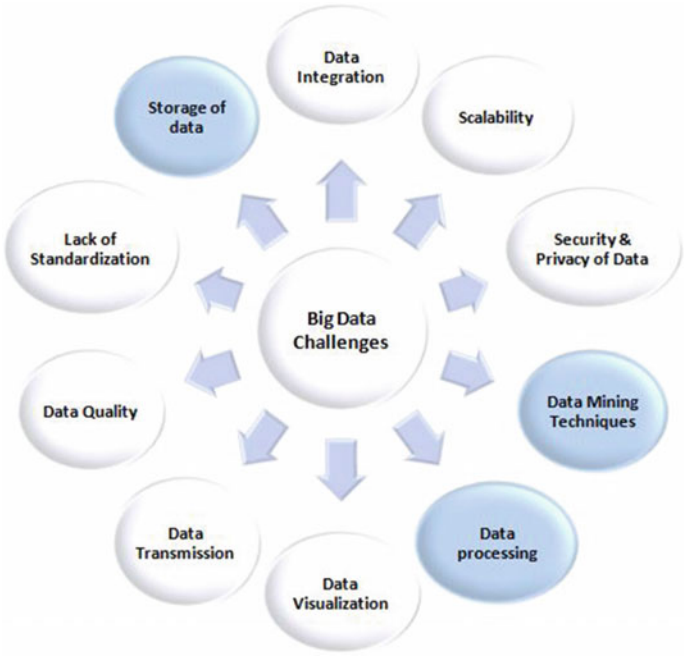

Fig. 1 Big data challenges

3.2 Data Storage

The Dominance of Relational database is slowly fading away with the rise of new type of databases. The relational database which uses strict schema based approach cannot efficiently store unstructured and semi structured data. The ever increasing volume of data causes the relational database technology performance to decline as it does not provide a scalable solution for handling the "big data" [14]. Dealing with Massive data sets requires large storage space. Two approaches to deal with it are Compression and Sampling [15]. By using Compression techniques time taken is generally more but less space will be utilized. While using sampling, information is lost by takes less space. Feldman et al. [3] in his work uses coresets which are small sets that approximate the original data for a given problem to reduce the complexity of Big Data with merge-and-reduce approach for problems such as K-means, PCA and Projective clustering. Cloud Computing can be used as a powerful assistance to store and process large amount of healthcare data. One of the commonly used solutions to handle this problem of big data is dimensionality reduction. Linear mapping methods, such as principal component analysis (PCA), singular value decomposition, as well as non-linear mapping methods, such as Sammon's mapping, laplacian eigenmaps and kernel principal component analysis have been used for dimensionality reduction [16].

3.3 Data Processing

Some optimal mechanism is required as near real time processing of information is the need of the hour. The In-Database processing and In-Memory computing technologies can be adopted by organizations to improve their processing speed. Many organizations are leveraging on hybrid transactional/analytical processing (HTAP) allowing transactions and processing to reside in the same in-memory database. Analytics with HTAP is much faster compared to the solutions already available [17].

Analyzing genomic data is a computationally intensive task and combining them with standard clinical data adds additional layers of complexity. This sort of data explosion has led to complexity in handling and analyzing data with respect to increasing volume, velocity and variety. The solution provided by Hadoop is extensively used to deal the big data problem, but Hadoop being a batch processing framework it does not cope with the need for real time analytics [18]. To deal with issues of the fundamental architecture, Nathan Marz came up with Lambda Architecture (LA) paradigm which is a scalable and fault tolerant data processing architecture. Lambda Architecture consists of batch layer, serving layer and speed layer that compute real time analytics to compensate for the slow batch layer. The LA is able to serve a wide range of workloads in which low-latency is required

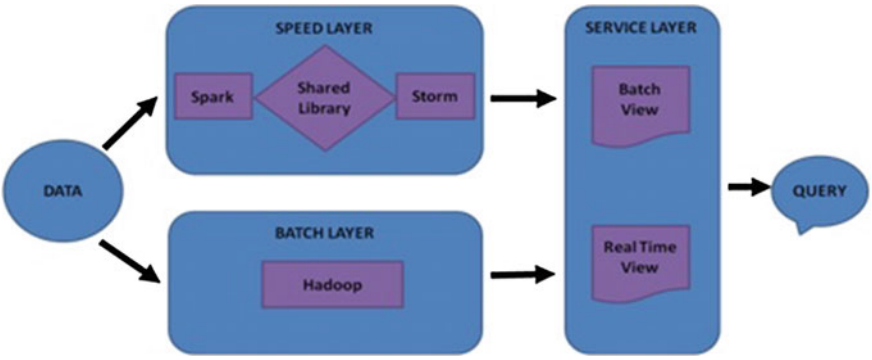

Fig. 2 Lambda architecture. (Based on Nathan Marz design)

[19]. Lambda Architecture is an approach that handles both batch and real time processing to achieve best of both world (Fig. 2).

Spark has the ability to address both batch and near real-time data processing and provides high performance using in-memory computing. It is a cluster computing framework that guarantees up to 100 times faster performance making it best suited for machine learning and graph data processing. It gives a comprehensive, unified framework to manage big data processing requirements with a variety of data sets that are diverse in nature (text data, graph data etc.) as well as the source of data (streaming data) [20]. Like Spark, Storm is also fault- tolerant, does distributed computation and has the ability to handle arbitrarily complex computation [21, 22].

3.4 Existing Algorithms Redesigned to Perform Big Data Mining

Many data mining techniques are not capable to carry out distributed mining. Data is ever evolving so the data mining techniques must be able to adapt to the changing needs. Efficient parallel algorithms and implementation techniques can greatly improve and address the performance and scalability requirements.

For example SVM classification is only directed to two-class tasks, so to solve multi-class tasks, algorithms that deal with reduction to several binary problems have to be applied. It suffers from scalability issues in both memory use and computational time, to overcome this issue Parallel SVM (PSVM) has been used [5]. Ke Xu et al. proposed PSVM based on iterative MapReduce for e-mail classification. They have used ontology based semantics to improve the accuracy of PSVM based on MapReduce. The use of MapReduce framework significantly reduces the training time. However, for data intensive mining purpose MapReduce in cloud computing proves to be effective than MapReduce in Hadoop [23].

Association Rule Mining algorithms are often used to find association between items in a dataset. The Apriori algorithm uses an iterative approach where results of previous iterations are used to find the frequent item sets for the current iteration. Main drawback with this approach is the candidate sets become large if the dataset is too huge and multiple data scans are required. Eclat algorithm deals with the issue of Apriori algorithm's multiple scan by making a single pass on the dataset there by making it a faster algorithm when compared to Apriori. However Eclat does not deal with the huge candidate set generated by Apriori, which is effectively handled by FP-Growth. FP-Growth eliminates the candidate generation step. It is faster than Apriori algorithm but as the size of data increases the efficiency of the algorithm drops. Thus to handle the issues relating to huge datasets parallel algorithms are introduced. Hence in order to scale across multiple machines fault tolerant framework for distributed application such as MapReduce is used. MapReduce based on Hadoop implementation of Apriori algorithm handle the large datasets in HDFS, for each iteration result is to be sourced from the HDFS which requires high I/O time leading to a reduced performance. Yet another frequent itemset Mining (YAFIM) Algorithm based on Spark RDD (Resilient Distributed Dataset) framework resolves the issues. The Apriori algorithm's implementation on spark platform speeds up to nearly 18 times on an average over different benchmarks. Results generated on real world healthcare data are observed to be many times faster when compared to MapReduce framework. Thus in an attempt to further improve the algorithm, Rathee et al. proposed Reduced-Apriori (R-Apriori) based on Spark RDD framework, it eliminates the time consuming part of the YAFIM algorithms second pass to further reduce the number of computations required thereby improving the speed [24].

Existing Clustering algorithms do not provide scalable solutions to handle the Big Data. A clustering method Fuzzy C-means based on the Google's MapReduce paradigm was proposed in [25]. The Experimental evidence of MapReduce-based Fuzzy C-Means Algorithm (MR-FCM) demonstrated that the algorithm scales well with increasing data sets. Another comparison between the two Mahout algorithms KMeans and Fuzzy KMeans (FKM) with MR-FCM for data set sizes varying showed that though FKM and MR-FCM are computationally quite similar, the Mahout FKM algorithm scales better than the MR-FCM algorithm.

High dimensionality data clustering methods are designed to handle data with hundreds of attributes, including DFT and MAFIA [26].

Algorithms such as SPRINT (Scalable Parallelizable Induction of Decision Tree algorithm) and SLIQ (Supervised Learning In Quest) are highly scalable, and has no storage constraint on larger data sets [26]. The drawback associated with SLIQ algorithm is that it requires computation of large number of Gini indices at each node of the decision tree to decide which attribute is to be split at each node. This computation is carried out for all attributes and for each successive pair of values. For huge datasets, lot of such computation is required. A performance enhancement to SLIQ algorithm, CC-SLIQ (Cascading Clustering and Supervised Learning In Quest) was proposed by Prasad et al. [27]. The CC-SLIQ algorithm constructs the binary decision tree by cascading two machine learning algorithms: k-means

clustering and SLIQ decision tree learning. The CC-SLIQ approach results in decision trees that have smaller sizes, fewer rules. It is proved to be useful in noisy environments when compared to standard methods.

A parallel version of the random forest using MapReduce programming framework on top of Hadoop has been developed [28] for large-scale population genetic association studies involving multivariate traits. The algorithm has been applied to a genome-wide association study on Alzheimer disease (AD) in which the quantitative characteristic consists of a high-dimensional neuroimaging phenotype describing longitudinal changes in human brain structure. Remarkable speed-ups in the processing were observed.

There are some limitations with the traditional association rule mining algorithms for large-scale data. The FP-Growth algorithm's success is limited by internal memory size because mining process is on the base of large tree-form data structure, which results in high computation time. The algorithm that constructs an Optimum pattern Tree with the node as the data item of the transaction has been proposed. This algorithm is implemented on Hadoop to reduce the computation cost and for handling the large data and processes them parallely to improve efficiency [29].

Traditional machine-learning frameworks such as Weka and Rapidminer do not scale to big data. Apache's Mahout is a framework which is open source and is a distributed framework to deal with big data. Mahout is a scalable machine-learning library used on top of Hadoop. Two approaches for performing machine learning algorithm on distributed framework using MapReduce are Mahout and Spark. In case of Mahout all iteration results are written to and read from disk while for Spark all iterations can be stored in memory. As processing of data can be done directly from memory there is a significant performance improvement when using Spark as opposed to Mahout [30].

3.5 Comparison Between MapReduce and Spark

Apache Spark is an open source framework for big data. It is a cluster computing framework that guarantees up to 100 times faster performance making it best suited for machine learning and graph data processing. It gives a comprehensive, unified framework to manage big data processing requirements with a variety of data sets that are diverse in nature (text data, graph data etc.) as well as the source of data (streaming data). The lack of speed and absence of in-memory queuing are considered to be the biggest drawback plaguing MapReduce. Apache Spark allows processing of data streams unlike MapReduce which processes data in batches causing queuing delays which are not acceptable in several real time applications. Real time studies conducted have proved Spark to sort 100 TB of data in just 23 min when compared to the 72 min taken by Hadoop to accomplish the same using a number of Amazon Elastic Cloud machines. Spark runs on Hadoop just the way MapReduce does but with the exception that MapReduce runs only on Hadoop

Table 1 MapReduce versus Spark

	MapReduce	Spark
Performance	Relatively slow	10 to 100 times faster than MapReduce engine
Processing	Majorly for batch processing	Streaming and batch processing
Compatibility	Hadoop	Mesos, YARN and Hadoop
Data store	Stores data on disk	Stores data in-memory
Implementation	Implemented using java	Using API's for Java, Scala and Python
Ease of use	Code is lengthy	Less word count in code compared to MapReduce
Failure tolerance	Slightly more failure tolerant	Less tolerant when compared to MapReduce
Security	More security features	Security aspects still in Infancy

while Spark on the other hand has the ability to run and exist without Hadoop. Spark also runs on Mesos, standalone, or in the cloud, making it the next big thing in data analytics [31].

Spark stores data in-memory whereas Hadoop stores data on disk. Hadoop uses replication to achieve fault tolerance whereas Spark uses different data storage model, resilient distributed datasets (RDD) that minimizes network I/O. It can access diverse data sources including HDFS, Cassandra, HBase, and S3 [20]. Spark permits writing applications in Java, Scala, or Python. It comes with a built-in set of over 80 high-level operators.

Some of the differences between MapReduce and Spark are highlighted in the table mentioned below [32, 33] (Table 1).

4 Conclusion

In Healthcare industry the medical professionals and patients are generating huge data for monitoring and improving patient. The biggest challenge faced by data scientists is how to integrate these diverse data and make use of this massive data effectively.

To overcome the Big Data challenge several new methods and technologies have been devised to resolve storage, processing and security related issues. Till date, the Hadoop ecosystem has proven to be the most mature framework for handling big data, but is restricted to batch processing. New technologies such as Storm, Spark, and Mahout are emerging to provide flexibility, support real-time processing and to run adhoc queries on large data sets. A major challenge for the next couple of years is to come up with robust implementations of analytical methods that are required for the biomedical and health domain, within these frameworks. This will require

transforming the existing analytical algorithms to fit into the distributed processing model.

Organizations must understand that big data solutions are not a replacement to the existing solutions but a complement to it for solving big data problems. It is essential for them to take an end to end solution leveraging from multiple big data and traditional solutions in order to obtain the desired Healthcare Outcome.

References

1. http://www.gartner.com/it-glossary/big-data
2. Chen, C.L.P, Zhang, C.-Y.: Data-intensive applications, challenges, techniques and technologies: a survey on Big Data. Inf. Sci. Elsevier (2014)
3. Feldman, D., Schmidt, M., Sohler, C.: Turning Big data into tiny data: Constant-size coresets for k-means, PCA and projective clustering. In: SODA '13 Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 1434–1453 (2013)
4. Raghupathi, W., Raghupathi, V.: Big data analytics in healthcare: promise and potential. Health Inf. Sci. Sys. http://www.hissjournal.com/content/2/1/3 (2014)
5. Yadav,C., Wang, S., Kumar, M.: Algorithm and approaches to handle large data—a survey. IJCSN Int. J. Comput. Sci. Netw. **2**(3) 2013. ISSN (Online):2237-5420
6. Jonker, D.: https://blogs.saphana.com/2013/05/09/saps-hadoop-strategy/
7. Wang, W., Krishnan, E.: Big Data and Clinicians: A Review on the State of the Science, vol. 2, no. 1 (2014)
8. Archenaa, J.: Big Data analytics for health care using hadoop. Int. J. Appl. Eng. Res. **9**(18), 3301–3308 (2014). Research India Publications, ISSN:0926-4513
9. Panahiazar, M., Taslimitehrani, V., Jadhav, A., Pathak, J.: Empowering personalized medicine with Big Data and semantic web technology: promises, challenges, and use cases. In: Proceedings of IEEE International Conference on Big Data, pp. 790–795, Oct 2014. doi:10.2409/BigData.2014.7004307
10. Kumar, V., et al.: Exploring clinical care processes using visual and data analytics: challenges and opportunities. http://dssg.uchicago.edu/kddworkshop/papers/kumar.pdf
11. Pratt, M.K.: No Quick Cure for Healthcare Systems Computerization is Slowly Improving the Healthcare System, But it's a Long Way From Living up to Expectations. Computer World, Healthcare IT
12. Mathew, P.S., Pillai, A.S.: Big Data Solutions in Healthcare: Problems and Perspectives. IEEE Xplore (2015). doi:10.2409/ICIIECS.2015.2293224
13. Banaee, H., Ahmed, M.U. Loutfi, A.: Data mining for wearable sensors in health monitoring systems: a review of recent trends and challenges. Sensors **13**(12), 12245–12900 (2013). doi:10.3390/s131212245
14. Sadalage, P.: NoSQL databases: an overview. https://www.thoughtworks.com/insights/blog/nosql-databases-overview (2014)
15. http://albertbifet.com/big-data-mining-tools
16. Hirak, K., Afzal, A.H., Nazrul, H., Swarup, R., Kumar, B.D.: Big Data analytics in bioinformatics: a machine learning perspective. J. Latex Class Files **13**(9) (2014)
17. http://www.computerworld.com/article/2690856/big-data/8-big-trends-in-big-data-analytics.html
18. Hausenblas, M., Bijnens, N., inspired by Marz, N.: Lambda Architecture (2015)
19. Laurent Bride.: Hadoop Summit 2015 Takeaway: The Lambda Architecture (2015)
20. Mohammed, J.: Is apache spark going to replace hadoop. http://aptuz.com/blog/is-apache-spark-going-to-replace-hadoop/ (2015)

21. Giamas, A.: Spark, Storm and Real Time Analytics (2014)
22. https://storm.apache.org/
23. Xu, K., Wen, C., Yuan, Q., He, X., Tie, J.: A MapReduce based parallel SVM for email classification. J. Netw. **9**(6), (2014)
24. Rathee, S., Kaul, M., Kashyap, A.: R-Apriori: an efficient apriori based algorithm on spark. In: PIKM'15, Melbourne, VIC, Australia. ACM, Oct 19 2015. ISBN:978-1-4503-3782-3/15/10. doi:http://dx.doi.org/10.1145/2809890.2809893
25. Ludwig, S.A.: MapReduce-Based Fuzzy C-Means Clustering Algorithm: Implementation and Scalability
26. Chen, F., Deng, P., Wan, J., Zhang, D., Vasilakos, A.V., Rong, X.: Data mining for the internet of things: literature review and challenges. Int. J. Distrib. Sens. Netw. **2015**, Article ID 431047, 14 (2015). http://dx.doi.org/10.1155/2015/431047
27. Narasimha Prasad LV., Naidu, M.M.: CC-SLIQ: performance enhancement with 2 K split points in SLIQ decision tree algorithm. IAENG Int. J. Comput. Sci. **41**(3), IJCS_41_3_02
28. Mohammed, E.A., Far, B.H., Naugler, C.: Applications of the MapReduce programming framework to clinical big data analysis: current landscape and future trend. BioData Min. **7**, 22 (2014). doi:10.1186/1756-0381-7-22
29. Shah, A.H., Patel, P.A.: Optimum frequent pattern approach for efficient incremental mining on large databases using MapReduce. Int. J. Comput. Appl. (0975–8887) **120**(4) (2015)
30. Aydin, G., Hallac, I.R., Karakus, B.: Architecture and implementation of a scalable sensor data storage and analysis system using cloud computing and Big Data technologies. J. Sens. **2015**, Article ID 834217, 11 (2015). http://dx.doi.org/10.1155/2015/834217
31. Suresh, R.: Apache spark and the future of big data analytics. http://suyati.com/apache-spark-and-the-future-of-big-data-analytics/ (2015)
32. https://www.xplenty.com/blog/2014/11/apache-spark-vs-hadoop-mapreduce/
33. http://www.dezyre.com/article/hadoop-mapreduce-vs-apache-spark-who-wins-the-battle/83