I.J. Modern Education and Computer Science, **2017, 3, 28-35**
Published Online March 2017 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2017.03.04

The Obstacles in Big Data Process

Rasim M. Alguliyev
Institute of Information Technology of Azerbaijan National Academy of Sciences 9,
B. Vahabzade str., Baku, AZ1141, Azerbaijan
Email: r.alguliev@gmail.com

Rena T. Gasimova, Rahim N. Abbaslı
Institute of Information Technology of Azerbaijan National Academy of Sciences 9,
B. Vahabzade str., Baku, AZ1141, Azerbaijan, GoEasy LTD, Canada, Mississauga L5B2N5
Email: {renakasumova@gmail.com, rahim.abbasli@gmail.com}

***Abstract*—**The increasing amount of data and a need to analyze the given data in a timely manner for multiple purposes has created a serious barrier in the big data analysis process. This article describes the challenges that big data creates at each step of the big data analysis process. These problems include typical analytical problems as well as the most uncommon challenges that are futuristic for the big data only. The article breaks down problems for each step of the big data analysis process and discusses these problems separately at each stage. It also offers some simplistic ways to solve these problems.

***Index Terms*—**Big data, big data analytics, database, management, NoSQL, MapReduce, Hadoop, cloud, data scientists.

I. Introduction

The advanced technologies has allowed companies to collect data from multiple sources to create a big data stream that was initially designed to be used to extract a valuable information to manage the business. The big data flood allows companies to build essential conceptual models that help to adjust to the new market trends and understand customer behavior. These models are used to differentiate the products offered by the company to match consumer’s expectations.

The information is also used to explore the niche markets before competitors enter the market. The leverage and the power that the big data offers has attracted many companies and scientists. However big data has also created challenges that must be solved to eliminate the gaps in the big data analysis process [1, 2].

The problems start right away during data acquisition, when the data tsunami requires us to make decisions, currently in an ad hoc manner, about what data to keep and what to discard, and how to store what we keep reliably with the right metadata. Much data today is not natively in structured format; for example, tweets and blogs are weakly structured pieces of text, while images and video are structured for storage and display, but not for semantic content and search: transforming such content into a structured format for later analysis is a major challenge. The value of data increases when it can be linked with other data, thus data integration is a major creator of value. Since most data is directly generated in digital format today, we have the opportunity and the challenge both to influence the creation to facilitate later linkage and to automatically link previously created data. Data analysis, organization, retrieval, and modeling are other foundational challenges. Data analysis is a clear bottleneck in many applications, both due to lack of scalability of the underlying algorithms and due to the complexity of the data that needs to be analyzed. Finally, presentation of the results and its interpretation by non-technical experts is crucial to extracting actionable knowledge.

The many novel challenges and opportunities associated with Big Data necessitate rethinking many aspects of these data management platforms, while retaining other desirable aspects. That appropriate investment in Big Data will lead to a new wave of fundamental technological advances that will be embodied in the next generations of Big Data management and analysis platforms, products, and systems. A major investment in Big Data, properly directed, can result not only in major scientific advances, but also lay the foundation for the next generation of advances in science, medicine, and business [3, 4].

The big data has created sophisticated analytical bottlenecks that cannot be solved with common tools and practices that are used in the industry today. Many counterintuitive approaches are taken to reduce the clout in the process and gain the beneficial competitive advantage that the big data analysis offers. Hence it has become a prerequisite to build the scientific approaches and theoretical models to tackle these problems.

II. The Big Data Analytical Problems

Through better analysis of the large volumes of data that are becoming available, there is the potential for making faster advances in many scientific disciplines and improving the profitability and success of many enterprises. The challenges include not just the obvious issues of scale, but also heterogeneity, lack of structure, error-handling, privacy, timeliness, provenance, and visualization, at all stages of the analysis pipeline from data acquisition to result interpretation. These technical challenges are common and therefore not cost-effective to address in the context of one big data alone. Furthermore, these challenges will require transformative solutions, and will not be addressed naturally by the next generation of industrial products. For this we need to encourage basic research in the direction of solving these technical problems that would achieve the promised benefits of big data.

The technologies used for big data analysis include MPP (Massively Parallel Processing) analytical platform systems, Cloud Services, Hadoop and MapReduce and NoSQL data warehouse management systems. The Hadoop systems that are part of Apache Software Foundation is one of the most common technologies used to analyze immense amount of data in distributed file systems. Hadoop consists of two main components; Hadoop MapReduce and Hadoop Distributed File Systems (HDFS). The MapReduce component is used in parallel calculations whereas HDFS is assisting in managing the distribution of the files within system [5].

The programming model used in Hadoop is MapReduce which was proposed by Dean and Ghemawat at Google. MapReduce is the basic data processing scheme used in Hadoop which includes breaking the entire task into two parts, known as mappers and reducers. At a high-level, mappers read the data from HDFS, process it and generate some intermediate results to the reducers. Reducers are used to aggregate the intermediate results to generate the final output which is again written to HDFS. A typical Hadoop job involves running several mappers and reducers across different nodes in the cluster.

A certain set of wrappers are currently being developed for MapReduce. These wrappers can provide a better control over the MapReduce code and aid in the source code development. The following wrappers are being widely used in combination with MapReduce.

Apache Pig is a SQL-like environment developed at Yahoo is being used by many organizations like Yahoo, Twitter, AOL, LinkedIn etc. Hive is another MapReduce wrapper developed by Facebook. These two wrappers provide a better environment and make the code development simpler since the programmers do not have to deal with the complexities of MapReduce coding. In addition to these wrappers, some researchers have also developed scalable machine learning libraries such as Mahout using MapReduce paradigm [6, 7].

The most advanced technologies are used in order to find a better way to extract the information from the big data. The process of analyzing the big data and extracting the essential information can be divided into four simple steps regardless of the purpose of the analysis:

- data collection;
- integration;
- analysis;
- real world application.

A. Data Collection

Color figures will be appearing only in online publication. All figures will be black and white graphs in print publication. Collecting the data from multiple sources is the first step of general big data schema (Figure 1). Challenges arise when the data sources are complex and sophisticated. The main source of data for Big Data stream is rapidly shifting from manual data entries to the data collected from sensors, social networks, automatic data collector machines that are triggered when a particular event happens, geographic information systems, automatic page scanners that enable to extract particular data characteristics form emails and online pages [8, 9].

Heterogeneity of the data sources is the most important problem in the data collection step. Heterogeneous data problems arises due to – Variety, Representation and Semantics of the data sources. Most of the data created nowadays fundamentally differ from the data types that the original systems were designed for [10, 11].

Semantic problems emerge due to the difference in the definition of the collected data between two parties. For example, the system designers might program the bank system to include some fees into total earnings whereas the data analysts would assume these fees to be reported separately. The analyst who uses the total earnings in calculations might not be aware of these definition, hence create blunder in reports.

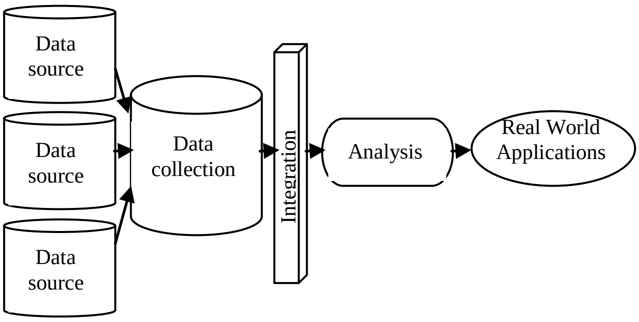


Fig.1. The Big Data Analytical Process Steps Problems

The data representation problems are similar in nature to the semantics problems. The data misrepresentation might be caused by different types of data that is used to show the same information. In the similar example we used above, even if both parties agree on the earnings definition, the data collector might capture the data in floating numbers where as it is required to be an integer for the other party in order to be able to merge the data into a bigger dataset. Common mistakes in representation is caused by date formats and character fields. Database designers might try to join datasets using name and surname of the customer to extract some essential information. Since character fields are case sensitive, even small misrepresentation, such as, using different capital letters, will make the search and joins inefficient.

Taking into account the commonality of the problem the most regularly used database tools such as SAS has created functions like “UPCASE” to capitalize all letters before making the comparison. SAS also uses

“DATEPART” function to generalize the date formats into single form before trying to match the observations. The easy way to solve the problem is to use ontologies agreed by both parties beforehand [12-13].

One of the most important problems in data collection is to collect the data required for the purpose. The immense amount of data acquisition requires to make instant decisions of what data to keep and what to discard. Due to the size of the big data this process usually takes enormous efforts and resources. Note that at this stage it is also important to delete or ignore the data deviations. Since most of the data collected nowadays is in the digital format, it is easy to link the data to each other or previously collected data. This creates false correlations that will be discussed in the next sections of the big data analysis process.

The outliers may also appear in dataset due to human or technical errors. Since the data collection process is separated from the data analysis, the systems and data collectors are not necessarily familiar with the purpose of the data that they are collecting. The detachment from the end result makes it hard for data collectors to pinpoint the outliers or data errors right away.

Another big problem is to transfer the data collected. Due to the size of the data the speed of the transfer may be a bottleneck in the process. Researchers work on creating the high speed fiber optics that can transfer the big data fast. Using new type of fiber optics, researchers in the Technical University of Denmark were able to transfer the data using single optics with the speed of 43 TBps. This is the highest speed achieved so far after successful attempt of scientists in İnstitute of the Technology Karlsruhe were able to achieve the speed of 32 Tbps [14].

Note that even though the fiber optics are developed to achieve high speeds in data transfer process, there are few successful attempts in creating the devices that can receive the data and store it with the same speed. At this stage of the big data process, the data protection and security problems must also be taken into consideration. Losing the high sensitive data during transition is one of the most common data security problems [15].

Many of the secure transmissions require some type of encryption agreed on beforehand by both parties. The multiple layered encryption codes are used by banks to transfer the sensitive customer information from one source to the other. Despite these security measures data losses happen in the system. Common practice is to create secure channels or SSL (secured socket layer) portals between parties to add additional layer of protection [16, 17].

B. Integration

The data that has been transferred must be stored in some form. Every day we create so much data that it costs companies fortune to store it in order for them to improve their business. The demand for storing the big data has increased so immensely and in such a fast pace that, new companies such as Switch has been created solely to help companies to resolve their problems with storing the data. According to the SuperNap, one of the biggest data centers in the world located in Las Vegas, US, Switch’s seven football court sized server helps Google, Morgan Stanley and the other big companies to store the data required for their business. The data storage market has grown to 70 billion dollar a year. According to Google’s fourth quarter fiscal year spending results In order to decrease their dependency on companies that focus on storing the data, Google spent 7.3 billion dollars in 2013 to invest in its infrastructure and data storage facilities [18, 19].

Storing such a huge sized data requires enormous amount of energy and resources. One of the problems of the Big Data is to find the best located servers to store the data. The server locations must also be energy efficient and scalable. The location is important due to the speed of transfer of the stored data to do the analyses. The problem of a storage location and the speed of transfer is more severe for companies that need to make instant decisions based on market fluctuations. Since most of the mathematical algorithms are designed to start the calculation of market variables at a given market prices, most brokers and stock market participants are vulnerable to the differences in information transfer speed. Most companies started to solve the problem using the cloud computing.

Cloud computing is a successful computational paradigm for managing and processing big data repositories, mainly because of its innovative metaphors known under the terms “Database as a Service” (DaaS) and “Infrastructure as a Service” (IaaS). DaaS defines a set of tools that provide final users with seamless mechanisms for creating, storing, accessing and managing their proper databases on remote (data) servers. Due to the na ïve features of big data, DaaS is the most appropriate computational data framework to implement big data repositories.

IaaS is a provision model according to which organizations outsource infrastructures used to support ICT operations. Due to specific application requirements of applications running over big data repositories, IaaS is the most appropriate computational service framework to implement big data applications.

Even though the cloud computing have multiple advantages it requires most small companies to “rent” the storage places from other companies. This creates business dependencies that only the big companies have resources to avoid.

Most data generated today is not generated with a metadata or is not transferred with a metadata attached. Omitted metadata creates problem during the integration process of large amount of data. The process requires to attach a meaning and assign the data to a field [20, 21].

Think about a case when customer information such as latest balance is transferred from one branch to another without indicating what the data is actually representing. This will obviously create confusion when attaching the data to the customer’s profile. When metadata does not exist or is not created during data collection process the stored data is basically useless for further analysis.

Integration of the data requires IT technicians to thoroughly understand the data transferred in order to store them in a meaningful fashion. It is not uncommon to spawn assumptions about the metadata when it does not exist.

However this approach might create false assumption and give wrong results in further analysis. The assumptions made on the metadata might also result in an increase in the scale of the data and create cross products during the join operations. The cross products will result in “useless” repetition of observations in dataset that should be avoided by any cost due to the size of the big data. Therefore there is no doubt that integrating multiple datasets in order to create a data storages is the hardest process at this step [22].The data storage for the Big Data must have some properties in order to serve the needs for long term and work efficiently (Figure 2).

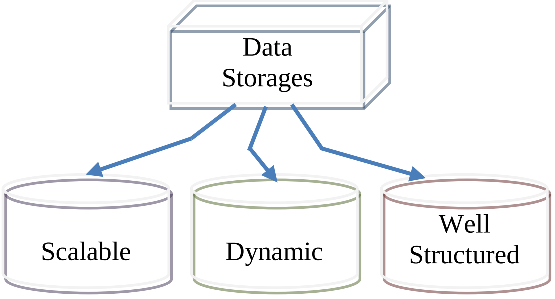


Fig.2. The Data Storages Characteristics

First of all the data storage must be well structured. Data structure is a specialized format for organizing and storing the data. The data structure must be easy to understand, easy to extract information from and easy to change when required. The structure of the data must be consistent [23]. Some of the data storages such as Oracle have physical and logical structure forms. Logical structure forms are not known to the operating systems. The data analyst or app developer might be aware of the logical structure which would help the analysis. It is harder to solve the problems when database administrator is not aware of the link between the physical structure and logical structure.

When dealing with Big Data the structure of the data storage might get too complicated. In that case many database administrators will try to create the data structure control files in order to find the required fields or understand the impact of the changes to the main structure. At this point it is also important to create the rollback codes. When the integration process fails and the data storage gets ugly after the changes has been made to the structure, data rollback files will help to undo the damage done to the database. Oracle has also created the redo files that would redo all the selected changes to the database [24].

Since the big data variables and data types can change actively the data storage structure must also be dynamic enough to accommodate the changes in the data received. The fast changing aspect of big data will also require the data storage to change its scale rapidly. Logical structure files are meant to help the data administrators to review the changes to the structure and adjust the data storages to eliminate any inconsistencies. One of the advantages of using clouds for storing the data is its dynamic and scalable properties. However most of the companies have raised their concerns for privacy issues of using clouds for storing the data. Most business units do not want to lose the control over the data. By using the stored data for marketing and advertising purposes, the companies offering cloud computing have lost their creditability. The low data protection and poor data backup policy issues have also caused the clients of cloud computing losses failure to protect customer sensitive data. These problems have created serious concerns over reliability and continuity of the services provided by the companies offering the cloud storage. Data protection laws have been passed in many countries to eliminate the above problems. However it is important to understand that the cloud storages and companies using them might be operating in different countries under different laws. Note that this type of storage of the big data requires high speed of connection for transfer of data extracted as discussed above [25-27].

Since storing the big data can be expensive, some companies have tried to store only part of the data collected. This approach has eliminated the need to outsource data storages and decrease the cost of storing the data. This simple technique might be helpful for the small companies that are focused only on one aspect of the customer service. However for most of the companies deciding on the data that needs to be stored can be more challenging than storing all of the data acquired. A complex analysis may require ad hoc information that can be hard to include in to the dataset if the data storage has not been designed to store that information [28].

C. Analysis

The problems that data analysts face when dealing with an average size of data set emerge in more severe form for the big data. Most business decisions need to be made in a timely manner. The companies that cannot modify their behavior to the changes in the market behavior in a timely manner have serious problems and will likely face severe problems in the future. The decisions made to adjust the company behavior must be based on the results of market realities. The challenges with the big data is that the data extracted for decisions may skew those realities by more than most companies can afford. The false correlations and unknown data links will create challenges in the interpretation of data extraction results. When the data scientist or the senior management do not have clear idea of how the data should look like, such biased results are hard to pinpoint [29].

False temporary correlation will most likely result in wrong decisions that can damage the company in the long term. The correlations are hard to remove unless it is extremely controversial. For example data analysis might show that the oil prices are highly correlated with the demand for a medicine to fight alcoholism. At this stage it is always good to keep in mind that correlation does not mean causation. Hence data analyst might want to ignore this trend due to the high unlikeliness or the counter intuition of results. However, even in such a controversial situation, market adjustments might create real data that might seem extreme, but could be real. The alcohol prices are indeed highly correlated with oil prices due to changes in the transportation cost, which affects the alcohol price and hence the consumptions which increases the demand for that medicine. It is indeed hard to find false correlations and links within big data. The links within the big data may be created due to the business practices or false data. Think about a new product introduced by the company. The product such as credit monitoring subscription can be bought with the new loan or transferred from the previous loan of the customer. During creation of metadata or database this link might be lost. Hence the data analyst that wants to understand the penetration rate or profitability on that product can ignore the transferred subscriptions. The result will most likely create a false view that can result in wrong decision such as cancellation of the product. Management might want to try to eliminate the problem using sanity checks before making big decisions. The management will require an alternative analysis that will support the decision. However this takes time that most companies cannot afford.

False results are not the only problem in big data analytics. The main problem is the velocity of the analysis. This is one of the three V's of the big data. Most define these V's as; Velocity, Variety and Volume. Variety and Volume has already been discussed during the data collection and integration process. Velocity of big data not only refers to the flow of data from sources to database but also the flow of the data from database to the end result in analytics. The speed of the data extraction and data analysis is the most important advantage that a company can get over competitors. Fast and correct strategic decisions will increase the return on investment and increase the market share of the company. However this requires to analyze the dynamic big data extremely fast [30, 31].

Think about the stock price fluctuations. The investor might want to short sell the stock if the implied volatility will decrease below certain level. The implied volatility might be calculated using Black Scholes model and then be adjusted using the average volatility of all bonds in that industry. The investor expects the calculations to be done instantly and buy the stocks at a given point in time. The investor will lose millions of dollars if such simple calculations will take more than coupe seconds. This process that might seem simple at first glance requires to analyse and extract immense amount of data. The problem can be partially solved using high end technologies. However such technological advanced products cost companies fortunes. Taking into account the abrupt changes in the technological trends and the pace of the technological innovations, the long term benefit from such programs is extremely low. It creates challenges for the small companies or start-ups to adjust their businesses swiftly and catch up to their competitors. The cost based optimization programs have been developed by big companies to increase the efficiency of the use of such technologies. For small companies such a simultaneous change to the new trend is nearly impossible. Most of these companies would prefer to adjust their capacity and enter into niche markets to survive. The new features in the technology have created new ways to spread the information from one end to another. It has created the market interdependencies which makes niche markets interconnected to bigger market events. Hence scraping all the information from the big data becomes reality and a need to survive [32].

It is also important to underline the models that are used to extract the information from the big data. Due to the size of the big data, common practice for most financial institutions would be to create the model using development sample and validate the results in the validation test sample. One of the problems in is to choose the development and the validation sample. Given the size of the big data and the computer capacity, the test sample might be less than 25% of the whole data.

The test sample is then divided into the development and validation sample which further decreases the model dataset. It must be noted that even though the test statistics could result in somewhat predictive models, the samples chosen are not enough to be fully sure in the decisions made. Of course even at this stage companies must make choice between how fast they can get the result versus how predictive they want their end result to be. Note that the big data collected might include thousands and thousands of variables. This feature of the big data makes it impossible to create the predictive models using all of these variables.

Companies bin the development datasets and use logistic regression on the development dataset to find a handful set of variables that can be used as predictive. It is extremely important to understand how these regression models work. Commonly used regression models- logistic models would choose one predictive variable and eliminate the variables that are correlated with that variable and have lower predictive power in comparison.

However we have already underlined the commonality of the false correlations in the big data. Hence this false correlation is not only analytical problem that arises at the final step of the big data analytics, it also causes problems during the model building process. The eliminated variables from the model due to the high correlation, could have been predictive ones that showed the false correlation and false linkage to the selected variable. Even after the modelling is complete, most senior management would chose to eliminate some of the predictive variables in order to make the model simplistic for operational purposes. It creates additional problems in the strength of the model built on the big data. Hence even though, the big data captures most of the variations in the customer behavior, and the models built on this data should be more predictive, the end result is not necessarily always true [33-35].

D. Real World Application Problems

The big data creates challenges at every step of the analytical process for data scientists and management. The companies are having bigger challenges in finding qualified data scientist that can work with the big data rather than the problems in the analytical process itself. Companies spend millions of dollars every year to train their staff to work with the big data. There are very few analysts in the job market who can work with the big data.

However, there are even fewer people in the market who can understand the data and the meaning underlying the numbers. Most analysts have hard time to understand and see the false data results. Data scientist must have ability to descend patterns where others do not see any. A study by McKinsey projects that “by 2018, the U.S. alone may face a 50 percent to 60 percent gap between supply and requisite demand of deep analytic talent” [36].

The shortage is already being felt across a broad spectrum of industries, including aerospace, insurance, pharmaceuticals, and finance. The negative trend has also been noted buy high ranking universities which now offer exclusive programs to train such analysts. The lack of data analyst will also create more problems for the companies that want to do ad hoc analysis on big data and implement the new models in the market. The problems discussed in this article will result in most companies losing their customers as a result of skew in the data and misinterpretation [37- 41].

III. Conclusion

The big data creates challenges at every step of extraction and analysis. Despite the challenges described in the article companies will not stop using big data for their business purposes. The companies and scientist will not stop trying to solve these problems using various analytical and scientific tools. The reality is that the future is depended on the big data.

The companies that want to survive and operate in the future will have to learn to work with the big data and solve these problems. Markets have already shifted to react to the big data trends. Even though the cost of big data analytics leaves less hope for small companies, most believe the small entrepreneurs and start-ups have higher chance to adjust their businesses due to the lack of a complicated hierarchical structure in those small companies. The companies also hope to attract the better data analysts buy offering them high salary. This might create new incentives for data analysts and scientists to implement the new ideas that might solve some of the most challenging problems in big data analytics.

Big data has already attracted a lot of attention and many work on solving fundamental problems that can change the way we perceive the reality right now. The technologies that understand and process huge amount of data to interact with humans are more and more of the reality and attract customers. Markets has already created the demand for such technologies. It is now up to the companies to find ways to satisfy that demand.

References

[1] L. Clifford, “Big data: How do your data grow?”, Nature, vol.455, 2008, pp.28–29.

[2] The digital universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. Study report, IDC, December 2012. http://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf

[3] V. Gopalkrishnan, D. Steier, H. Lewis, J., “Guszcza Big data, big business: bridging the gap” in Proceedings of the 1st International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications (BigMine '12), NY, USA, 2012. pp. 7–11.

[4] S. Madden, “From Databases to Big Data”, IEEE Internet Computing, vol.16, no.3, 2012, pp. 4–6.

[5] K-H. Lee, Y-J. Lee, H. Choi, Y.D. Chung, B. Moon, “Parallel data processing with MapReduce: a survey”, ACM SIGMOD Record, vol.40, no.4, 2011,pp.11–20.

[6] K.H. Lee, Y.J. Lee, H. Choi, Y.D. Chung, B. Moon “Parallel data processing with MapReduce: a survey” ACM SIGMOD Record, 2012, vol. 40, no. 4, pp. 11–20.

[7] Y. Chen, S. Alspaugh, R.H. Katz, “Interactive analytical processing in big data systems: A cross-industry study of mapreduce workloads” Proceedings of the VLDB Endowment (PVLDB), 2012, vol. 5, no. 12, pp. 1802–1813.

[8] W. Shang, Z.M. Jiang, H. Hemmati, B. Adams, A.E. Hassan, P. Martin, “Assisting developers of big data analytics applications when deploying on hadoop clouds”, in Proceedings of the 2013 International Conference on Software Engineering (ICSE '13), NJ, USA, 2013, pp.402–411.

[9] C.L.P. Chen, C.-Y. Zhang, “Data-intensive applications, challenges, techniques and technologies: A survey on Big Data, Information Sciences”, vol. 275, no. 10, 2014, pp. 314–347.

[10] C. Statchuk, M. Iles, F. Thomas, “Big data and analytics”, in Proceedings of the 2013 Conference of the Center for Advanced Studies on Collaborative Research (CASCON '13), USA, 2013, pp. 341–343.

[11] V.Mayer-Schonberger, K. Cukier, “Big Data: A Revolution That Will Transform How We Live”, Work and Think, Pub.: John Murray, 2013, p. 256.

[12] R. Birke, M. Björkqvist, L. Y. Chen, E. Smirni, T. Engbersen, “(Big)data in a virtualized world: volume, velocity, and variety in cloud datacenters” in Proceedings of the 12th USENIX conference on File and Storage Technologies (FAST'14), USENIX Association Berkeley, CA, USA, 2014, pp.177–189.

[13] Big Data - What Is It? 2013, http://www.sas.com/big-data/what-is-big-data.html

[14] SAS 9.2 Language Reference: Dictionary 4th Edition, Publisher SAS Institute Inc, Cary, NC, USA, 2011, p. 2356. https://support.sas.com/documentation/cdl/en/lrdict/64316/PDF/default/lrdict.pdf

[15] K. Munir, M. Odeh, R. McClatchey, S. Khan, I. Habib, “Semantic Information Retrieval from Distributed Heterogeneous Data Sources”, Presented at the 4th International Workshop on Frontiers of Information Technology (FIT 2006), Islamabad, Pakistan, 2006, pp. 1–6. http://arxiv.org/ftp/arxiv/papers/0707/0707.0745.pdf.

[16] O. Leif Katsuo, H. Hao, “World data transfer record back in Danish hands”, Technical University of Denmark (DTU), 2014, online resource,

http://www.dtu.dk/english/News/2014/07/Verdensrekord-i-dataoverfoersel-paa-danske-haender-igen?id=bed76c33-c9da-4214-91f3-c9ed3f8a0e24

[17] A. Cuzzocrea, “Privacy and Security of Big Data: Current Challenges and Future Research Perspectives”, in Proceedings of the First International Workshop on Privacy and Security of Big Data (PSBD '14), NY, USA, 2014, pp. 45–47.

[18] R.T. Gasimova, “Security of global domain infrastructure in the Internet”, Journal Problems of İnformation Technology, "İnformasiya Texnologiyaları" Publishing house, 2015, no. 2, p. 61–67. http://jpit.az/storage/files/article/71c96379ecf1714a60247e0206a0ba4b.pdf

[19] DigiCert is a U.S.-based Certificate Authority. It provides SSL Certificates and SSL management tools, online resource, https://www.digicert.com/ssl.htm

[20] S.V. Stacey, “Big Data creates big industry for storing data”, online resource, http://www.marketplace.org/topics/business/big-data-creates-big-industry-storing-data

[21] Google Inc. Announces Fourth Quarter and Fiscal Year 2013 Results http://investor.google.com/pdf/2013Q4_google_earnings_release.pdf

[22] T. Mastelic, A. Oleksiak, H. Claussen , I. Brandic, J-M. Pierson, V.A. Vasilakos, “Cloud Computing: Survey on Energy Efficiency”, Journal ACM Computing Surveys (CSUR), NY, USA, vol. 47, no.2, 2015, pp. 1–36.

[23] K. Smith, L. Seligman, A. Rosenthal, C. Kurcz, M. Greer, C. Macheret, M. Sexton, A. Eckstein, “"Big Metadata": The Need for Principled Metadata Management in Big Data Ecosystems”, in Proceedings of Workshop on Data analytics in the Cloud (DanaC'14), NY, USA, 2014, pp. 1–4.

[24] R.M. Alguliev, R.T. Gasimova, “Identification of Categorical Registration Data of Domain Names in Data Warehouse Construction Task” Intelligent Control and Automation, vol.4, no.2, 2013, pp. 227–234.

[25] M. L. Haas, “The Power Behind the Throne: Information Integration in the Age of Data-Driven Discovery”, in Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data (SIGMOD '15), NY, USA, 2015, p. 661.

[26] Oracle Database Online Documentation 11 g Release 2 (11.2), E10897-10, 2012, Primary Author: Bert Rich. http://docs.oracle.com/cd/E11882_01/server.112/e10897.pdf

[27] D. Lin, A. Squicciarini, “Data protection models for service provisioning in the cloud”, in Proceedings of the 15th ACM symposium on Access control models and technologies (SACMAT '10), NY, USA, 2010, pp.183–192.

[28] M. L. Kaufman, “Data Security in the World of Cloud Computing”, Journal IEEE Security and Privacy, vol.7, no. 4, 2009, pp. 61–64.

[29] C. Marinescu Dan. Cloud Computing: Theory and Practice. Publisher: Morgan Kaufmann, 1 edition, San Francisco, CA, USA, 2013, p. 416.

[30] D. Assunção Marcos, N. Rodrigo, Bianchi Silvia, A.S. Netto Marco, Buyya Rajkumar, “Big Data computing and clouds: Trends and Future Directions” Journal of Parallel and Distributed Computing, vol.79, 2015, p. 3–15.

[31] B. Marr, “Big Data: Using Smart Big Data, Analytics and Metrics to Make Better Decisions and Improve Performance”, Pub. John Wiley & Sons, Ltd.; 1 edition, 2015, p. 258.

[32] D-H. Tran, M.M. Gaber, K-U. Sattler, “Change detection in streaming data in the era of big data: models and issues”, ACM SIGKDD Explorations Newsletter - Special issue on big data, vol. 16, no. 1, 2014, NY, USA, pp. 30–38.

[33] L. Doug, “3D Data Management: Controlling Data Volume, Velocity and Variety”, Technical report, META Group, Inc (now Gartner, Inc.), February 2001, pp.1–3. http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

[34] K. Slagter, C-H. Hsu, Y-C. Chung Zhang Daqiang, “An improved partitioning mechanism for optimizing massive data analysis using MapReduce” Journal of Supercomputing, vol.66, no.1, 2013, pp.539–555.

[35] A. Ashraf, B. Shivnath, “Workload management for big data analytics”, in Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data (SIGMOD '13), NY, USA, 2013, pp. 929–932.

[36] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, A. Hung Byers, “Big data: The next frontier for innovation, competition, and productivity”, Analyst report, McKinsey Global Institute, May 2011. online resource, http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

[37] H. Chen, R.H.L. Chiang, V.C. Storey, “Business intelligence and analytics: from big data to big impact”, Journal Management Information Systems Quarterly, vol.36, no.4, 2012, pp.1165–1188.

[38] R. Ramasamy, “Towards big data analytics framework: ICT professionals salary profile compilation perspective”, in Proceedings of the 8th International Conference on Theory and Practice of Electronic Governance (ICEGOV '14), NY, USA, 2014, pp. 450–451.

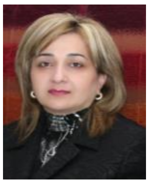
[39] A. Labrinidis, H. V. Jagadish, “Challenges and Opportunities with Big Data”, Proceedings of the VLDB Endowment, vol. 5, no.12, 2012, pp. 2032–2033.

[40] A. Baaziz, L. Quoniam, “How to use Big Data technologies to optimize operations in Upstream Petroleum Industry”, International Journal of Innovation, 2013, vol. 1, no. 1, pp. 19–29.

[41] K. Karthik, G. Kollias, V. Kumar, A. Grama, “Trends in Big Data analytics” Journal of Parallel and Distributed Computing, 2014, vol. 74, no. 7, pp. 2561–2573.

Authors' Profiles

Rasim M. Alguliyev. He is director of the Institute of Information Technology of Azerbaijan National Academy of Sciences (ANAS) and academician-secretary of ANAS. He is professor and full member of ANAS. His research interests include: Information Security, E-government; Information Society, Social Network Mining and Analysis, Cloud Computing, Evolutionary and Swarm Optimization, Data Mining, Text Mining, Web Mining, Social Network Analysis, Big Data Analytics, Scientometrics and Bibliometrics.



Rena T. Gasimova. She is head of sector at the Institute of Information Technology of ANAS. Her research interests include: Data Mining, Big Data Analytics, Domain name system, Decision Support Systems, Data Warehouse.

Rahim N. Abbasli. He is Senior Risk Analyst at one of the leading Canadian organizations specializing in subprime credits. He is developing credit models, analyzing big data, developing algorithms and adjusting underwriting rules according to the results.

How to cite this paper: Rasim M. Alguliyev, Rena T. Gasimova, Rahim N. Abbaslı,"The Obstacles in Big Data Process", International Journal of Modern Education and Computer Science(IJMECS), Vol.9, No.3, pp.28-35, 2017.DOI: 10.5815/ijmecs.2017.03.04