Contents lists available at www.innovativejournal.in

Asian Journal of Computer Science And Information Technology

Journal Homepage: http://innovativejournal.in/ajcsit/index.php/ajcsit

ISSN 2249 - 5126

BIG DATA ANALYTICSTOOLS, METHODS & FRAMEWORKS: A COMPREHENSIVE REVIEW

Mukesh Shukla, Dr. P.K. Rai, Rinku Singh

A. P. S. University, Rewa(MP) 486003, India

ARTICLE INFO

Corresponding Author:
Mukesh Shukla,
A. P. S. University, Rewa(MP)
486003, India
needmukesh@gmail.com

Keywords: Big Data, Map Reduce, Hadoop, KDD, ETL, Spark, Data Mining.

DOI:http://dx.doi.org/10.15520/ajcsit.v7i3.66

ABSTRACT

Big data refer to the collection of new information which must be made handy to high numbers of users close to real time, based on gigantic data inventories from multiple sources, with the goal of speeding up critical competitive knowledge discovery processes. Massive amounts of data have become accessible on hand to data miners which is making analysis and decision making task much more challenging and tedious. Considering the massive volume and variety of data, the analyses, predictive and behavioral exploration of situations and business intelligence workloads are beyond the capabilities of existing tools &methods. In recent years a number of Big Data tools & methods have been suggested to handle these massive quantities of data. The objective of this paper is to study and to get the in-depth understanding of the various attributes of big data science, engineering, tools &techniques. This study also analyze the several frameworks suggested by researchers and abilities of these frameworks to revolutionize knowledge discovery process for enhancing the decision making process. This objective is considered via wide ranging review of literature.

1. INTRODUCTION

Big Data refers to data that exceeds the typical processing, storage and computing capacity of traditional databases and techniques used for data analysis. One of the buzzwords in IT during the last few years is 'Big Data'. Organizations which had to handle the fast growth data, they initially shaped it for processing data resulting from scientific or business simulations, web data or data from other sources. Fundamental business model of some of those companies are rely on indexing and using this large amount of data. Google developed the the Google File System and MapReduce for handling the sheer volume data available on web. These technologies are available as open source software as Apache Hadoop and the Hadoop File System. All these efforts laid the foundation for technologies summarized today as 'big data'. Later big giants IBM Oracle, HP, Microsoft, SAS and SAP in information management field stepped in and invested to extend their business and build new products especially aimed at Big Data analysis. At the same time many start-ups like Cloudera entered the scene. Considering the trends analysts expectBig Data impact onto business and the praise they sing on 'big data', it was obvious for these big players to get part of the big data.

As per the IDC predictions digital data created and consumed per year will grow up to 40.000 exabyte by 2020, from which a third 2 will potentially valuable to organizations if processed using big data technologies. IDC also declared that in 2012 only 0.5% of potentially valuable data were examined, calling this the 'Big Data Gap'. McKinseyGlobal Institute also predicts 40% annual growth in global data per year. They describe big data trends in terms of monetary figures and see Big Data market of 300 billion $ in US health care sector and 250 billion in European public sector and a potential improvement of margins in the retail industry by 60%.

As a resource, there is need of Big Data tools and methods that can be used to analyze and extract patterns from large-volume data. Increased computational processing power, increased data storage capabilities and availability of increased volumes of data are the major features of Big Data. The size, variety, and rapid change of such data require a new type of big data analytics, as well as different storage and analysis methods. Such large amounts of data need to be properly analyzed, and pertaining information should be extracted.

The contribution of this paper is to offer an analysis of the available literature on Big data analytics concepts, frameworks, methods and available implementing tools. This study presents comprehensive survey of big data attributes with discussion of some of the various big data tools, methods, and frameworks which can support future need of discovering knowledge from massive data.

2. Definitions, Tools &Methods

Academicians, Industry R&D experts and other prominent stakeholders positively agree that big data has become a big game changer in most in most of the knowledge discovery process over the last few years. With

this in mind, having a bird's eye view of big data and its frameworks implemented for different areas help us better appreciate what will be the direction and trends of future research across different domains. In this section, we discuss literatures which examine different tools, techniques that are available for big data analytics tosolve domain specific challenges.

In a 2011 Gartner report [36] Doug Laney explains the concept of Volume, Variety and Velocity in data management. These are known as the 3V's and characterize the concept of Big Data.In 2012, Gartner revised and provided a more detailed definition "Big Data are high-volume, high-velocity, and/or high-variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization. More generally, a data set can be termed Big Data if it is difficult to perform capture, analysis, curation and visualization on it at the present technologies."

Ruixan [34] presents Bibliometrical Analysis on the Big Data Research in China and summarizes research characteristics in order to study Big Data in-depth development and the future development of Big Data. They also suggest reference information for studies related to Library and Information Studies. They find that research based on Big Data has taken shape though most of these papers in the theoretical stage of exploration, lack adequate practical support and therefore recommend intensifying efforts based on theory and practice.

Fernández et al. [26] focus on systems for large-scale analytics based on the MapReduce scheme and Hadoop. They find several libraries and software projects that have been developed for aiding experts to address this new programming model. They also examine the benefits and drawbacks of MapReduce, in contrast to the classical solutions in this field. They also suggest a number of programming frameworks that have been proposed as an alternative to MapReduce, developed to solve the limitations of this model in certain scenarios and platforms.

Polato et al. [29] have conducted a systematic literature review to assess research contributions to Apache Hadoop. The goal of their study was to identify gaps, providing motivation for new research, and outline collaborations to Apache Hadoop and its ecosystem, categorizing and quantifying the main topics addressed in the paper.

Wu and Yamaguchi [32] presents a survey of Big Data in life sciences, Big Data related projects and Semantic Web technologies. Theirstudy helps to understand the role of Semantic Web technologies in the Big Data era and how they provide a promising solution in life science big data analytics.

Kambatla et al. [28] provide an overview of the state-of-the-art and focus on emerging trends to highlight the hardware, software, and application landscape of big-data analytics.

Hashem et al. [12] have assessed the rise of big data in cloud computing. In their study they provide the definition, features, and grouping of big data along with some discussions on cloud computing. They also discussed the relationship between big data and cloud computing, big data storage systems, and Hadoop technology. In their study, various research challenges are examined, with focus on scalability, accessibility, data integrity, data transformation, data quality, data heterogeneity, privacy, legal and regulatory issues, and governance.

Chen and Zhong [10] have done a comprehensive survey of Big Data technologies, techniques, challenges and applications. They provide a close view of Big Data applications opportunities and challenges as well as techniques that is currently adopted and used to solve Big Data problems.

Gandomi and Haider [27] present a consolidated description of Big Data by integrating definitions from practitioners and academics. Their study focuses on the analytic methods used in Big Data. Significance of this paper is its focus on analytics related to unstructured data, which according to these authors constitute 95% of Big Data.

Wang and Krishnan [31] present a review with an objective to provide an overview of the features of clinical Big Data. They explain a few generally employed computational algorithms, statistical methods, and software tool kits for data manipulation and analysis, and discuss the challenges and limitations in this realm.

Tsai C. et al. [23] reviewed studies on the data analytics from the traditional data analysis to the recent big data analysis. They used the KDD process as the framework for these studies and summarizedit into three parts: input, analysis, and output. Their discussion is focused on the performance-oriented and results-oriented issues from big data analytics perspective. This paper provides a brief introduction to the data and big data mining algorithms from the Perspective of mining problem, To better know the value added by the big data, this paper is focused on the data analysis of knowledge discovery in databases from the platform/framework to data mining.

Zuech R. et al. [25] studied the problem of heterogeneous data and in particular Big Heterogeneous Data. They discussed the particular issues of Data Fusion, Intrusion Detection, and Security Information and Event Managementsystems, along with the discussion on areas where more research opportunities exist.

Paper presented by Singh D. et al. [30] surveys different h/w platforms available for big data analytics and assesses the gains and downsides of each of these platforms based on various metrics like scalability, I/O rate, fault tolerance, real-time processing, data volume supported and iterative task support. Other than the hardware, a detailed portrayal of the software frameworks used within each of these platforms is also discussed along with their strengths and shortcomings. Some of the critical features described in their paper can potentially help people for making an informed decision about the right choice of platforms depending on their needs. A rigorous qualitative comparison between different platforms is also explainedusing a star ratings tablefor each of the six features that are critical for the algorithms of big data analytics. In order to provide more insights into the value of each of the platform in the context of big data analytics, detailed implementation level details of the widely used k-means clustering algorithm on various platforms are also described in the form pseudo code.

Sakr et al. [35] provide a comprehensive survey for a family of approaches and mechanisms of large-scale data processing mechanisms that have been implemented based on the original idea of the MapReduce framework and are currently gaining a lot of attention in both

research and industrial communities. They also suggest a set of introduced systems that have been implemented to provide declarative programming interfaces on top of the MapReduce framework. They evaluate several large-scale data processing systems that resemble some of the ideas of the MapReduce framework for different purposes and application scenarios.

3. Big Data Frameworks for Knowledge Discovery using data mining techniques

Kaur A. et al. [14] presented SUBSCALE, a novel clustering algorithm to find non-trivial subspace clusters with minimal cost and it requires only k database scans for a k-dimensional data set. Their algorithm scales very well with the dimensionality of the dataset and is highly parallelizable. They experimented with upto 6144 dimensional data and safely claimed that it will work for larger datasets too by adjusting the split factor.The main cost in SUBSCALE algorithm is the computation of the candidate one dimensional dense units. SUBSCALE supports a high degree of parallelism as there is no reliance in computing dense units across multiple scopes.

Yang Y. et al. [24] presented a 2D approach for visualizing big hierarchical data, called Cabinet Tree. Cabinet Tree uses the enclosure and orthogonal drawing methods and performs a space-optimized layout for leaves and explicit branches with carefully designed color schemes for appealing and rich visualization. Some features like Color-coded sorting, contrast-enhanced color strategy and labeling techniques all make full usage of display properties. Cabinet Tree also providesuninterrupted node selection using the mouse wheel and focus context view using the detail window. They providequantitative evaluations which show that CabinetTree is capable of visualizing large datasets. It is predicted that with greater screen resolutions, trees of hundreds of millions of nodes can be visualized on a single display. Cabinet tree is highly scalable for increased resolutions and data volumes and high layout speed and it is an effective tool for visualizing huge hierarchical structures in a wider range of applications.

Liu X. et al. [17] presented a Meta-MapReduce algorithm MMR implemented with MapReduce. Their algorithm tackles the difficulties for supporting iterations in Hadoop. Theirinvestigation results display that the error rates of MMR are lesser than the results of a single node on 9 out of 11 datasets. The comparison between MMR and the parallelized AdaBoost.PL algorithm displays that PML has lesser error rates than AdaBoost.PL on 7 out of 8 datasets. The speedup performance of MMR evidences that MapReduce improves the computation complexity substantially on large datasets. Proposed MMR algorithm has the ability to reduce computational complexity significantly, while producing slighter error rates.

Husain S. et al. [13] developed a mechanism to integrate dispersed multi-source data and service the mashed information via human and machine interfaces in a secure, scalable manner. Their mechanismsimplifies the exploration of subtle associations between variables, population strata, or clusters of data elements, which may be impervious to standard independent inspection of the different sources. They developed a new platform includes a device agnostic tool for graphical querying, navigating and exploring the multivariate associations in complex diverse datasets. Their article explains this core functionality and service oriented infrastructure using healthcare data as well as Parkinson 's disease neuroimaging data.

Depeige A. et al. [11] introduced a new framework for knowledge management in cloud computing environment, explained its specificities, as well as functional principles, and studied the role of knowledge analytics to drive the application of, and support this framework. Their study contributes to enhance our understanding of knowledge management practices in cloud computing environment, by identifying K-analytics that enable to support the Actionable Knowledge as a Service Framework.

Lourenço J. et al. [18] created a concise and up-to-date comparison of NoSQL engines, recognized their most favorable use case scenarios from the software engineer point of view and the quality features that each of them is most appropriate to. They discussed main features and types of NoSQL technology while approaching different sides that highly pay to the use of those systems. They also provided the state of the art of non-relational technology by discussing some of the most significant studies and performance tests.

Olshannikova E. et al. [20] found relevant Big Data Visualization methods classification and have suggested the modern inclination towards visualization-based tools for business support and other noteworthy fields. Past and current states of data visualization were explained and supported by analysis of benefits and drawbacks. The approach of utilizing VR, AR and MR for Big Data Visualization is offered and the benefits, drawbacks and potential optimization strategies of those are discussed. They explained the promising utility of Mixed Reality technology mixing with applications in Big Data Visualization. Placing the most vital data in the central area of the human visual field in Mixed Reality would allow one to achieve the presented information in a short period of time without substantial data losses due to human perceptual issues. Furthermore, they studied the impacts of innovative technologies, such as Virtual Reality displays and Augmented Reality helmets on the Big Data visualization as well as to the grouping of the main challenges of integrating the technology.

Trifunovic N. et al. [9] presented the Maxeler Application Gallery project, its vision and mission, as well as a selected number of examples using a uniform template, which enables an easy comprehension of the new dataflow paradigm underneath. Solution suggested by them could be used in education, in research, in development of new applications, and in signifying the advantage Application Gallery creates a number of synergistic effects.

Khalilian M. et al. [4] presented DCSTREAM method using the vector model and k-Means divide and conquer approach. Experimental results presented by them display that DCSTREAM can achieve richer quality and performance than STREAM and ConStream methods for abrupt and gradualreal world datasets. Their study display that the use of batch processing in DCSTREAM and ConStream is time consuming compared to STREAM but it avoids further analysis for spotting outliers and novel micro-clusters.

Silva B. et al. [21] performed parameter sensitivity analysis and experiments show that UbiSOM outperforms existing solutions in continuously modeling possibly non-stationary data streams, converging quicker

to stable models when the underlying distribution is stationary and responding accordingly to the nature of the change in continuous real world data streams.

Pirouz M. et al. [6] presented an algorithm called optimized relativity search to reduce the number of nodes in a graph when attempting to decrease the running time for personalized page rank estimation. In their paper, the weighted page rank method was mixed with the Monte-Carlo technique and a local update algorithm over a reduced map space; this algorithm was developed to obtain a more precise and quicker search method than FAST PPR. The experimental results displayed that for nodes with a high degree of input nodes, the speed of estimation was twice than FAST PPR, at the expense of a little precision.

Steininger T. et al. [8] introduced d2o, a Python module for cluster-distributed multi-dimensional numerical arrays. This module provides a layer of abstraction between the algorithm code and the data distribution logic. The main objectiveof their study toobtain usability without dropping numerical performance and scalability. Its global interface is similar to the one of a numpyndarray, whereas the cluster node's local data is directly available for use in customized high-performance modules. They developed this module in Python which makes it portable and easy to use and maintain.

Brahim M. et al. [1] designed a framework to tackle the spatial data retrieval within Cassandra NoSQL database by extending the CQL with spatial queries. He defined a CQL-like syntax to allow spatial functions while keeping the native CQL query syntax. The proposed framework is tested against various data set volumes. The experimental results confirm the efficiency of using an aggregation algorithm in order to moderate the number of queries sent to the cluster and avoid making hot-spot nodes, despite its additional cost in terms of run time. The significance of paralleling queries in non-blocking way to avoid needless idle time is also highlighted through the performance results. The suggested framework displayed the feasibility of approach where basic spatial queries are underpropped and the query response time is decreased by up to 70 times for a fairly large area.

Selim H. et al. [7] have proposed in their paper a fast method and an algorithm for the approximation and reduction of a large network and a fast estimate calculation for the shortest path, in addition to previously proposed algorithms for distance estimation. They discussedasimple yet powerful approach to the graph data that will operate on the collected similarity graph data that is computed by the reduction algorithm and obtaining at query execution time that beat's up the traditional Dijkstra's shortest path algorithm with large datasets.

Pääkkönen P. [5] Performed feasibility analysis of technologies for stream-based processing of semi-structured data. He did feasibility analysis of a Big Data management system for semi-structured data compared to Spark streaming, which has been mixed with Cassandra NoSQL database for persistence. They did tweet analysis in Eucalyptus cloud computing environment on a distributed shared memory multiprocessor platform. The experimental results show that AsterixDB improves performance significantly both in terms of throughput and latency. Data feed functionality of AsterixDB is effective when stream processing has been implemented with Java. AsterixDB also scaled on the same level or better, when the volume of nodes on the cloud platform was increased.

Liehr A. [15] introduced the concept of process evolution functions and event reduction policies, which allow for the time resolved visualization of an unlimited number of concurrent workflows by means of aggregated task views. The theoretical foundation of this concept is applicable for workflows represented by DAG(directed acyclic graphs). It is described on the basis of a simple IO-workflow model, which is normally found for distributed resource management systems used for many-task computing.

The task of selecting machine learning tools for big data tedious and time consuming. The available tools have benefits and drawbacks, and many have overlapping uses.Landset S. et al. [16] in their paper provides a list of criteria for making selections along with an analysis of the benefits and downsides of each. They discussed the gains and drawbacks of three different processing models along with a comparison of engines that implement them, including MapReduce, Spark, Flink, Storm, and H2O. They observed at machine learning libraries and frameworks including Mahout, MLlib, SAMOA, and assess them based on criteria such as scalability, ease of use, and extensibility. They conclude that there is no single toolkit that truly embodies a one-sizefits-all solution.

Nagwani N. K. et al. [19] provided a novel framework based on MapReduce technology for summarizing large text collection. The suggested technique is designed using semantic similarity based clustering and topic modeling using Latent Dirichlet Allocation (LDA) for briefing the large text collection over MapReduce framework. The briefing task is done in four phases and offers a modular implementation of multiple documents summarization. They assessedthis technique in terms of scalability. Different text summarization parameters like, compression ratio, retention ratio, ROUGE and Pyramid score are also used for evaluation. The advantages of MapReduce framework are clearly visible from the experimental result and it is also confirmed that MapReduce provides a faster implementation of summarizing large text collections and is a powerful tool in Big data Text analysis.

Trifunovic N. et al. [22] in their paper discusses the shift in the computing paradigm and the programming model for Big Data problems and applications. They evaluate Dataflow and Control Flow programming models through their quantity and quality characteristics. They offered a new methodology for benchmarking, which not only use the execution time, but also the power and space, needed to accomplish the task. Their research displays that if the TOP500 ranking was based on the new performance measures, Dataflow machines would outperform Control Flow machines. To support the above claims, they presented 8fresh implementations of different algorithms using the Dataflow paradigm, which displaysubstantial speed-ups, power reductions and space savings over their implementation using the Control Flow paradigm.

Firmani D. et al. [3] discussed the application of concept of data quality to big data by highlighting how much complex is to define it in a general way. It is difficult to characterize multidimensional data quality concept, even in the case of well-structured data. Big data add two more dimensions of complexity:being"very" source specific and being highly unstructured and schema-less.

After discussing data quality in traditional contexts, they study big data by providing insights into the UNECE classification, and then, for each type of data source, they choose a specific instance of such a type) and descibed how quality dimensions can be defined in these cases.

Cui B. et al. [2] presented a high-level abstraction system on Storm, called POS. The proposed system delivers a Pig Latin-like language on top of the Storm execution engine. Developers can code POS program and compiled executable code run over Storm.

Gorodov and Gubarev [33] have done a review of methods for visualizing data and provided a classification of visualization methods in application to Big Data.

4. Conclusions

In this paper, we have studied the pioneering topic of big data analytics, in recent time which has gained lots of interest due to its perceived exceptional opportunities and benefits. Industry influencers, academicians, and other prominent stakeholders undoubtedly agree that big data has become a big game changer in most, if not all, types of modern industries over the last few years. As big data continues to permeate our daily lives, there has been a substantial shift of focus from the propaganda surrounding it to finding real value in its use. Our study provides an analysis of the big data analytics concepts which are being explored, as well as their importance to knowledge discovery. Consequently, some of the big data analytics tools, frameworks and methods in particular were evaluated. In addition, we have examined different framework which are proposed to do knowledge discovery using various mining techniques on Big Data platform.

In our study we also found that various open source tools and techniques are available for big data analytics framework implementations and existing data mining algorithms are being integrated within these frameworks. The future scope of research work is to explore most effective framework for addressing various issues like performance, quality, feasibility, scalability etc. Also scope of future research is to integration of leading data mining, ETL & Business Intelligence tools with big data analytics tools.

5. References

1. Brahim Mohamed Ben, DriraWassim, Filali Fethi, Hamdi Noureddine.:Spatial data extension for Cassandra NoSQL database.Big Data (2016) 3:11 DOI 10.1186/s40537-016-0045-4
2. Cui Bin · Jiang Jie · Huang Quanlong · Xu Ying · Gui Yanjun · Zhang Wenyu.:POS: A High-Level System to Simplify Real-Time Stream Application Development on Storm.Data Sci. Eng. (2016) 1(1):41–50 DOI 10.1007/s41019-015-0002-9.
3. Firmani Donatella · Mecella Massimo · Scannapieco Monica · Batini Carlo.:On the Meaningfulness of "Big Data Quality".Data Sci. Eng. (2016) 1(1):6–20 DOI 10.1007/s41019-015-0004-7.
4. Khalilian1Madjid * , Mustapha Norwati and Sulaiman Nasir.:Data stream clustering by divide and conquer approach based on vector model.Big Data20163:1DOI: 10.1186/s40537-015-0036-x
5. Pääkkönen Pekka.:Feasibility analysis of AsterixDB and Spark streaming with Cassandra for stream-based processing. Big Data (2016) 3:6 DOI 10.1186/s40537-016-0041-8
6. Pirouz Matin, Zhan Justin.:Optimized relativity search: node reduction in personalized page rank estimation for large graphs.Big Data (2016) 3:12 DOI 10.1186/s40537-016-0047-2
7. Selim Haysam and Zhan Justin.:Towards shortest path identification on large networks.Big Data (2016) 3:10 DOI 10.1186/s40537-016-0042-7
8. Steininger Theo, Greiner Maksim,Beaujean Frederik, Enßli Torsten.:d2o: a distributed data object for parallel high-performance computing in Python.Big Data20163:17,DOI: 10.1186/s40537-016-0052-5
9. TrifunovicNemanja, Veljko Milutinovic *, Nenad Korolija and Georgi Gaydadjiev.:An AppGallery for dataflow computing.Big Data (2016) 3:4 DOI 10.1186/s40537-015-0038-8
10. Chen, J., Ma, J., Zhong, N., Yao, Y., Liu, J., Huang, R.,Li, W., Huang, Z., Gao, Y., Cao, J.: Waas: Wisdom as a service. IEEEIntelligentSystems 29(6),40–47(2015). DOI 10.1109/MIS.2014.19
11. Depeige Audrey * and Doyencourt Dimitri.:Actionable Knowledge As A Service (AKAAS): Leveraging big data analytics in cloud computing environments.Big Data (2015) 2:12 DOI 10.1186/s40537-015-0023-2.
12. Hashem, I.A.T., Yaqoob, I., Anuar, N.B., Mokhtar, S., Gani, A., Khan, S.U.: The rise of "big data" on cloud computing:Review and open research issues. Information Systems 47, 98–115 (2015). DOI 10.1016/j.is.2014.07.006
13. Husain Syed S, Kalinin Alexandr, Truong Anh and Dinov Ivo D *.:SOCR data dashboard: an integrated big data archive mashing medicare, labor, census and econometric information.Big Data (2015) 2:13 DOI 10.1186/s40537-015-0018-z.
14. Kaur Amardeep * and Datta Amitava.:A novel algorithm for fast and scalable subspace clustering of high-dimensional data. Big Data (2015) 2:17 DOI 10.1186/s40537-015-0027-y.
15. Kempa-Liehr Andreas.:Performance analysis of concurrent workflows.Big Data (2015) 2:10 DOI 10.1186/s40537-015-0017-0.
16. Landset Sara, Khoshgoftaar Taghi M., Richter Aaron N. * and Hasanin Tawfiq.:A survey of open source tools for machine learning with big data in the Hadoop ecosystem.Big Data (2015) 2:24 DOI 10.1186/s40537-015-0032-1
17. Liu Xuan *, Wang Xiaoguang, Matwin Stan and Japkowicz Nathalie.:Meta-MapReduce for scalable data mining.Big Data (2015) 2:14 DOI 10.1186/s40537-015-0021-4.
18. Lourenço João Ricardo * ,Cabral Bruno, Carreiro Paulo, Vieira Marco and Bernardino Jorge.:Choosing the right NoSQL database for the job: a quality attribute evaluation. Big Data (2015) 2:18 DOI 10.1186/s40537-015-0025-0.
19. Nagwani N K.:Summarizing large text collection using topic modeling and clustering based on MapReduce framework.Big Data (2015) 2:6 DOI 10.1186/s40537-015-0020-5.
20. Olshannikova Ekaterina *, Ometov Aleksandr, Koucheryavy Yevgeni and Olsson Thomas.:Visualizing Big Data with augmented and virtual reality: challenges and research agenda.Big Data (2015) 2:22 DOI 10.1186/s40537-015-0031-2.
21. Silva Bruno * and Marques Nuno Cavalheiro.:The ubiquitous self-organizing map for non-stationary data

streams.Big Data20152:27DOI: 10.1186/s40537-015-0033-0
22. Trifunovic Nemanja *, Milutinovic Veljko, Salom Jakob and Kos Anton.:Paradigm Shift in Big Data SuperComputing: Dataflow vs. ControlFlow.Big Data (2015) 2:4 DOI 10.1186/s40537-014-0010-z.
23. Tsai Chun-Wei, Lai Chin-Feng, Chao Han-Chieh and Vasilakos Athanasios V. *.:Big data analytics: a survey.Big Data20152:21DOI: 10.1186/s40537-015-0030-3
24. Yang Yalong, Zhang Kang, WangJianrong * and Nguyen Quang Vinh.:Cabinet Tree: an orthogonal enclosure approach to visualizing and exploring big data. Big Data (2015) 2:15 DOI 10.1186/s40537-015-0022-3.
25. Zuech Richard * , Khoshgoftaar Taghi M and Wald Randall.:Intrusion detection and Big Heterogeneous Data: a Survey.Big Data (2015) 2:3 DOI 10.1186/s40537-015-0013-4.
26. Fernández, A., del Río, S., López, V., Bawakid, A., del Jesus, M.J., Benítez, J.M., Herrera, F.: Big data with cloud computing: An insight on the computing environment,mapreduce, and programming frameworks 4, 380– 409– (2014). DOI 10.1002/widm.1134/abstract
27. Gandomi, A., Haider, M.: Beyond the hype: Big data concepts, methods, and analytics 35, 137–144– (2014)
28. Kambatla, K., Kollias, G., Kumar, V., Grama, A.: Trends in big data analytics 74, 2561–2573– (2014)
29. Polato, I., Ré, R., Goldman, A., Kon, F.: A comprehensive view of hadoop research - a systematic literature review 46,1–25– (2014)
30. Singh Dilpreet and Reddy Chandan K.:A survey on platforms for big data analytics.Big Data20142:8 DOI: 10.1186/s40537-014-0008-6.
31. Wang, W., Krishnan, E.: Big data and clinicians: A review on the state of the science 16, – (2014)
32. Wu, H., Yamaguchi, A.: Semantic web technologies for the big data in life sciences 8, 192–201– (2014)
33. Gorodov, E.Y., Gubarev, V.V.: Analytical review of data visualizationmethodsin application tobigdatapp. – (2013)
34. Ruixian, Y.: Bibliometrical analysis on the big data research in china 11, 383–390– (2013)
35. Sakr, S., Liu, A., Fayoumi, A.G.: The family of mapreduce and large-scale data processing systems 46, – (2013)
36. Laney, D.: 3D data management: Controlling data volume,velocity, and variety. http://blogs.gartner.com/douglaney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf