Asian Journal of Research in Computer Science

6(3): 1-11, 2020; Article no.AJRCOS.61470
ISSN: 2581-8260

A Review of Data Science and Big Data Computing

Wajid Ali[1*], Muhammad Usman Shafique[2], Muhammad Arslan Majeed[1], Muhammad Faizan[3] and Ahmad Raza[4]

[1]Department of Computer Science, Government College University, Lahore, Pakistan.
[2]Department of Computer Science, University of Engineering and Technology, Lahore, Pakistan.
[3]Department of Computer Science, University of Punjab, Lahore, Pakistan.
[4]Department of Computer Science, The Superior College Lahore, Pakistan.

Authors' contributions

This work was carried out in collaboration among all authors. Author WA designed the study, performed the statistical analysis, wrote the protocol and, wrote the first draft of the manuscript. Authors MUS, MAM, MF and AR managed the analyses of the study. Author WA managed the literature searches. All authors read and approved the final manuscript.

Article Information

DOI: 10.9734/AJRCOS/2020/v6i330158
Editor(s):
(1) Dr. Omidiora, Elijah Olusayo, Ladoke Akintola University of Technology (LAUTECH), Nigeria.
(2) Dr. G. Sudheer, GVP College of Engineering for Women, India.
(3) Dr. Hasibun Naher, BRAC University, Bangladesh.
Reviewers:
(1) M. Rao Batchanaboyina, Acharya Nagarjuna University, India.
(2) Rodrigo Hipólito Roza, Pontifical Catholic University of Campinas, Brazil.
(3) Totohasina André, ENSET- University of Antsiranana, Madagascar.
Complete Peer review History: http://www.sdiarticle4.com/review-history/61470

Review Article

Received 10 August 2020
Accepted 13 October 2020
Published 02 November 2020

ABSTRACT

Data Science emerged as an important discipline and its education is essential for success in almost every aspect of life. Here comes the age of Big data. Big data impacts all aspects of our lives and society is admitting it. Data processing and other techniques are combined to convert abundant data into valuable information for society, organizations, and people. Specific strategies and approaches are needed to provide better to educate future data scientists to overcome the challenges of Big data. In this paper, we discussed the general concept of data science, Big data, and areas of Big data computing.

Keywords: Data science; big data; 5vs of big data; big data computing; big data technologies.

**Corresponding author: E-mail: gcu.wajid.ali@gmail.com;*

1. INTRODUCTION

Recent years have seen the rise of data science as a significant discipline.

Consider it as an amalgamation of disciples of statistics, artificial intelligence, and computer science with its sub-disciplines including machine learning. To turn abundantly available data into value for individuals, organizations, and society existing approaches need to be combined. Also new challenges arose, such as “Big data” [1].

Data science is a connective tissue data analytics (including big data) technology [3].

Big data organizes and collects valuable knowledge from potentially increasing, large quantities, various formats, and constantly evolving data sets obtained by various sets and autonomous sources using various computational and machine learning techniques in the shortest possible time [4].

Big data is all around in our lives, if we look around starting from weather, to location, using the phone, and the social media platform, from medicine to network monitoring. If we talk about research and development, financial and businesses system, and banking systems Big data is everywhere. The way education has become so easy that almost everyone can reach any data regarding education.

Big data is benefitting our education system by providing all the information about relevant study programs and the right jobs relevant to individuals’ skills and is helping the world at the governmental level.

Big data has its challenges like size, credibility, quality, and diverse demands which needed to be focused.

This work has been structured as follows:

Section 2 is covering the data science, disciplines of data science, types of data, pillars of data science, and the fundamental concepts of data science. Section 3 covered Big data, 4 is on dimensions of Big data, 5 is about big data technologies, while section 6 is all about areas of Big data computing including education, and many more and in section 7 the challenges of Big data that cannot be overlooked.

2. WHAT IS DATA SCIENCE?

Data science at a high level is a collection of basic concepts that endorse and direct the ethical extraction of data information and expertise. Probably the most directly connected field of data science is data mining, the practical exploitation of data information by technology implementing such concepts [3].

Data science is about information extraction, preparation, analysis, visualization, and maintenance. It is a cross-disciplinary field that draws insights from data using scientific methods and computer algorithms. A field that combines statistics, computer science, and mathematics to evaluate data and using machine learning algorithms to forecast possible case events [6].

Data science is a field of study which devotes vast data volume to find unseen trends, Obtaining meaningful information, and decision making in business using modern tools and technology. In data science, we construct a predictive model using complicated machine learning algorithms [23]

In this day-to-day industry data science is one of the most discussed subjects. It has become popular over the years and businesses have started to introduce data science technology to grow their industries and customer satisfaction [23].

Fig. 1 shows the machine learning, data science, and statistical research relation

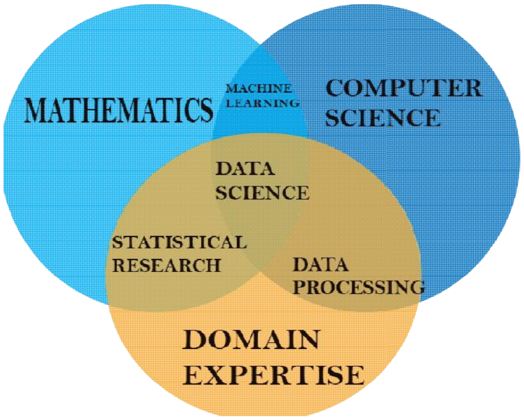

Fig. 1. Machine learning, data science, and statistical research relation [22]

Data science uses both structured and unstructured data, the science of nature conduct an in-depth statistical study on the data, utilizes innovative analysis and machine learning (ML) and forecasting, combines previous and present evidence to estimate future success and effects. [23]

2.1 Discipline of Data Science

The disciplines are overlapping. Take an example of the distinction between machine learning, data mining, and statistics: their origins are quite different but disciplines are definitely overlapping [1].

Statistics is the root of data science, categorized into descriptive statistics and inferential statistics. The Former summarize data using mean, standard deviation, and frequency while the latter one estimate characteristics of data [1].

Data mining identifies unsuspected relationships and summarizes data in novel ways by analyzing data sets so it is understandable, where input is given as a table and the output in the form of a graph, tree structure, and clusters. Data science depends on statistics, databases, and algorithms [1].

Machine learning gives the computer the ability to learns without having a program specifically or an artificial intelligence branch developing an application that learn from data and enhances their performance over time that focus on. This field has emerged from techniques such as neural networks from within artificial intelligence [1, 11].

2.1.1 How machine learning works

- Select and prepare a training data set [11]
- Choose an algorithm to run on the training data set [11]
- Training algorithms to create the models [11].

Visualization: data is most often processed to be visually analyzed. By the deliberate use of graphs, maps, and diagrams, people are able to understand data better [10].

Things that we don’t know we don’t recognize and analyze heavily rely on the human intuition and interaction with the data, using the right visualization will leverage the vision capacities of the human cognitive system [1].

Database and process mining: Database is used for data collection, and is one of the cornerstones of the data science. Programmers need not think about data storage due to this technology Process [1].

Process perspective is being added to data mining and machine learning. Process mining explores the collision between event data and process models, where the former is observed behavior, and the latter is made by hand or automatically discovered. Event data is compared to specific process models. E.g. Petri nets.

Fig. 2 shows the disciplines of data science

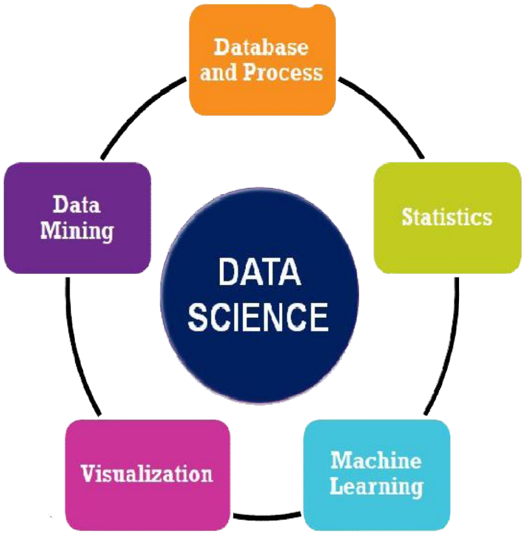

Fig. 2. Disciplines of data science [20]

2.2 Types of Data

Data science is a field to convert data into real value. Data may be structure, unstructured, and semi-structured.

Structured data is commonly labeled as relational. E.g. names, addresses, dates, locations. Structured data is highly ordered by machine language and readily interpreted. Many who work in relational databases can easily insert, scan, and manipulate structured data [19].

Unstructured data is qualitative data which cannot be interpreted and analyzed using traditional instruments and techniques. E.g. email, visual, audio, telephone activity, social media activity, satellite imagery, and more. Unstructured data is difficult to understand because it does not have a predefined format, so it cannot be arranged into a relational database [19].

Semi-structured is a mixture of both structured and unstructured data. Actually, it is unstructured data while structured data is often added to it. When you take an image from a smartphone shutter records details about light reflection in form of 1's and 0's (binary), this data has no form (unstructured) but the camera keeps more details like when the image was shot, last change, and image size, which is organized (structured) [10].

2.3 Three Pillars of Data Science

The three pillars of data science include data, technologies, and people and elaboration for three of them is as follows:

Data: refers to domain fields such as relational data and non-relational data (e.g. unstructured and semi-structured data) [5].

Technologies: includes Hadoop ecosystem, NoSQL, in-memory computing, data processing, cloud computing and machine learning [5].

People: include informatics technicians, statisticians, subject specialists, computer scientists, and market analysts [5].

Fig. 3 shows the pillars of data science including data, technologies, and people below

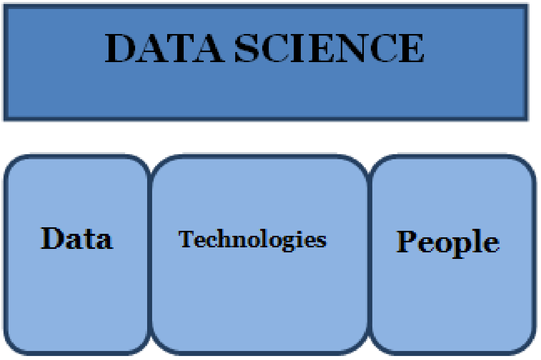

Fig. 3. Pillars of data science [5]

There is also a clear human dimension of human science (e.g. criteria analysis, user interface design, concept evaluation, and collaboration with domain experts), which is closely linked to behavioral discipline [5].

Among all three pillars the 3rd one "people" is the most important one, we can buy more computers, storage, and tools to handle Big data efficiently but human capacity is not scale-up, educating people called data scientists is essential to solving the Big data era problems [5].

2.4 Fundamental Concepts of Data Science

I. By following a method of relative well-defined steps, the extraction of valuable information from data to solve market challenges can be dealt with systematically one codification of this process is a cross industry-standard framework for data mining(CRISP-DM), will guide us about data analytic issues [3].

II. Evaluating the findings of data science requires a close analysis of the context in which they are applied [3].

III. It is useful to break the business problem into components to estimate the probability and analyze the value, together with the structure of recombining components [3].

IV. The notion of discovering similarities is one of the first data analysis ideas found in the business-analytics scenario. "Correlation" is also used loosely to mean data items and provide knowledge about other data items specifically, known quantities and minimize our confusion about unknown quantities [3].

V. Computing similarity is one of the data sciences principal methods. Entities which are similar in terms of known features or attribute are often similar in comparison to unknown features or attributes [3].

VI. Mining data techniques can be very powerful, and one of the most important concepts to grasp when applying data mining tools to a real problem is the need to detect and avoid overfitting. **Overfitting:** If you dig into a data set so closely you'll discover something but it won't generalize to the data you are examining is called overfitting a dataset [3].

VII. For a casual conclusion**,** the Existence of mitigating variables, perhaps unknown ones, must be paid very careful attention. Not always enough to find associations in data, one must use their model to direct decisions about how to affect the action that generates the data [3].

3. BIG DATA

Big data organize and extract valued information from rapidly growing, large volumes, various forms and frequently changing data sets collected from multiple and autonomous sources, using multiple statistical and machine learning techniques [4].

Big data define as different numbers of Vs, from 3Vs (high volume, high velocity, and high variety) to 4Vs including veracity, and then to 5Vs in which value was included [5].

4. BIG DATA DIMENSIONS

Volume: means data size which scales to a terabyte, petabyte, and even more. In past, it was a challenge storing it but now data lakes and Hadoop Kind of sites have eased the pressure [5, 4].

Velocity: Speed, at which data is analyzed, stored, processed, and generated [5].

Data streams to the business at an unprecedented rate with the growth in the internet of things and must be handled in a timely manner, sensors and smart meters drive the needle to handle torrents of data in almost real-time [4].

Variety: data comes in many forms organized, numerical records, to unstructured text, images, audios, and financial records [4].

Veracity: veracity stands for data quality, reliability, and uncertainty. We see veracity as a challenging research dimensions because this is an area that still needs to be researched more thoroughly, particularly on the impacts of the veracity on data integration and analytics [5].

Value: value signifies the discovery of workable intelligence and is the most challenging research dimensions from all the 5Vs [5]. We start to see a significant implication of Big data in every area of life and culture. Fig 4 illustrates the idea of Big data's 5Vs, where value is at the center of the diagram and intersects with the other 4Vs

Fig. 4 shows the 5Vs of big data.

5. BIG DATA TECHNOLOGIES

- Until the realization of the amplitude of the data passing through the internet, specialists began to think about how to handle such an amount of data. To mine the knowledge and to obtain good insight they wanted to build software that can construct the expected result. A common deployment that handles the big data is MapReduce [6].
  MapReduce consists of two things: "mapping" and "reducing". By mapping, any dataset is restructured to different values and reducing take mapped values to output and form smaller sets of tuples [6].

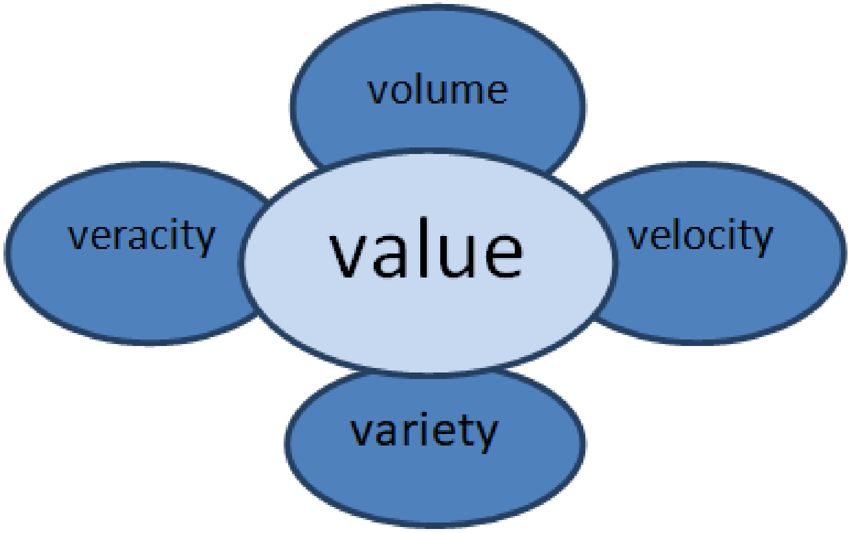

Fig. 4. 5Vs of Big data [5]

Hadoop is the most famous technology to mine and sort the Big data. Being open-source tech, Hadoop is the most common approach implemented for handling large data, has adequate flexibility for working with many, or even assembles data sources so you can do significant scale processing. HDFS (Hadoop Distributed File System) used by Hadoop has the feature that divides the data into smaller blocks and distributes it over the whole cluster [6].

Fig. 4 shows the mapping and reducing of data.

- **Microsoft HD insight:** Microsoft powered Big data system operated by apache Hadoop and is available in the cloud as service, uses Azure blob storage as a default file structure, also offers high accessibility at low cost. Also allows us to connect Big data stored in the Azure cloud using the power query option [7].
- **NoSQL:** NoSQL database holds unstructured data; a row can have its own sets of values for columns and provides excellent efficiency in storing a vast volume of data [7].
- **Hive:** Hive is alike SQL Bridge that connects to Hadoop to allow traditional applications to conduct Queries, and is used for data mining purposes [6].

Fig. 5 shows the mapping and reducing big data

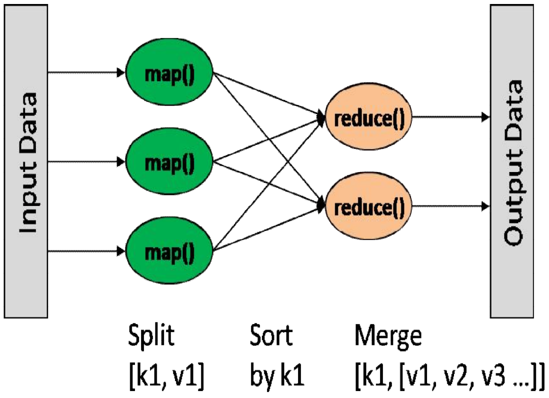

Fig. 5. mapping and reducing of data [21]

- **Scoop:** A data conversion method that links Hadoop to different relational databases. Effectively this can be used to move structured data to Hadoop and hive, import data from MySQL, Oracle to Hadoop HDFS, exports data from Hadoop file system to relational databases [7, 8].
- **PolyBase:** Operates on parallel Data Warehouse (PDW) for SQL server 2012 used for viewing data stored in PWD (a data warehousing system designed to process any amount of relational data and offers connectivity with Hadoop so that we can access non-relational data) [7].
- **Big data in Excel:** Microsoft excel helps the user to leverage “Big data” techniques for testing, reviewing, and data research, it’s data mining feature is user friendly and users can use these resources to identify hidden connections and patterns [9].
  You can also link the saved data to Hadoop using excel 2013. Hortonworks, which actually works to provide enterprise apache Hadoop, offers an alternative to use, excel 2013 to view Big data in their Hadoop database. Excel’s function of power view can be used to summarize data [7].
- **Presto:** A query engine (SQL-on-Hadoop) is created by Facebook and open-sourced to handle petabytes of data, unlike hive not based on MapReduce and can access data easily [7].

6. AREAS OF BIG DATA COMPUTING

As domains of engineering, science, health, and business are growing a new system for each domain needs to converge relationships to be redefined among all the stakeholders, not possible to rely on experiences, however, the use of critically important data sources for decision making is often necessary. Many fields of big data computing is as follows [3].

6.1 Scientific Exploration

Data obtained from different sensors are processed to obtain the relevant knowledge for the well-being of society, such as for physics and astronomy a huge number of scientists are working together to develop, run and analyze the properties of sensor networks and detectors for experimental purposes. Earth observation networks collecting information and theoretical methods on earth chemical, biological and physical using remote sensing systems. To boot social and economic wellbeing and application for the forecast of weather, tracking, and reaction to natural disasters, forecast of climate change [3].

6.2 Healthcare

Healthcare organizations would like to forecast the areas from which the disease spread to avoid further spread. However, it would not be possible to forecast the origin of the disease precisely until observational data from different sources become available. When a new flu virus H1N1 emerged in 2009 Google not only expected but released a report in the science journal Nature, looking at what people was searching on the internet [3].

Healthcare using big data:

1) Product development: It takes incredible time and effort to find new medication and other health-related products. Big data can lead to reducing time in numerous ways. Naturally, this lowers the prices [12].
2) Patient outcome: Big data advances healthcare because it makes diagnosis and procedures for physicians and other medical providers more reliable and precise. With advanced technology, doctors will look forward to treatments that were incurable or dangerous cases [12].
3) Operational efficiency: Collecting employee data will increase the success of team members of medical institutions or including hospitals and pharmaceutical firms. Managers will update the workflow more effectively and allocate the resources where they are most needed [12].
4) Driving innovation: One of the greatest applications of healthcare with big data, there will be no improvement in medicine without invention. Big data would increase the pace by which the new therapies would be introduced and also the standard of treatment [12].

6.3 Governance

In order to forecast traffic flows and monitor public transit plans, transportation agencies using real-time. Traffic info, intelligence agencies evaluate photographs from social media, news, and artifacts. State organization classifies fraudsters and prosecution of support, by review of complicated identities and tax returns, sensor applications for promoting sanitation and fire suppression [3].

6.4 Financial and Business Analytics

Customer satisfaction and customer preferences are financial firms' most significant problems. In some sectors like the travel industry for the optimum cost storage industry as well as retail goods for prospective consumers, sentiment analyses and predictive analysis will play a key role.

Big data is widely adopted by the banking industry and the reason is fraud prevention. History and activity records of clients can be used to identify some suspicious activity [14].

6.4.1 Retailors

Analysis, diversity, and price efficiency of store action, product positioning design, value enhancement, the efficiency of labor input, delivery, and logistic optimization [13].
Loyalty cards and credit cards are given to customers are not courteous. The information gathered by the cards is stored on a broad database that enables retailors to make smarter pricing decisions, monitor stock, and incentives for customers. Big data is important because the number of customers' visits, customer preferences, and product options, the number of shops, and internet sales are very rapidly accrued. [14].

6.4.2 Marketing

Marketers love Big data, the more knowledge you get the more you know. What Big data advertisers have not had before is the opportunity to describe their offering in the smallest terms regarding consumers' behavior. Marketing business 360i claims Big data is helpful [14].

- Customer acquisition and upselling [14].
- Latest customer identity [14].
- Reveal new opportunities for marketing [14].
- Conduct rentable publicity [14].
- Measure more reliably the effect of campaigns [14].

6.5 Web Analytics

Millions of people visit different websites every day and produce a vast variety of content. Companies are constantly searching for this knowledge to consider their sites' shortcomings and to deliver more tailored advertising etc. This includes methods for complex data processing, well beyond the memory of a single device or even a system cluster. Helps determine traffic and patterns in trends important for market research [3].

6.6 Personal Locations

New business models, advanced navigation, regional, or emergency response ads. [13].

6.7 Social Media

Big data research platform helps businesses, advertisers, and organizations to assess consumers' satisfaction in near-real-time, as well as how the consumer feels about goods or services in their social network posts. The convergence of Big data with social web allows businesses to evaluate the most active consumers among other members of the social network [14].

6.8 Research and Development

Big data helps businesses, institutions, and government departments to benefit from the potential of Big data to inhale huge volumes of unstructured knowledge and to give scientists a deeper view of what is happening [14].

6.9 Network Monitoring

Just about all the data logs for computer and networking equipment, the quantity of data logged is unmanageable. Big data can easily accommodate the data size, helps administrators to track network behavior and detect issues, or check for such trends of malicious activity.

Big data also helps a machine to handle loads between electronic systems. A Virtual computer can handle a service request from a remote location rather than relying on a single computer process. It will help manipulate data to allow network management teams to concentrate and building techniques to prevent unwanted data streams. Efficient time scheduling allows automatic solutions that promote protection and minimize downtime [15].

6.10 Education

Cloud computing is another technology based on Big data. This technology can improve educational services, enabling both young, and adult students' access to cheap content, online teachers, and fellow learning communities [16].

Big data will promote a classical education system, which allows teachers to evaluate what students understand and which strategies are more appropriate for each student, which also allows teachers to learn new approaches for their curriculum [16].

Data mining and data processing techniques will also provide students and teachers with instant input on their academic success. These techniques will offer a detailed study and draw useful information about such schooling patterns. Collective and broad data will forecast who needs more support from the education system by lowering and minimizing the risk of failure [16].

Online education has advanced very quickly in recent years and is having a growing effect on the education market [16].

The spatial definition of the ecosystem facilitates the learning process in higher education is exceedingly complicated and nuanced [16].

6.10.1 Challenges of implementing big data in higher education

Regardless of the accelerated introduction of big data techniques implementation in higher education, there is still a need for caution [17].

- Financial spending may be one of the key hurdles to the introduction of bid data research in higher education. Many institutions see analytics not as an opportunity but as a costly endeavor. The biggest issue of affordability is the suspected criteria for pricy data collection processes [17].

  More investments are needed in analytical professionals who are able to make proper use of large data and analytics [17].
- Another challenge in the implementation of the use of Big data in higher education is data collection, data security, and learners' rights to their individual conduct records. Critical issues need to be taken into account: can students articulate themselves as being followed up on their activities? For teachers, students, and parents, Scholarships, how much knowledge is required? Do students' need to request assistance? [17].

The data at its source should be masked to resolve these problems, which will protect the confidentiality of students and teachers' information [17].

6.10.2 Benefits of Big data and open data in education

For young ones and for adults Big data may bring new learning experiences. Students should also exchange experience with schools in order to improve awareness and capacity. Universities and educational institutions will support growing potential students [16].

Open data is actually Big data but smaller in size and all the information is accessible for everyone. Open data is primarily from government datasets other agencies and organizations or from people [16].

Fig. 6 shows how Big data and open data.

- **Improved instructions:** It will enhance the success and willingness of the students to personalize their lessons. The lesson can be analytically modified by the teachers [16].
- **Matching students to programs:** Open data will help students and parents to identify the right career and or best school plan [16].
- **Matching students to employment:** Companies and applicants may use better and more powerful ways to use open data to assess their talents for necessary qualifications. Students can identify and apply more effectively than ever for jobs that complement their skills [16].
- **Transparency in education financing:** Enabling students to get access to education who previously did not have the opportunity to engage in educational activities. In addition, you can do something about higher education to choose the most appropriate educational program [16].
- **Efficiency in system administration:** School education programmers should build a knowledgeable school supply that can allow the administration to produce more [16].

7. IMPORTANT ISSUES REGARDING BIG DATA

There are many new possibilities posed by Big data from the data science viewpoints. Any of them are not really new but are problems that are not taken care of. Some interesting issues are:

Data size: on one hand developing “one-pass learning” algorithms with limited storage while attempt to make smaller components of significant data from Big data [18].
Data trust: data can be retrieved more easily and efficiently, but taking care of the source of data is important, also if the information is correct and useful too [18].

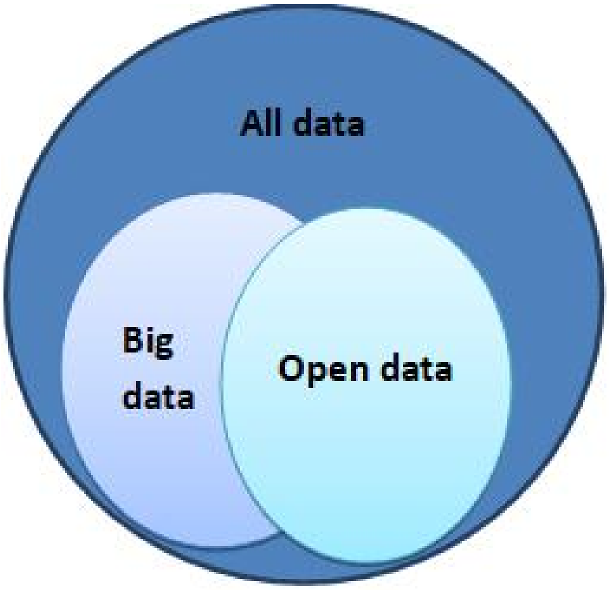

Fig. 6. Big data and open data relation [16]

Distributed existence: different kinds of access rights of different data parts are so whole data cannot be accessed and exploited. But different sources can have different quality thus resulting in the noise at the end [19].

Diverse demands: people may have diverse specifications and demands, while for each demand the high cost for building separate models is disabled. Is it possible to build one model for a variety of demands? We need to work on that [18].

Sub models: for different demands it may happen that we have to build a lot of smaller models instead of one large one [18].

Big optimization: global optimization algorithms were popular in academic research but not in the industry. One major issue is the high computational cost for engineering systems; the advent of massive data technology will eliminate this barrier [18].

Complex optimization: the formulation of the optimization problem itself becomes a dynamic optimization for the optimization of complex systems. Big data can give us a new perspective, new methods, and effective solutions for formulating optical mixing problems. Thus Big data will provide varied possibilities but no particular methodology will satisfy all demands. Big data also provide a “mix balance” between technologies and research [18].

8. CONCLUSION

This article is about data science, Big data, and the areas of Big data computing. How data science and Big data is helping us improve our lives. How the problems from every corner of the world can be solved with the best possible solution. This article also discusses the technologies used to collect, store, analyze, and visualize data. What volume and variety of data can be processed at what speed. While we talk about a decade ago Big data covered few areas but if we talk about the year 2020 and in the near future Big data is going to impact how we live our lives, improved healthcare facilities, better to do online shopping, way better banking system.

Big data is contributing towards our safety and education is much less costly than it was before, students and teachers are much at ease with the help of Big data being applied to education.

As opportunities come along with the challenges, so is it with Big data, as to know if the data is credible, and diverse demand of big data models costs much which need to be resolved and for security concerns when data from different sources are accessed the quality may vary resulting in the noise. The challenges must be focused so Big data will be easy to handle and implement in every department.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Van Der Aalst, will M.Pprocess mining, data science in action 2nd edition; 2016.
2. Fawcett, Tom. & Provost, Foster. Data Science and its Relationship to Big data and Data-Driven Decision Making; 2013. Accessed at https://www.liebertpub.com/doi/full/10.1089/big.2013.1508
3. Kune, Raghavendra, Konugurthi, Pramod kumar, Agarwal Arun, Chillargi, Raghavendra Rao, Buyya, Rajkumar. The anatomy of Big data computing. A Survey paper, Wiley online library; 2015.
4. Data flair team. (2019, March 14) what is data science?. Data Flair; 2020. Avaialable:https://dataflair.training/blogs/what-is-data-science/
5. Song Il-Yeol, Zhu Yongjun. Big data and data science: what should we teach? An Article, Wiley online library; 2015.
6. Tole, Alexandru, Ardian. Big data Challenges. dbJournal; 2013. Accessed at http://www.dbjournal.ro/archive/13/13_4.pdf
7. Ramachandran, Madhivanan. (2013, November 26). Top Big data technologies used to store and analyze data. Crayon; 2020. https://www.crayondata.com/blog/top-big-data-technologies-used-store-analyse-data/#:~:text=BIG%20DATA%20is%20a%20term%20used%20for%20a,of%20challenges%20relating%20to%20its%20volume%20and%20complexity
8. Sqoop tutorial, tutorials point, https://www.tutorialspoint.com/sqoop/index.htm
9. Check point learning. Thomson Reuters, Excel: Big data tools checkpointlearning.thomsonreuters.com/CourseFinder/CourseDetails/Excel-Big-Data-Tools/9360
10. Castrounis Alex. (2016, November). Types of Data and Data Sets. KDNuggets; 2020.

Available:www.kdnuggets.com/2016/11/big-data-data-science-explained.html

11. IBM Cloud Education, (2020, July 15). How machine learning works; 2020. Available:https://www.ibm.com/cloud/learn/machine-learning
12. Turea, Marina. (2019, November 19). How health care uses Big data; 2020. Available:https://healthcareweekly.com/big-data-in-healthcare/
13. SAGIROGLU, Seref, & SINANC, Duygu. Big data: A review. conference, IEEE; 2013. Accessed at https://ieeexplore.ieee.org/abstract/document/6567202
14. Kasnner, Michael. (2015, October 2). 5 key areas where big data is making a major impact. Techopedia; 2020. Available:https://www.techopedia.com/2/30244/technology-trends/big-data/big-data-is-changing-lives-lifestyles-and-the-way-we-work
15. (2016, may 18). How Big data is being applied network monitoring. Digital Edge; 2020. availabhttps://www.digitaledge.org/the-current-state-of-big-data-network-monitoring/
16. Drigas Athanasios S, Leliopoulos Panagiotis. The Use of Big data in Education. Article, Researchgate; 2014. Accessedhttps://www.researchgate.net/profile/Athanasios_Drigas/publication/274890131_The_Use_of_Big_Data_in_Education/links/552b90bb0cf2e089a3aa4526/The-Use-of-Big-Data-in-Education.
17. Klašnja-Milićević, Aleksandra, Ivanović Mirjana, Budimac Zoran. Data science in education: Big data and learning analytics. Review, Wiley. Online Library; 2017. Available:https://doi.org/10.1002/cae.21844
18. Zhou Zhi-hua, Chawala Nitesh V, Jin, Yaochu, William Graham J. Big data Opportunities and Challenges: Discussions from Data Analytics Perspectives. Journal and magazine, IEEE, 2014. Accessed at https://ieeexplore.ieee.org/abstract/document/6920114
19. Pickle Devin. (2018, November 16). Structure and unstructured data what is the difference?; 2020. Available:https://learn.g2.com/structured-vs-unstructured-data
20. Yoo Taesun. (2019, July 4). Why is it so hard to become data scientist; 2020. Available:https://towardsdatascience.com/how-to-become-a-data-scientist-3f8d6e75482f
21. Ho Ricky. (2008, November 25). Map/Reduce function; 2020. Available:http://horicky.blogspot.com/2008/11/hadoop-mapreduce-implementation.html
22. Cook, Kimberly. (2018, October 6). How to Make A Career In These Field - Data Science, Machine Learning and Big data?; 2020. Available:http://houseofbots.com/news-detail/3686-4-how-to-make-a-career-in-these-field-data-science-machine-learning-and-big-data
23. Simplilearn team. (2020, September 9) What is Data Science: A Comprehensive Guide for Beginners? Simplilearn; 2020. Available:https://www.simplilearn.com/tutorials/data-science-tutorial/what-is-data-science?source=sl_frs_nav_playlist_video_clicked

Peer-review history:
The peer review history for this paper can be accessed here:
http://www.sdiarticle4.com/review-history/61470