Major Big Data Challenges in Most Industries and Innovative Solutions

Shanmugasundaram Palanimalai, R.Velusamy, P.Vijaykumar

Abstract: ***The term "Big data" refers to "the high volume of data sets that are relatively complex in nature and having challenges in processing and analyzing the data using conventional database management tools". In the digital universe, the data volume and variety that, we deal today have grown-up massively from different sources such as Business Informatics, Social-Media Networks, Images from High Definition TV, data from Mobile Networks, Banking data from ATM Machines, Genomics and GPS Trails, Telemetry from automobiles, Meteorology, Financial market data etc. Data Scientists confirm that 80% of the data that we have gathered today are in unstructured format, i.e. in the form of images, pixel data, Videos, geo-spatial data, PDF files etc. Because of the massive growth of data and its different formats, organizations are having multiple challenges in capturing, storing, mining, analyzing, and visualizing the Big data. This paper aims to exemplify the key challenges faced by most organizations and the significance of implementing the emerging Big data techniques for effective extraction of business intelligence to make better and faster decisions.***

Keywords: Big Data, Hadoop, HDFS, MapReduce, No-SQL

I. INTRODUCTION

In the Big data era, the inordinate growth of data on its volume, variety and velocity, significantly raises the importance of handling Big data with an innovative approach to reveal the new insights and make use of it, for better and faster decisions. The largest retail companies and social-networks such as Wal-Mart, Google, Amazon, Face book and Twitter deal with millions of Petabyte and Exabyte data today in mounting prevalence, that impulsively demands the attention on the research and exploration to handle the Big data magnificently. It is quite fair that the emerging and advancement technologies helps organizations to analyze the Big data and extract better value to run more efficiently and profitably. In this paper, we have exemplified the various challenges faced by typical organizations in Big data space. We also have attempted to provide an outline and recommendations on some of the salient Big data techniques to perform higher optimized data analysis to extract value.

Revised Manuscript Received on December 09, 2019.
***** Correspondence Author
Dr.Shanmugasundaram Palanimalai *, Project Manager – Information Technology, Bangalore, India. Email:shan.palanimalai@gmail.com
R.Velusamy, Research Scholar, Bharathiar University Arts & Science College, Mudakkuruchi, Erode, India. Email: velusamy.msc75@gmail.com
Dr.P.Vijaykumar, AP, Bharathiar University Arts & Science College, Mudakkuruchi, Erode, India. Email Email: vijaykumar.msc78@gmail.com

II. BIG DATA CHALLENGES

In below, we have illustrated the key challenges identified by most organizations in BIG data phenomenon. Big data is measured by its magnitudes such as Volume, Varity, Velocity and Veracity. These factors are generally used to measure the aspects of Big data, and understand the nature of data & how it is exploited. These 4Vs are continued to be the challenges in Big data processing and analytics [1] [21] [14] [2] [12] [5]. The main factors to portray the components and comprehend nature of Big data are described in the below Fig-1.

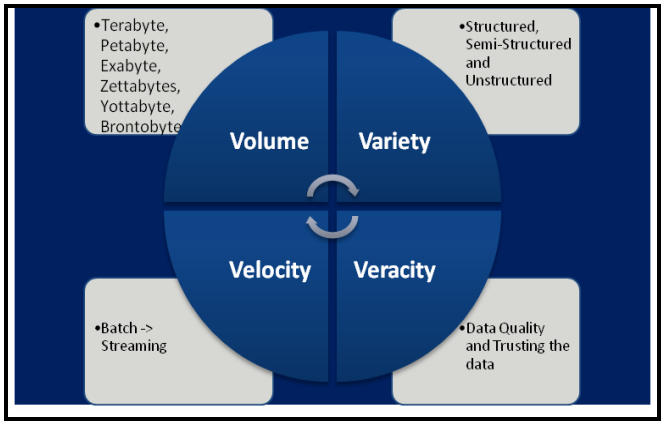

Fig.1. Big Data - 4V Challenges

A. Volume

Today, the volume of data that we deal with is growing towards Petabyte and Exabyte due to multiple factors. The unstructured data from the large internet companies & social network, also the growing demand for the sensors and M2M data (Machine-to-Machine) significantly contributes to the growth in data volume. With the enormous data growth, it is obligatory for any organization to analyze data volume growth trend for reducing the storage cost, create innovations and value from the relevant data which would fit best for the organizations. Managing large data with the enormous growth trend becomes one of the key challenges for most of the organizations. As result of the high volume of data growth, organizations need to focus more on their architecture and infrastructure quite often since the growth trend is unpredictable. It also leads to complexity in data management perspective. The below graph shows, the Digital Universe data growth trend in Exabyte based on IDC survey. [1] [17] [5] [7] [11].

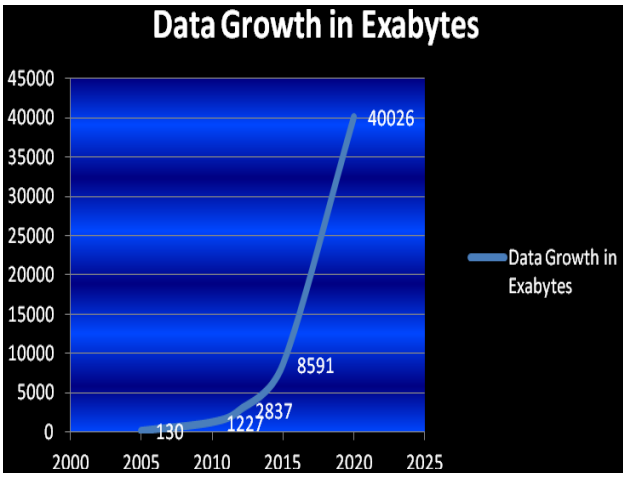

Fig.2. Big Data - 4V Challenges

Majority of the information are coming from digital TV, Social Media, videos and images from mobile phone etc. As per data analysts, data growth trend may even go beyond Moore's prediction [17] [9] [20].

B. Variety

We have all the different types of data today i.e. unstructured, structured and semi-structured format such as, emails, photos, machine data, audio-visual, PDF files, and HTML text. For example, there are millions of people using smart phone and sending variety information to IT network, therefore combining and managing the high volume of data with different format increases the complexity to the organizations on Data modeling & Data analysis standpoint [13].

C. Velocity

It is imperative to analyze, the data speed and frequency that are coming in and out, also the time required for the application system to process the data in relevant. In other words, processing the incoming data quickly and providing answers for decision making, for example in a Fraud management system, if any fraudulent Transaction (or) deal is suspected, it should be immediately flagged before the next transaction takes place. Thus, Organizations must analyze and handle the data streaming in a timely manner it as appropriately. This provides, challenges to the organizations in providing timely decisions, data relevance and speed.

D. Veracity

It is one of the key challenges for any organization to judge the data accuracy or quality of data. In other words, it is an aspect of questioning, whether the available data is accurate or not. Analyzing the data to make timely decision provides challenges to judge the accuracy of the information. Thus significant attention should be given on analyzing how to handle missing values, misstatements or untruths, imprecision, also uncertainty in data.

It is compulsory to do more accurate Big data analysis to make confident decisions, which would result in greater operational efficiencies, cost reductions and reduced risks. Apart from the 4Vs, there are several additional challenges and issues, which should be taken into consideration when an organization steps into perform the Big data analysis. We have listed below, the various other challenges, involved in Big data space [23] [19].

E. Data Discovery and Complexity

Discovering the high quality data from the large volume of data becomes another challenge, with respect to the data quality relevance for a specific activity. It is difficult to relate the different large data that organizations have; therefore, organizations need to emerge an in-depth data analysis with a new way of thinking. [31].

F. Scalability, Performance and Time line

The enormous growth rate on the volume and variety of data leads to complexity on the scalability & performance standpoint. As a result of this, organizations frequently need to focus profoundly on the scalability of their Big Data environment and architecture. Even though organizations can increase the hardware, having multiple data centers if required cloud, etc., it will lead to its own limitations. For example, scalability can be increased by having more hardware. It may provide performance results for the business or end user, but it does not mean that, adding hardware will help the organizations to focus on delivering up-to-the-minute information. Additionally, there are various challenges involved in performing high performance data analysis due to high data volume and timeline [22].

G. Heterogeneity and Data incompleteness

Analyzing incomplete data is one another challenge for the organizations. For example, a Knowledge Base can be created by collecting data from various heterogeneous sources such as social-networks (e.g., Twitter & Face-book), e-commerce (e.g., eBay & Amazon), Internet movie databases (e.g., IMDB), and in many database applications. This data is assumed to be standardized information; however, it is relatively possible that, the heterogeneous data may have incomplete information. Developing relationship with heterogeneous information becomes constrain in structuring, guiding the exploration of the semantics data. Even though there are several analysis and techniques available on mining the homogeneous information such as sequencing, community deduction and link the prediction and etc., most of these approaches can't be applied straight as heterogeneous data may have missing semantic [12] [31] [33].

H. Data Integration and Quality

In the integration standpoint, integrating the enterprise applications, databases, increase the complexity in optimizing each ETL operations, which are running on a very large data for example, optimizing the joins in large databases also when integrating the global information systems, with several information that are collected from multiple heterogeneous sources. Data warehouse requires extensive support for data cleaning. The load and continuous refresh on the high volume of data will lead to data quality and processing issue due to the probability that some of the data are incomplete from the source itself [24] [26].

III. EXTRACT BI BY IMPLEMENTING BIG DATA TECHNIQUES

In below, we have highlighted the Big Data techniques to extract BI.

A. High Performance Big Data Platform

Big data demands a high-performance platform to handle high volume of data, velocity and, variety. Organizations need the right platform, architecture, tools to capture and organize the Big data and reveal the new insights and higher business value. It is therefore a specialized hardware and infrastructure/platform must be required to administrate, integrate and support the Big Data environment [4] [6] [15] [16].

We suggest the below High Performance Big data Platform/Architecture using the emerging Big data techniques i.e. Hadoop and No-SQL. It is proven that, Hadoop has arisen as the principal system for handling Big data and relational databases i.e. to process the large data sets and analyze the structured & unstructured data [21].

Figure-3 in below describes Big Data High Performance Platform. The Big data high performance platform provides a high-performance distributed runtime with Hadoop MapReduce, HDFS in multiple applications and file systems with resource availability and predictability. This will deliver operational maturity and high resource utilization with the multi-application architecture [21] [4] [15] [6] [10] [30]. The below diagram describes the High Performance Big data platform & solutions. In integration perspectives, integrating enterprise applications, databases, increase the complexity in optimizing.

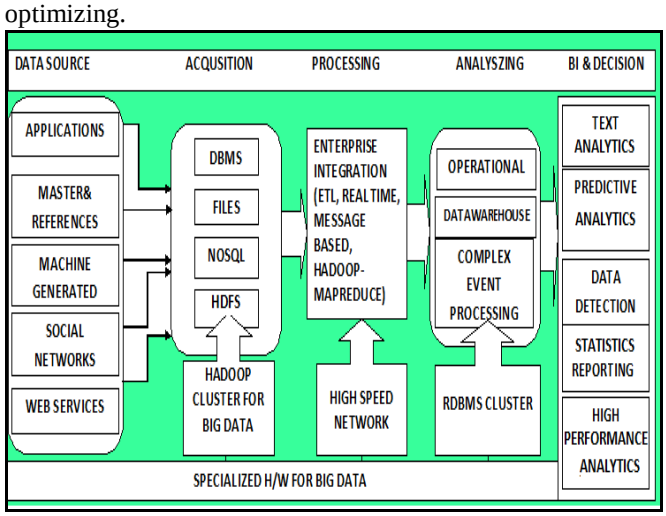

Fig.3. Big Data - High Performance Platform

We propose the above stated High Performance Platform or architecture to organizations for capturing, processing, and analyzing the data, for acquiring the data and discovering new insights to extract value and make repeatable decisions. Also based on the analysis of multiple research reports, we observe that "Predictive Analytics" is one of the recommended ways to leverage all the information, gain substantial new insights, and stay ahead of the competition. Since Predictive analytics is the sensible outcome of Big Data and business intelligence (BI) on what we need to while organizations collect overwhelming volumes of new data. It is fundamentally a process of learning from past behavior about how to do certain business processes healthier and bring new insights into how organizations should function. According to according Allison Snow, Senior Analyst of B2B Marketing at Forrester Report, *"Predictive analytics is about recognizing patterns in data to project probability".* [34].

B. Hadoop

We illustrate the methodical significance of employing Hadoop, HDFS, MapReduce and No-SQL in Big data platform to process and transform Big data. However, it also requires an infrastructure, which would upkeep the distributed processing and real-time demands on Big data space [21] [8].

Hadoop enables to process the large data sets in multiple commodity servers in the distributed data computing architecture. As the data storage & computations are managed and processed in multiple servers, it facilitates to achieve, faster result set, when running a query on the large data set. The query is being executed in all the local machines aging the large dataset, where in results, consolidated, and returned with faster result sets. Hadoop reduces the demand for large power servers as the data is stored and computed in multiple servers [8] [19] [18]. Naturally, Hadoop is a foundation step for a data analytics with its two major components such as HDFS and MapReduce [8].

- **Hadoop Distributed File System (HDFS):** A highly scalable & portable file system for storing data. In HDFS, the data is being stored in blocks on the different notes of the Hadoop clusters. This helps for the reliable and faster retrieval of data, because HDFS develops many replications of data blocks and distributes those data in multiple clusters. Since data is spread across multiple nodes, even if one fails, the data can be retrieved from other nodes. A typical HDFS block size is 128MB, which is relatively larger when comparing conventional file systems. HDFS makes many replications of the data blocks and allocates accordingly in cluster for the faster retrieval. This is made with objective to reduce the % seek time, compared to the % of transfer time. This will reduce the number of seeks in comparison to the amount of data to be transferred [21] [7] [8] [25].

- **Map-Reduce:** The Map-Reduce framework allows data analysis in the distributed computing with the highly scalable approach. It enables the large data sets to process in parallel, with the distributed algorithm on the Hadoop cluster. In technical stand point, the libraries in MapReduce model, handle the various activities by itself, so that the IT development team do not need to pay attention to take care of the various untidy details, for example library in this model takes care of parallelization, fault tolerance, data distribution, load balancing etc.

-

The Job Tracker ensures the large data sets run successfully and returned to the client. The Task Tracker performs the map and reduce-tasks, assigned by the Job-Tracker [8].

C. No SQL Database

The term NOSQL refers to "Not Only SQL Databases". NoSQL systems are distributed, non-relational databases, designed for large-scale data storage and for parallel data processing through a large number of commodity-servers. NOSQL plays a significant focus on supporting predictive analytics, i.e. data transformation & OLTP activities. NoSQL significantly provides higher contribution behind the analytical capabilities such as contextual search applications [10] [16] [30].

IV. BIG DATA ANALYTICS TECHNOLOGIES

In Big data phenomenon, Big Data Analytics signifies "applying the elevated analytical techniques on the high volume, and variety of data which are structured and unstructured nature, as well as batch and streaming". We suggest the organizations that, the highly developed emerging technologies such as Text Analysis, Predictive Analysis, Data detection and Mining, Statics Reporting, for the efficient data analysis and better & faster decisions [6] [21] [33].

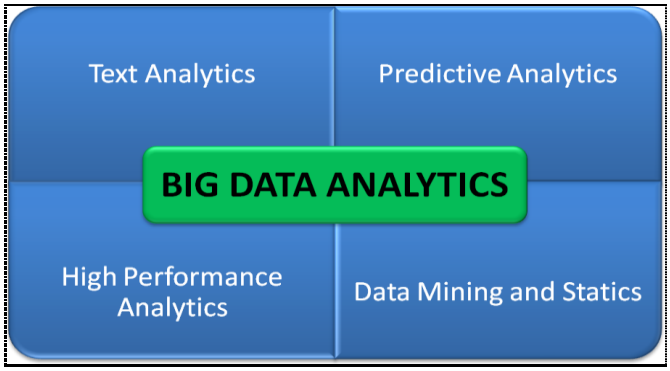

Fig 4: Big Data Analytics - Technologies

A. Text Analytics

Text Analytics: Text Analytics is nothing but a software or application, which has the Text Mining procedures to answer (or) solve a business issue. This approach enables the organization to get meaningful information from the unstructured & Machine generated data. Text analytics aids the organization to gain new insights into content-specific values such as sentiment, intensity and relevance. Text analytics applications can also aid the organization in converting words & phrases of the un-structured data in to numerical values, thus, the numeric values can be associated with structured data in a database for better analysis and comparison.

B. Predictive Analytics

Predictive analytics include variety of statistical & analytical methods that are used to develop models and predict the future. Predictive analytics with Data mining techniques can be used to analyze & compare the historical data with the current data to make prediction on trends & relationship patterns, and predict future events & behavior [18] [23].

C. Data Mining and Statistics Reporting

Mining insights from the large volume of data sets and statistical Reporting will aid the organization to make database decisions. This also enables for finding patterns, trends, behavior and to perform deep data analysis [2] [17] [29].

D. High Performance Analytics

The High performance analysis can be achieved by incorporating the below primary technologies.

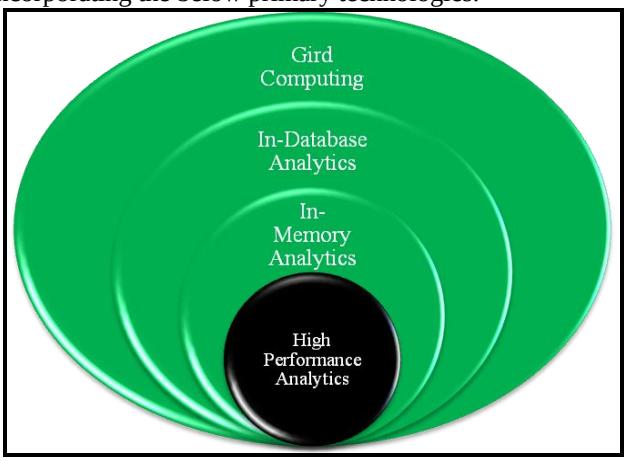

Fig.5. Big Data - High Performance Analytics

- **Grid Computing:** Grid computing combines computing resources to execute computationally difficult jobs. A centralized grid infrastructure enables the organization to balance the workload dynamically and achieve parallel processing for data management. It also enables for high availability and data analytics & reporting. [11] [15].

- **In-Database Analytics:** "In-Database analytics", the data processing and analytics are carried-out in the database itself. This helps to achieve scalability, parallel processing and partitioning. Handling the data movement and analytics tasks, in Database itself will improve the speed (or) rapidity in mining the insights. In-database processing significantly reduces the time required to prepare data & build, deploy & update analytical models. [22] [24].

- **In-Memory Analytics:** "In-memory analytics" enables to query the data from the main memory of the computer storage (RAM) instead of retrieving (or) querying the data from the physical storage. It is relatively fair that, querying data from main memory is faster when comparing to the disk databases, as the data seek time in memory is comparatively less than the disk storage. In-Memory analysis supports the business, to resolve difficult problems since analytical computations as the query or insights are gathered much faster [17] [18] [22] [23].

V. CONCLUSION

In this paper, we have discussed about how the emerging Big data techniques and tools, can help organizations to meet their various Big data challenges. We have also put forward the Big data solutions for analysis to extract business value for better and faster decision-making. It is obvious that technology should respond to the changing nature of business requirements in the Big data

context so that it can derive value for the better analysis patterns, more accurately, timely and intelligence decisions. Organizations need to have a structured set of solutions for capturing, processing, and analyzing the data, from data acquiring to discovering new insights for making repeatable decisions and scaling the connected information systems.

REFERENCES

1. X. Wu, X. Zhu, G. Q. Wu, et al., "Data mining with big data," IEEE Trans. on Knowledge and Data Engineering, vol. 26, no. 1, pp. 97-107, January 2014.Rajaraman and J. D. Ullman, "Mining of massive datasets," Cambridge University Press, 2012.
2. Z. Zheng, J. Zhu, M. R. Lyu. "Service-generated Big Data and Big Data-as-a-Service: An Overview," in Proc. IEEE BigData, pp. 403-410, October 2013. A . Bellogín, I. Cantador, F. Díez, et al., "An empirical comparison of social, collaborative filtering, and hybrid recommenders," ACM Trans. on Intelligent Systems and Technology, vol. 4, no. 1, pp. 1-37, January 2013.
3. T. C. Havens, J. C. Bezdek, C. Leckie, L. O. Hall, and M. Palaniswami, "Fuzzy c-Means Algorithms for Very Large Data," IEEE Trans. on Fuzzy Systems, vol. 20, no. 6, pp. 1130-1146, December 2012.
4. D. Markonis, R. Schaer, I. Eggel, H. Muller, and A. Depeursinge, "Using MapReduce for large-scale medical image analysis," in Proceedings of the 2nd IEEE International Conference on Healthcare Informatics, Imaging and Systems Biology (HISB '12), p. 1, IEEE, San Diego, Calif, USA, September 2012.
5. A. Sandryhaila and J. M. F. Moura, "Big data analysis with signal processing on graphs: representation and processing of massive data sets with irregular structure," IEEE Signal Processing Magazine, vol. 31, no. 5, pp. 80–90, 2014.
6. Cannataro M, Congiusta A, Pugliese A, Talia D, Trunfio P. Distributed data mining on grids: services, tools, and applications. IEEE Trans Syst Man Cyber Part B Cyber. 2004;34 (6):2451–65
7. Harati A, Lopez S, Obeid I, Picone J, Jacobson M, Tobochnik S. The TUH EEG CORPUS: A big data resource for automated EEG interpretation. In: Proceeding of the IEEE signal processing in medicine and biology symposium; 2014. pp. 1–5.
8. Yuan LY, Wu L, You JH, Chi Y. Rubato db: A highly scalable staged grid database system for OLTP and big data applications. In: Proceedings of the ACM international conference on conference on information and knowledge management; 2014. pp. 1–10.
9. Zhang L, Stoffel A, Behrisch M, Mittelstadt S, Schreck T, Pompl R, Weber S, Last H, Keim D. Visual analytics for the big data era—a comparative review of state-of-the-art commercial systems. In: Proceedings of the IEEE conference on visual analytics science and technology; 2012. pp. 173–182
10. Tsai C-W, Lai C-F, Chiang M-C, Yang L. Data mining for internet of things: a survey. IEEE Commun Surv Tutor. 2014;16(1):77–97.
11. Kollios G, Gunopulos D, Koudas N, Berchtold S. Efficient biased sampling for approximate clustering and outlier detection in large data sets. IEEE Trans Knowl Data Eng. 2003;15 (5):1170–87
12. Zhang L, Stoffel A, Behrisch M, Mittelstadt S, Schreck T, Pompl R, Weber S, Last H, Keim D. Visual analytics for the big data era—a comparative review of state-of-the-art commercial systems. In: Proceedings of the IEEE conference on visual analytics science and technology; 2012. pp. 173–182.
13. Laney D. 3D data management: controlling data volume, velocity, and variety. META Group, Tech. Rep. [Online]. http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.
14. Safavian S, Landgrebe D. A survey of decision tree classifier methodology. IEEE Trans SystMan Cyber. 1991;21(3):660–74.
15. Cooper BF, Silberstein A, Tam E, Ramakrishnan R, Sears R. Benchmarking cloud serving systems with YCSB. In: Proceedings of the ACM symposium on cloud computing; 2010. pp. 143–154.
16. Ghazal A, Rabl T, Hu M, Raab F, Poess M, Crolotte A, Jacobsen HA. BigBench: towards an industry standard benchmark for big data analytics. In: Proceedings of the ACM SIGMOD international conference on management of data; 2013. pp. 1197–1208
17. Malewicz G, Austern MH, Bik AJ, Dehnert JC, Horn I, Leiser N, Czajkowski G. Pregel: A system for large-scale graph processing. In: Proceedings of the ACM SIGMOD international conference on management of data; 2010. pp. 135–146.
18. MH. Kuo, T. Sahama, A. W. Kushniruk, E. M. Borycki and D. K. Grunwell, Health big data analytics: current perspectives, challenges and potential solutions, International Journal of Big Data Intelligence,1 (2014), pp.114-126.
19. X. Liu, G. Huang, and H. Mei, "Discovering homogeneous web service community in the user-centric web environment," IEEE Trans. on Services Computing, vol. 2, no. 2, pp. 167-181, April-June 2009.
20. Computer Research Association 2012. Challenges and Opportunities with Big Data - A community white paper developed by leading researchers across the USA.
21. Dewitt and Stonebreaker Microsoft Research, Seattle Oct, 2009. Hadoop Architecture and its Usage at Facebook Presented at Microsoft Research Seattle.
22. Gartner Group Press Release, STAMFORD June 27, 2011. Gartner's Pattern-Based Strategy to Gain Value in Big Data.
23. Gartner Group Press Release, STAMFORD June 27, 2011. Gartner's Pattern-Based Strategy to Gain Value in Big Data.
24. Gali Halevi & Dr.Henk F.Moed, Sep 2012. The Evaluation of BIG Data as Research and Scientific.
25. John Gantz and Davic Reinsel Dec, 2012. The Digital Universe in 2020: BIG Data, BIGger Digital Shadows, and Biggest Growth in the Far East
26. Sherif SKR, Senior Research Scientist, National ICT, Australia, May 07, 2011. An Introduction to Info Streams - Analyzing Big data in motion.
27. Steve Lohr. New York Times, Feb 11, 2012. The Age of BIG data. http://www.nytimes.com/2012/02/12/sunday-review/BIG-datas-impact-in-the-world.html
28. Yizhou Sun and Jiawei Han, Dec 2012. Mining Heterogeneous Information Networks: A Structural Analysis Approach.
29. Teradata, 2012. Biggest Challenges in BIG data Analysis http://www.teradata.com/News-Releases/2012/Sold-Out-Crowds-at-BIG-Analytics-Roadshow-Explored-How-to-Cut-Through-BIG-Data-Hype/
30. Oracle Feb, 2013. Information Management and BIG Data A Reference Architecture.
31. Computer Cramming More Components onto Integrated Circuits - GORDON E. MOORE, LIFE FELLOW,IEEE.
32. Hortonworks May, 2012. 7 Keys drives for the BIG Data Market. http://hortonworks.com/blog/7-key-drivers-for-the-BIG-data-market/
33. IBM BIG data Analytics: http://www.ibm.com/analytics/us/en/what-is-smarter-analytics/BIG-data-analysis.html
34. McKinsey Global Institute May, 2011. The next frontier for innovation, competition, and productivity.
35. B2B Marketing at Forrester Research Report, Oct (2018). https://www.forrester.com/B2B-Marketing

.