# Ins and Outs of Big Data: A Review

Hamidur Rahman(✉), Shahina Begum, and Mobyen Uddin Ahmed

School of Innovation, Design and Engineering,
Mälardalens University, Västerås, Sweden
{hamidur.rahman, shahina.begum,
mobyenuddin.ahmed}@mdh.se

**Abstract.** Today with the fast development of digital technologies and advance communications a gigantic amount of data sets with massive and complex structures called 'Big data' is being produced everyday enormously and exponentially. Again, the arrival of social media, advent of smart homes, offices and hospitals are connected as Internet of Things (IoT), this influence also a lot to Big data. According to the study, Big data presents data sets with large magnitude including structured, semi-structured or unstructured data. The study also presents the new technologies for data analyzing, collecting, fast searching, proper sharing, exact storing, speedy transferring, hidden pattern visualization and violations of privacy etc. This paper presents an overview of ins and outs of Big Data where the content, scope, samples, methods, advantages, challenges and privacy of Big data have been discussed. The goal of this article is to provide big data knowledge to the research community for the sake of its many real life applications such as traffic management, driver monitoring, health care in hospitals, meteorology and so on.

**Keywords:** Big data issue · Framework · Analytics · Challenges · Tools

## 1 Introduction

The 'Big data' term has come into the research community more clearly during 2013 and afterwards. Several authors have tried to explain the definition and the possible issues, technologies, challenges and privacy of big data in a concise way [1–5]. For example in 2001, Laney et al. have highlighted the challenges and opportunities generated by increased data through a 3Vs model, i.e., increases in volume, velocity and variety [6]. In recent years, the world has become so much digitalized and interconnected and as a result the amount of data has been exploding. Therefore, to manage the massive amount of records it requires extremely powerful business intelligence. The problem may arise even more during data acquisition if the amount of data is too large and then it may have a confusion level that what data to keep and what to discard and how to store the data in a reliable way. A clear definition of Big data has been using for the accumulation of different sort of huge amount of data since last 2–3 years. In 2015, the digital world expanded to 5.6 exabytes ($10^{18}$ bytes) of data created each day. This figure is expected to double by every 24 months or so [7]. As a result, storing, managing, sharing, analyzing and visualizing information via typical database software tools is not only so difficult but also very hazardous task. Big data can be structured,

M.U. Ahmed et al. (Eds.): HealthyIoT 2016, LNICST 187, pp. 44–51, 2016.
DOI: 10.1007/978-3-319-51234-1_7

semi-structured and unstructured in nature but it could help in businesses by producing automated services to target their potential partners, agents or customers.

There are some Big data review articles available in online but most of them have emphasized on specific area e.g., big data framework, big challenges, big data applications etc. but almost all of them have failed to provide complete overview of Big data [2, 8, 9]. In this paper, we have presented a complete overview of Big Data and its present state-of-the-art. Additionally, we have tried to find out big data important characteristics, Big data frameworks and analytic, challenges of big data and possible solutions, big data tools and its applications in famous companies. This article will be very helpful for new researchers specially data scientist, research institutes and companies to get insights view and latest technologies of big data for their research planning, business activities and future demand for handling massive amount of data.

## 2 Materials and Methods

The Big data is relatively a new topic and the amount of research articles published so far is limited in this area. Around 60 Big data related articles have been collected from different online sources where IEEE explore, Research Gate and Google Scholar databases are the privileged sources. Some of the articles were searched using google Chrome search engine and during the searching period different key words were used such as ‘big data’, ‘big data issues’, ‘big data challenges’, ‘big data analytics’, ‘recent trend of big data’ etc. and it was also considered the most recent articles which are available in online database. In our case we only considered the articles published in between 2013 to 2016. About 70% of the collected articles were considered for detailed study through the paper and remaining 30% of the articles are excluded due to similarity with considered articles and less important for the study.

## 3 Big Data Characteristics

Big data is usually characterized by the three dimensions or 3 V called Volume, Velocity and Variety [6]. However, other dimensions presented in Fig. 1 such as variety, validity and value can be at least equally important. According to the study, additional three dimensions Veracity, Validity and Value have considered and presented in Fig. 1 with “6 Vs” of big data. The 1st V is Volume which concerns the fact of amount of generated data that is increasing tremendously in each day. The second V is Velocity which has come into light due to more and more data and is provided to the users immediately whenever required for real time processing. Variety is the 3rd V considered due to the tremendous growth in data sources which are needed for analysis.

Veracity which includes trust in the information received, is often cited as an important 4th V dimension in addition to big Data. Validity does not only involve ensuring accurate measurements but also the transparency of assumptions and connections behind the process. Value refers to recent large volumes of data measured in exabytes, petabytes or higher ranked of data and highly valuable for research institutes and industries.

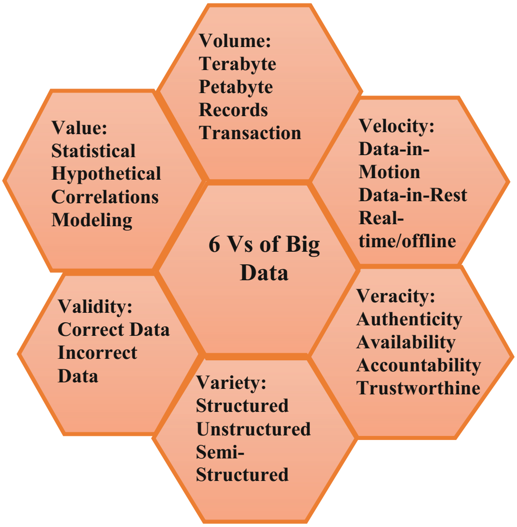

**Fig. 1.** The important characteristics of Big data (6 V's)

## 4 Big Data Analytic and Frameworks

In general view, data analytics is one of the major part in Big data environment which is responsible to simplify complexity of the data and calculation for achieving of expected pattern of data sets and outcome. As a whole, there are 3 main tasks in Big data framework which includes initial planning, implementation and evaluation and all the tasks have 8 layers as described in Fig. 2 [1].

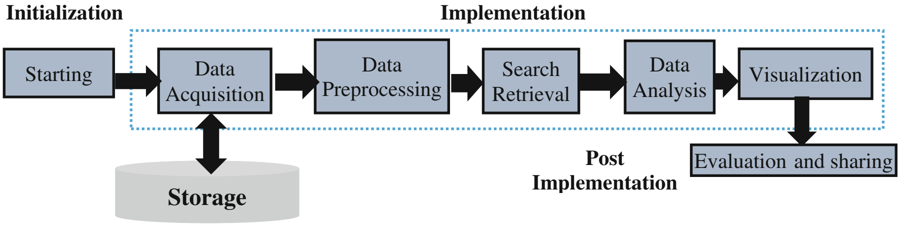

**Fig. 2.** Layers of big data analytic task and framework

**Initialization:** The first layer of any Big data framework is the primary planning which requires new investments for big changes and the changes basically include installation of a new technological infrastructure and a new way to process and control data [10]. At the beginning it is extremely important to find problems that needs a solution and decision whether they could be solved using new technologies or just with available software and techniques. These problems could be large volume challenges, real time processing, predictive analytics, on-demand analytics and so on.

**Implementation:** The second task is implementation and there are several activities such as data storage, pre-processing, search retrieval, analysis and visualization. To overcome the storage capacity problem cloud computing technology has a great advantage [11, 12] and provides easy access for applications from different corners of the world. Data analysis is one of the most important steps in implementation where various preprocessing operations are necessary to address different imperfections in collected raw data. As the data sources are different multi-source data fusion technology such as [13–15] could be applied. After that all the data must be pre-processed to avoid similarity, remove noise and delete unwanted signals [16–20]. For example, data can have multiple formats as heterogeneous sources are involved. It can also mix with noises where unnecessary data, errors, outliers etc. are included. Additionally, it may subsequently necessary to fit requirements of analysis algorithms. Therefore, data preprocessing includes a wide range of operations such as cleaning, integration, reduction, normalization, transformation, discretization etc.

When the pre-processing stage is done then the search retrieval is performed to extract values for further analysis for companies and institutions. Advance analytics is one of the most efficient approaches which provides algorithms such as descriptive analytics, inquisitive analytics, predictive and prescriptive analytics to perform complex analytics on either structured or unstructured data. When the analyzing part is done then the visualization is very important where it guides the analysis process and presents the summary of the results in a transparent, understandable and meaningful way. For the simple graphical representation of data, most software packages support classical charts and dashboards.

**Evaluation and Sharing:** The third task of the big data framework is evaluation the outcome and sharing it among the agents [10]. To evaluate a Big data project, it is necessary to consider a range of diverse data inputs, quality of data and expected results. To develop procedures for Big Data evaluation, the project first needs to allow real time stream processing and incremental computation of statistics. There is also necessity to have parallel processing and exploitation of distributed computing so that data can be processed in a reasonable amount of time. It is also considering that the project can easily be integrated with visualization tools. Finally, it should perform summary indexing to accelerate queries on big datasets to accelerate running queries.

## 5 Big Data Challenges and Inconsistencies

Context awareness is one of the major analytic challenge that focuses on some portions of data and which is useful for resource consumption [3]. Another crucial challenge is visual analysis that how data seems to be for the perception of human vision. Similarly, data efficiency, correlation between the features of data and contents validation are notable challenges. Data privacy, security and trust are also major concern among organizations. When volume of data grows, it is difficult to gain insight into data within time period. Processing near real time data will always require processing interval in order to produce satisfactory output. Transition between structured data- stored in well-defined tables and unstructured data (images, videos, text) required for analysis will affect end to end processing of data. Invention of new non-relational technologies will provide some flexibility in data representation and processing.

In circumstances where big data are collected, aggregated, transformed or represented inconsistencies invariably find their way into large datasets [21]. This can be attributed to a number of factors in human behaviors and in decision-making process. When datasets contain a temporal attribute, data items with conflicting circumstances may coincide or overlap in time. The time interval relationships between conflicting data items can result in partial temporal inconsistency. Spatial inconsistencies can be arisen from the geometric representation of objects, spatial relations between objects or aggregation of composite objects. As big datasets are increasingly generated from social media, blogs, emails, crowd-sourced ratings, inconsistencies in unstructured text and messages become an important research topic. If two texts are referring to the same event or entity, then they are said to be of co-reference. Event or entity co-referencing is a necessary condition for text inconsistencies.

## 6 Big Data Domain, Technology, Tools and Solution

Big Data domain has no boundary including retail application to governmental works. A big data might be petabyte (1024 terabyte) or Exabyte (1024 petabyte) of data consisting of billions to trillions of records of millions of people from different sources like educational institutions, research institutions, medical hospitals, small or multinational company data, customer care, weather records, demographical data, social media records, astronomical data etc. [21]. These massive data sets and its applications include technologies such Mathematics, Artificial Intelligence Especially Machine Learning, Data Mining, Cloud Computing, Real Time Data Streaming technology and so on [21]. Time series analysis is also very useful and of course there are many visualization technologies that can be used in Big data.

Today most of the renowned companies are using big data tools for their special needs. For example, Hadoop[1] and MongoDB[2] are the two best data storage and management tool used by Google, Amazon and MIT, MTV respectively. For data cleaning, Stratebi and Platon companies are using DataCleaner[3] tool. Teradata[4] is another big data tool for data mining used by Air Canada or cisco. Autodesk company uses Qubole[5] for their data analysis. For big data visualization, Plot.ly[6] is one of the greatest and many renowned companies like Google, Goji, VTT are using this tool. For the data integration, Pentaho[7] is one of the best tool used by CAT, Logitech etc.

[1] http://wiki.apache.org/hadoop/PoweredBy#G.

[2] https://www.mongodb.com/industries.

[3] https://datacleaner.org/testimonials.

[4] http://www.teradata.se/customers-list/browse/?LangType=1053&LangSelect=true.

[5] https://www.qubole.com/customer/?nabe=5695374637924352:1.

[6] https://plot.ly/#trusted-by.

[7] http://www.pentaho.com/customers.

**Table 1.** Big data tools used by renowned companies

| No | Big data tools | Where it is used |
|---|---|---|
| 1 | Hadoop | Google, Amazon, Alibaba, Facebook etc. |
| 2 | MongoDB | citiGroup, MIT, GOV.UK, ebay, MTV etc. |
| 3 | DataCleaner | Stratebi, Platon, BestBrains etc. |
| 4 | Teradata | Air Canada, cisco, Coca-Cola, Coop, Dell, Daimler etc. |
| 5 | Qubole | Autodesk, Answers.com, Capilary, Quora, Nextdoor, etc. |
| 6 | Plot.ly | Google, Goji, VTT, U.S. Air Force etc. |
| 7 | Pentaho | CAT, Nasdaq, Logitech, U.S. Navy etc. |
| 8 | Python | Forecastwatch.com, AstraZeneca, Carmanah etc. |
| 9 | Import.io | Quid, Nygg, OpenRise, University of Houston etc. |

Python[8] is widely used as a Big data language for company like AstraZeneca, Carmanah etc. As a big data collection tool Import.io[9] is pioneer and used by Quid, Nygg, OpenRise etc. A list of Big data tools used by famous companies are listed in Table 1.

There are thousands of Big data tools both available in the market to buy and also for free trial for extraction, storage, cleaning, mining, visualizing, analyzing and integrating. Table 2 shows the most popular big data tools.

**Table 2.** A number of popular big data tools (https://www.import.io/post/all-the-best-big-data-tools-and-how-to-use-them/).

| No | Big data area | Tools |
|---|---|---|
| 1 | Data Storage and Management | Hadoop, Cloudera, MongoDB, Talend |
| 2 | Data Cleaning | OpenRefine, DataCleaner |
| 3 | Data Mining | RapidMiner, Teradata, FramedData, Kaggle |
| 4 | Data Analysis | Qubole, BigML, Statwing |
| 5 | Data Visualization | Tableau, Silk, CartoDB, Chartio, Plot.ly, |
| 6 | Data Integration | Blockspring, Pentaho |
| 7 | Data Languages | R, Python, RegEx, XPath |
| 8 | Data Collection | Import.io |

## 7 Conclusion

A general overview and concept of the Big data has been discussed in this article including Big data 6 V, it's framework and analytic issues. Additionally, the difference between big and small data, popular tools, inconsistencies and challenges also have been reviewed. Due to management and analysis of petabytes and exabytes of data, the big data management system cooperates and ensures a high level of data quality, accessibility and helps to locate valuable information in large set of unstructured and

[8] https://www.python.org/about/success/#engineering.

[9] https://www.import.io/.

unplanned data. This review of different techniques can be applied to various fields of engineering, industry and medical science. Some real life applications such as autonomous driving, smooth transaction for semi-autonomous driving or driver monitoring in context of big data analysis will be presented as future work.

**Acknowledgement.** The authors would like to acknowledge the Swedish Knowledge Foundation (KKS), Swedish Governmental agency for innovation Systems (VINNOVA), Volvo Car Corporation, The Swedish National Road and Transportation Research Institute, Autoliv AB, Hök instrument AB, and Prevas AB Sweden for their support of the research projects in this area.

## References

1. Tekiner, F., Keane, J.A.: Big data framework. In: 2013 IEEE International Conference on Systems, Man, and Cybernetics, pp. 1494–1499 (2013)
2. Sagiroglu, S., Sinanc, D.: Big data: a review. In: 2013 International Conference on Collaboration Technologies and Systems (CTS), pp. 42–47 (2013)
3. Katal, A., Wazid, M., Goudar, R.H.: Big data: issues, challenges, tools and good practices. In: 2013 Sixth International Conference on Contemporary Computing (IC3), pp. 404–409 (2013)
4. Xiong, W., Yu, Z., Bei, Z., Zhao, J., Zhang, F., Zou, Y., et al.: A characterization of big data benchmarks. In: 2013 IEEE International Conference on Big Data, pp. 118–125 (2013)
5. Lu, T., Guo, X., Xu, B., Zhao, L., Peng, Y., Yang, H.: Next big thing in big data: the security of the ICT supply chain. In: 2013 International Conference on Social Computing (SocialCom), pp. 1066–1073 (2013)
6. Laney, D.: 3-D data management: controlling data volume, velocity and variety. META Group Original Research Note (2001)
7. Ahmed, F.D., Jaber, A.N., Majid, M.B.A., Ahmad, M.S.: Agent-based big data analytics in retailing: a case study. In: 2015 4th International Conference on Software Engineering and Computer Systems (ICSECS), pp. 67–72 (2015)
8. Gupta, P., Tyagi, N.: An approach towards big data-a review. In: 2015 International Conference on Computing, Communication & Automation (ICCCA), pp. 118–123 (2015)
9. Rout, T., Garanayak, M., Senapati, M.R., Kamilla, S.K.: Big data and its applications: a review. In: 2015 International Conference on Electrical, Electronics, Signals, Communication and Optimization (EESCO), pp. 1–5 (2015)
10. Mousannif, H., Sabah, H., Douiji, Y., Sayad, Y.O.: From big data to big projects: a step-by-step roadmap. In: 2014 International Conference on Future Internet of Things and Cloud (FiCloud), pp. 373–378 (2014)
11. Huang, G., He, J., Chi, C.H., Zhou, W., Zhang, Y.: A data as a product model for future consumption of big stream data in clouds. In: 2015 IEEE International Conference on Services Computing (SCC), pp. 256–263 (2015)
12. Khan, I., Naqvi, S.K., Alam, M., Rizvi, S.N.A.: Data model for big data in cloud environment. In: 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom), pp. 582–585 (2015)
13. Suciu, G., Vulpe, A., Craciunescu, R., Butca, C., Suciu, V.: Big data fusion for eHealth and ambient assisted living cloud applications. In: 2015 IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom), pp. 102–106 (2015)

14. Yang, L.T., Kuang, L., Chen, J., Hao, F., Luo, C.: A holistic approach to distributed dimensionality reduction of big data. IEEE Trans. Cloud Comput. **PP**, 1 (2015)
15. Zheng, Y.: Methodologies for cross-domain data fusion: an overview. IEEE Trans. Big Data **1**, 16–34 (2015)
16. Pandey, S., Tokekar, V.: Prominence of MapReduce in big data processing. In: 2014 Fourth International Conference on Communication Systems and Network Technologies (CSNT), pp. 555–560 (2014)
17. Wang, J., Song, Z., Li, Q., Yu, J., Chen, F.: Semantic-based intelligent data clean framework for big data. In: 2014 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC), pp. 448–453 (2014)
18. Biookaghazadeh, S., Xu, Y., Zhou, S., Zhao, M.: Enabling scientific data storage and processing on big-data systems. In: 2015 IEEE International Conference on Big Data (Big Data), pp. 1978–1984 (2015)
19. Diao, Y., Liu, K.Y., Meng, X., Ye, X., He, K.: A big data online cleaning algorithm based on dynamic outlier detection. In: 2015 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), pp. 230–234 (2015)
20. Taleb, I., Dssouli, R., Serhani, M.A.: Big data pre-processing: a quality framework. In: 2015 IEEE International Congress on Big Data, pp. 191–198 (2015)
21. Zhang, D.: Inconsistencies in big data. In: 2013 12th IEEE International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC), pp. 61–67 (2013)