Essential Learning Components from Big Data Case Studies

Pranjal Shikhar Sinha, Vineet Singh, Pallavi Asthana and Pragya

Abstract Global information generates trillions bytes of data every single day via e-mails, chats, e-commerce and media feeds. This structured and unstructured data is often referred as Big Data. The ability of big data lies in analyzing and capturing the information and quickly converting it into actionable insights. Big Data identifies business use cases with measurable outcomes to develop organization-wise big data strategy, through right tools and architecture for implementation with existing data for quick success. This paper presents the multiple case studies in some of the biggest data-generating platforms like e-governance, retail sector, healthcare sector, social networking sites, and astronomy. It also discusses important software-based tools that are employed for analyzing such vast amount of data for predicting the future performance of these platforms based on feedback and popularity.

Keywords Data mining · Predictive analysis · Hadoop · Data visualizations

1 Introduction

Data is the structure on which any association flourishes. The size, assortment and the quick difference in such data require another sort of big data analytics, just as various storage and analysis techniques. Anything extending from customer names and addresses to things available, to buy made, to workers hired and so forth has turned out to be fundamental for daily agreements. Such sheer measures of big data should

P. S. Sinha (✉) · V. Singh · P. Asthana
Amity University, Lucknow Campus, Lucknow, India
e-mail: shikhar.pranjal3@gmail.com

V. Singh
e-mail: vsingh@lko.amity.edu

P. Asthana
e-mail: pasthana@lko.amity.edu

Pragya
MVPG College, Lucknow, India
e-mail: de.pragya2011@gmail.com

P. K. Singh et al. (eds.), *Proceedings of First International Conference on Computing, Communications, and Cyber-Security (IC4S 2019)*, Lecture Notes in Networks and Systems 121, https://doi.org/10.1007/978-981-15-3369-3_66

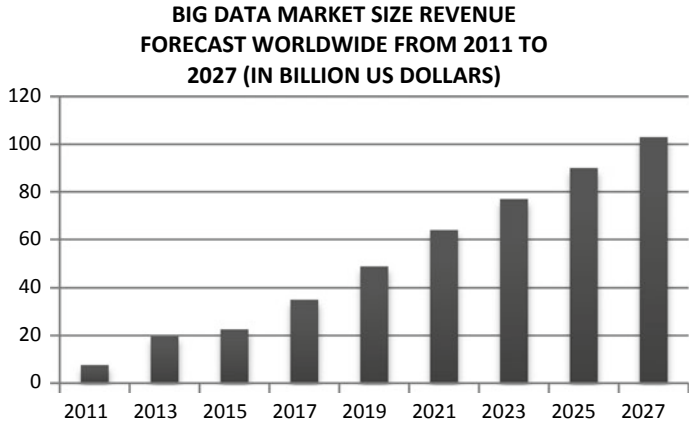


Fig. 1 Graph shows the big data market size revenue forecast worldwide [3]

be appropriately analyzed, and relating data should be separated [1]. "Big data" is a field that treats ways to deal with separate, proficiently extricate information from or by and large oversee informational indexes that are too much enormous or complex to be managed by conventional data-preparing application programming. Datasets with numerous cases offer more noteworthy measurable power, while datasets with higher complication may prompt a higher false discovery rate. Enormous information challenges consolidate getting information, information accumulating, information assessment, search, sharing, move, portrayal, addressing, invigorating, information security and information source. Big data was initially associated with three key concepts: volume, variety and velocity. Other thoughts later attributed to big data are veracity (i.e., how much noise is in the data) and value [2] (Fig. 1).

2 Literature Review

Faster and better decision making, with the speed of Hadoop and in-memory investigation, united with the ability to separate new wellsprings of information, associations can break down news rapidly—and choose decisions subject to what they have understood. New products and services, with the ability to quantify customer needs and satisfaction through examination, comes with the capacity to give customers what they need.

Big Data can be summarized as:

- Volume—It refers to the sheer size of the consistently detonating data of the registering scene. It brings up the issue of the amount of data.
- Velocity—It refers to the handling speed. It brings up the issue of at what rate the data is handled.

- Variety—It refers to the kinds of data. It brings up the issue of how unique the data organizations are.

Sources of Big Data: Machine Learning, a particular subset of AI that prepares a machine how to learn, makes it conceivable to rapidly and naturally produce models that can analyze more significant, progressively complex data and convey fast, increasingly precise outcomes—even on an exceptionally colossal scale. Furthermore, by exact structure models, an association has a superior shot of recognizing beneficial chances—or keeping away from unknown dangers. Data management, the information ought to be of high caliber and well spoken to before it will, in general, be continuously investigated. With information persistently spilling all through an affiliation, it is essential to set up repeatable systems to manufacture and keep up models for information quality. At the point, when information is reliable, associations should develop a piece of ace information, the executives' program that gets the entire undertaking in understanding. Data mining, information mining innovation allows a lot which makes you take a gander at a great deal of information to discover structures in the news, and this information can be used for further examination to help answer complex business questions. With information mining programming, you can channel through all the loud and repetitive tumult in information, pinpoint what is vast, that information can be used to assess likely outcomes, and a while later, enliven the pace of choosing instructed decisions [4, 5].

3 Case Studies

Definition of economy has evolved drastically over the years; dependent variables in monetary equations have changed with the advent of fourth industrial revolution. Online commerce has impacted in a big way, many vital constants contributing majorly for economy have been either changed or altered. Consider following case studies to build upon a predictive model with necessary components extracted from the case studies.

1. **E-Governance in India using Big Data**

- Analytics and data mining to detect tax evaders—Until now, the Indian government was shy about using machine learning and big data. With its help, it became very much more comfortable to catch the tax evaders in the year 2017. Nearly about 10 billion rupees are being found from the evaders [6–9].
- Track the flow of goods—As G.S.T. was introduced with the help of big data analysis, and with the use of G.S.T. and Railways, the flow of goods is being checked.

- Know the public mood—With the help of website *mygov.in*, the Prime Minister Office tries to know the sentiments of the public. By performing analysis, the Modi government was filtering the point to raise in the election campaign, and with the help of social media like Facebook and Twitter, they got to know the mood of the public [10, 11].
- Promising the tech-friendly future—With the help of big data, the government can track tax evaders, which proved that this government is going to give the tech-friendly future [12].

In the latest announcement, they try to make a national policy used in education, agriculture, retail, etc. Figure 2 depicts big data in Indian elections and described the platforms on which it has been used.

2. **5G Technology**

Dissimilar to 4G/LTE, 5G will be something beyond a pipe, it speaks to a reason-constructed innovation, and it is structured and built to encourage associated gadgets just as computerization frameworks. From numerous points of view, 5G will be a facilitator and a quickening agent of the following modern unrest, frequently alluded to as Industry 4.0 [13]. 5G guarantees to convey high data rates (in the scope of Gbps) with ultra-low inertness (not as much as millisecond delay) for applications in industrial automation, tactile Internet, robotics, AR/VR apps, and so on.

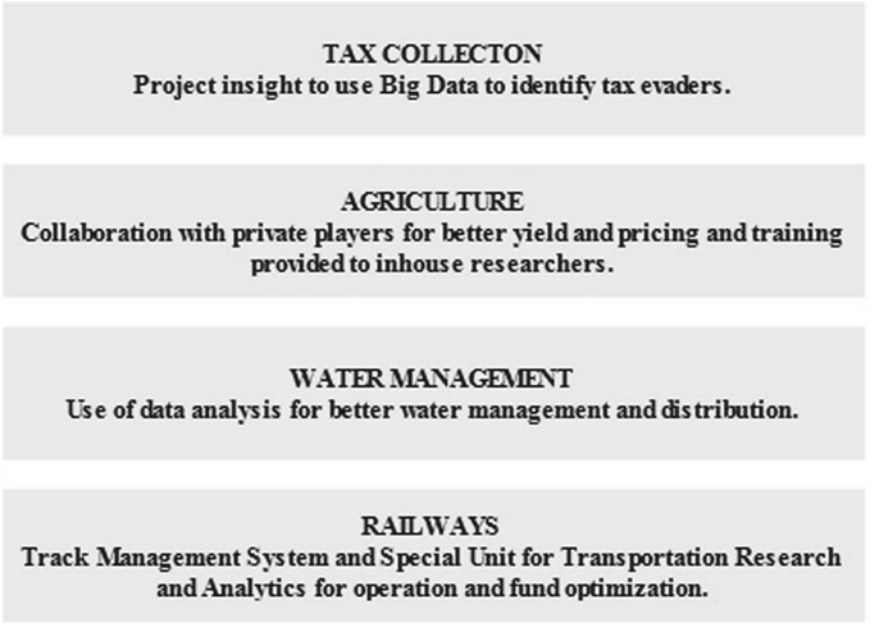


Fig. 2 How big data has been used in planning in the Indian election

Data analytics is at the sweet-spot exploiting 5G [14] arranges qualities, for example, high-transfer speed, low-latency, and mobile edge computing (MEC). 5G's capacity to help vast network crosswise over assorted gadgets (sensors/entryways/controllers), upheld by the appropriated register designs, makes the capacity to decipher the big data very still and the data-in-movement into constant bits of knowledge with unique insight.

Key Technology and Business Drivers are

- IoT over 5G (industrial IoT)—The mass measure of data being made by the IoT which can reform everything from assembling to human services to the format and working of shrewd urban communities—enabling them to work more effectively and gainfully than any time in recent memory. A fleet management company, for example, found that it had the option to lessen the expense of dealing with its armada of 180,000 trucks from 15 pennies for every mile to only 3 pennies [15].
- Data monetization—Telco's until 4G/LTE has been just utilizing information to improve administration quality and client experience. However, with the numerous conceivable outcomes of 5G system administrations joined with IoT and AI, they will investigate new business models of adaptation, for example, shrewd endeavor application administrations. For telco's, business openings lie in adapting information as well as the worth conveyed to ventures through application and system knowledge layers [16].
- Predictive maintenance—Predictive maintenance is the primary use-case of Industry 4.0. According to a statistical surveying report, predictive maintenance is an $11B showcase opportunity in the next five years. Predictive maintenance helps in predicting collapse before they happen by using AI [17].

Technical challenges and path to 5G involve following key factors:

- High-speed data-in-motion—On a very high-scale industrial IoT, smart cities and autonomous cars can pump the petabyte data within a minute. The 5G connectivity and the low-latency transmission will sum up the data throughput. Advanced cloud framework will be expected to help exceptionally fast read/write with the low-latency figure and the storage facility on the cloud.
- End-to-End security—Big data brings up various security issues, likewise with any applications today. In this way, it is essential to protect the client's security or enterprise data without any compromise. Building a robust, secure foundation from systems to applications will be critical in 5G structure and engineering [18].
- Real-time actionable insights—While low-latency is a characteristic for 5G systems, it turns into an extremely basic prerequisite for 5G to help quick data moves into the cloud, analytics at the edge and progressively, and continue the data at ultra-low-latencies to take ongoing activities in mission-critical applications, viz public safety, emergency care, and security surveillance.

3. **Walmart**

Walmart is one of the largest retailing company, which is based in America and was established in 1962, by Sam Walton. Until now, Walmart has 11,348 stores globally, and the annual turnover is $514.4 billion.

Walmart and Big Data—Walmart has a tremendous big data environment. The big data works in the petabytes. The gathered massive informational indexes are breaking down and mined toward the prescient examination, for streamlining the activities and business by the forecast of the habits for the clients. The retailers planned to attain maximum profit by knowing about the customers and, the factors and the consequences on the sales [19], the examinations cover millions of products and consumers. Walmart records a rise of 10% to 15% in online sales, which gives a turnover of about $1 billion. Big Data analysts very easily recognized this change in sales before and after the big data analysis.

Some techniques used by Walmart are:

- Savings catcher. The company notifies to customers, with the help of an application when its competitor reduces the price of the item. This application sends a gift voucher to its particular customer.
- Mapping. It works on Hadoop, to look after the Walmart stores most recent maps globally. This application can recognize even a small candy in the store globally [20–22].
- Track customer data. With the help of data mining technique, the sales pattern of Walmart has been observed. This pattern gives suggestions for the products to the customers, based on their previous purchase. The big data algorithm studies customer's interests, their purchase in-store and online, and also look after what is trending in the market.
- Launch new products. Walmart is utilizing Internet-based life information to discover about the slanting items with the goal that they can be acquainted with the Walmart stores over the world.

  – Social genome. It reads a large number of messages on Facebook and Tweets, watches YouTube videos and numerous blogs and analyzes the customer's requirements and the trending topics in the market, which help it to lead in the market.

Mobile big data solutions. More than 75% of customers are smart phones users. Moreover, with its help, they can have five more trips to the store online, which is about 70% more than the store visits [23]. With the help of cookies, the big data algorithm in mobile sections works more actively as per the user's previous activity, and there is no hassle created. Figure 3 depicts how Walmart used big data to become the leading retail store globally.

4. **Healthcare**

The healthcare field is vast, and it becomes essential to take the patient care and innovate medicines; so, the new technologies have been accepted by the industries. Big data in healthcare means gathering, dissecting, and utilizing the patient's physical and clinical data. According to research, the big data in the healthcare market is to be $34.27 billion by 2022 [24].

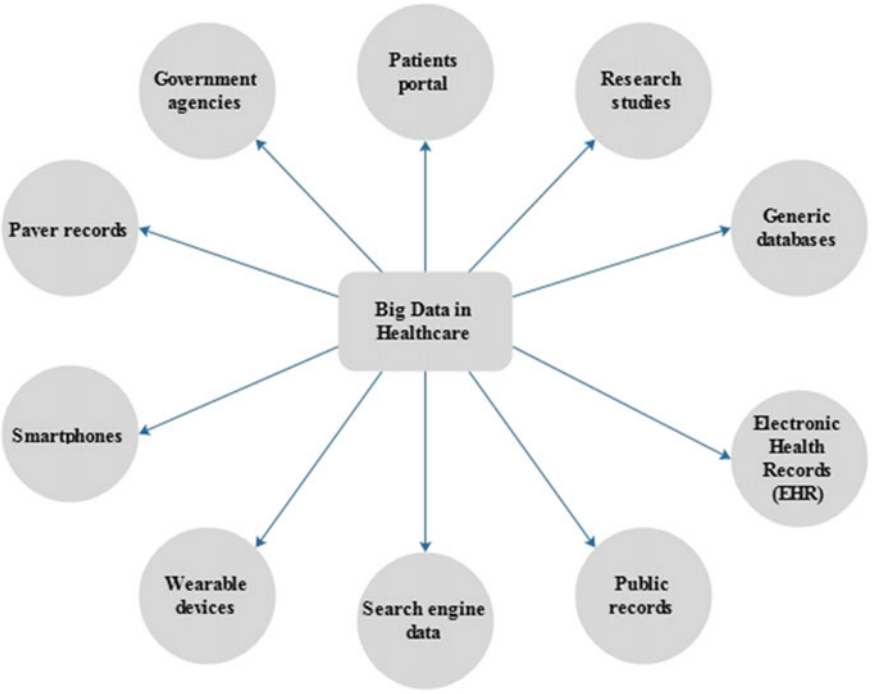


Fig. 3 Use of big data in the healthcare sector

Big Data can change the entire world of the healthcare field in the following ways:

- Tracking health. Big Data analytics and Internet of Things (IoT) can easily track anyone's statistics. Despite various wearable, there are many inventions which can detect the patient's blood pressure, heart rate, and other physiological parameters. With continuous monitoring of the body, people can be aware of their disease before the situation gets worst [25, 26].
- Cost reduction. With the help of big data, there could be a control over the staff members, which will decrease the investment rate in the staff. It will save time and money of the patients, and all the facilities will be easily available for the patients because of less crowd [27].
- Take care of high-risk patients. As all the data are digitalized, so it will be very helpful for the doctors to study the pattern of the high-risk patients, who are suffering from any chronic disease [28].
- Errors prevention. Any error due to wrong diagnosis in the medicine or wrong reporting may lead to even loss of lives sometimes. With the help of Big data techniques, medical practitioners would be able to counter check the prescriptions but making the treatment more effective and automated.

- Advancement in Healthcare Department. With the help of artificial intelligence (AI), it will be very easy to search for various data within seconds. Diagnosis of diseases and providing appropriate treatment will become very easy.

Figure 3 depicts big data in the healthcare department, so that the things do not become a hassle.

5. **Facebook**

It has been 11 years since Facebook started, and it is still so vigorously expanding that it has near about 1.59 billion accounts, which is about 1/5th of the world's total population. Facebook gathers a tremendous measure of information—a specific server sends out tens or hundreds of measurements that can be diagramed. This is not simply framework-level things like CPU and memory, and its additionally application-level measurements to comprehend why things are occurring.

Developing with Big Data, Facebook is one of the big data specialists and deals with the data in petabytes, including the analysis and the analytics. When people are coming closer on the platform of Facebook, it develops an algorithm to detect those relations. Facebook examines every bit of data to provide better services [29, 30].

Technologies behind Facebook's Big Data—Hadoop—Facebook holds the largest cluster of Hadoop in the world. It works beyond 4000 machines and stores data in thousands of petabytes [31, 32]. The developers can freely write map-reduce programs in any language.

Most of the files, *n* Hadoop are in tabular form. So, it becomes very easy for the developer to manage the subsets of SQL.

As Hadoop for Facebook is very efficient and reliable, it includes searching, video and image analysis, and data warehousing. The Facebook Messenger, developed by Facebook, is based on the Hadoop database, i.e., Apache HBase, which has an architecture that supports the overflow of messages.

- Scuba—Facebook faces numerous unstructured data on daily basis. To handle it, SCUBA was developed, which could help the Hadoop designers plunge into the enormous informational indexes and continue impromptu investigations progressively. Initially, Facebook was not able to handle such a huge amount of data single-handedly, and which cause the entire platform to crash. So, another platform for big data, Scuba allows the developer to store big data sets [33].
- Hive—In the world of unstructured data, Facebook became very popular by using Hive and the subsets of SQL, which makes Facebook run very fast and smooth without being a crashed [34, 35].
- Corona—It performs multiple tasks at the same time on a single Hadoop cluster without being crashed. Map-Reduce was designed based on pull-based scheduling model, which lags in performing the small tasks. Hadoop was limited by its slot-based resource management model, which was wasting the slots each time. Developing Corona helped in separating the clusters' job coordination.

6. **Astronomy**

Astronomy is the study of the physics, chemistry, and evolution of celestial objects and phenomena that originate outside the Earth's atmosphere, including supernovae explosions, gamma-ray bursts, and cosmic microwave background radiation.

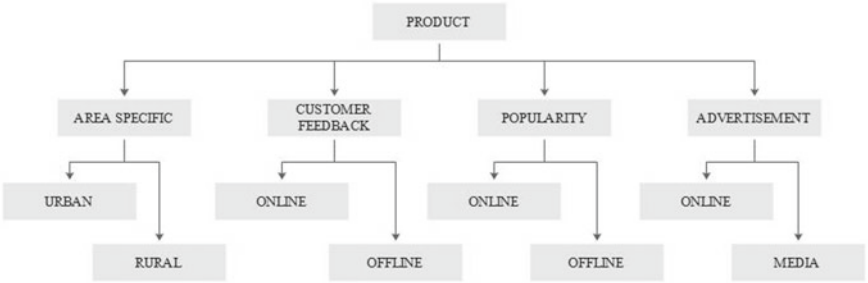
Data mining innovation makes you take a gander at a great deal of information to discover structures in the news, and this information can be used for further examination to help answer complex business questions [36].

Some of the data mining software and tools which are used in the Astronomy:

- *StatCodes*—It is a Web meta Webpage that gives hypertext connects to countless factual codes helpful for space science and related fields. It is being kept up at the Center for Astrostatistics site.
- *VOStat*—It is a GUI wrapper in the R language [37]. It does not just perform different examinations, including plotting, information smoothing, spatial investigation, time arrangement examination, outline, fitting appropriation, relapse, various kinds of factual testing, and multivariate procedures; however, it likewise plots intuitive 3D visualizations [38]. The main goals of VOStat are to encourage astronomers to use statistics and spread the use of R among astronomers.
- *Weka*—It executes machine learning algorithms for different information mining undertakings, for instance, information pre-preparing, characterization, relapse, bunching, affiliation standards, and perception. Likewise, it can grow new machine learning plans [39–41].
- *AstroML*—It is a Python module produced for AI and information mining, which is based on numpy, scipy, scikit-learn, matplotlib, and astropy and is disseminated under the three-condition BSD permit. To adequately break down cosmic information, it incorporates a developing library of measurable and AI schedules in Python and a few transferred open galactic datasets and gives an enormous suite of instances of examining and envisioning galactic datasets. The objective of AstroML is to give a network storehouse to quick Python usage of basic devices and schedules utilized for measurable information investigation in space science and astronomy and to give a uniform and simple to-utilize interface to openly accessible galactic datasets [42].

4 Proposed Predictive Model

The main proposal of predictive modeling is the use of Big Data analytics with the help of various technologies in various sectors like media, government, manufacturing, healthcare department, etc., and establishing a better future by analyzing and visualizing the data and predicting the nearest best possible outcomes. Implementing the various technologies in the analysis so that there will be a bright future and better outcomes to be predicted for the better results, in the respective fields,



1. **Area Specific**.
    a. **Urban**. Analyzing the types of jobs which can be required for the better establishment of the cities, and planting the factories regarding it.
    b. **Rural**. Big Data can be used in the agricultural field, analyze and assure the quality and variety of grains for the particular geographical conditions; corresponding the output of every respective grain.
2. **Customer Feedback**. To improve the services based on the customer feedback. Feedback can be online or off-line and can be further analyzed for improvement in any particular area of service.
3. **Advertisement**. Based on the online activities of a customer, choice of products is streamlined and advertisement is customized according to choose of customers.
4. **Popularity**.
    a. **Online**. Using big data analytics and data visualizations can raise the popularity of the product can be analyzed using online platforms, reading the messages on Facebook, checking the tweets on Twitter, watching the videos on YouTube, and many more.
    b. **Off-line**. The feedbacks of the specific product by the neighbor or the relatives, and comparing with other similar products of the same retail price, get to know which has better assurance and is economical.

Proposed Algorithm:
Step 1: *Identify the specific area of Big Data*
Step 2: *Data analytics applied for,*

a. *Rural*
b. *Urban*

Step 3: *Obtain customer feedback,*

a. *Online*
b. *Off-line*

Step 4: *Evaluate popularity,*

a. *Online*

b. *Off-line*

Step 5: *Advertisement methodology,*

a. *Online*
b. *Media*

Step 6: *From analytics, the outcome is predicted employing,*

a. *Data visualizations*
b. *Hadoop*
c. *Map-Reduce*
d. *A.I.e.*
e. *Machine learning*.

5 Conclusion

Big data which is generated through the online services can be further utilized for creating some meaningful results. Through the use of low-cost technologies, big data analytics can predict the future that helps in making important decisions in terms of right advertisement, finding precise market for products and in improving the services of existing platforms. It is a revolution where technology and management combine for predictive analysis and business intelligence. We have successfully presented the case studies of the most popular platforms generating the Big Data and also elaborated on various software tools that are utilized in the analysis of Big Data.

References

1. Elgendy, N., Elragal, A.: Big data analytics: a literature review paper (2014)
2. Singh, D., Reddy, C.K.: A survey on platforms for big data analytics (2014)
3. Liu, S.: 09 Aug 2019. https://www.statista.com/statistics/254266/global-big-data-market-forecast/
4. Bifet, A.: Mining big data—current status, and forecast to the future (2013)
5. Wu, X., Zhu, X., Wu, G.Q., Ding, W.: Data mining with big data (2014)
6. Banumathi, A., Pethalakshmi, A.: A novel approach for upgrading Indian education by using data mining techniques (2012)
7. Li, H., Nie, Z., Lee, W.C.: Scalable community discovery on textual data with relation. http://www.ics.uci.edu/~mlearn/MLRepository.html. University of California, Department of Information and Computer science, Irvine, CA
8. Guha, S., Rastogi, R., Shim, K.: CURE: an efficient clustering algorithm for large databases. In: Proceedings of 1998 ACM6SIGMOD International Conference Management of Data (SIGMOD'98), pages 73–84 (1998); Zhang, S., Zhu, C., Sin, J.K.O., Mok, P.K.T.: A novel ultrathin elevated channel low-temperature poly-Si TFT. IEEE Electron Device Lett. **20**, 569–571 (1999)

9. Zhang, C., Xia, S.: K-means clustering algorithm with improved Initial center. In: Second International Workshop on Knowledge Discovery and Data Mining (WKDD), pp. 7906792 (2009)
10. Sharma, P., Moh, T.S.: Prediction of Indian election using sentiment analysis on hindi twitter (2016)
11. Almatrafi, O., Parack, S., Chavan, B.: Application of location-based sentiment analysis using twitter for identifying trends towards indian general elections 2014. In: Proceedings of the 9th International Conference on Ubiquitous Information Management and Communication, Article No. 41 (2015)
12. Sawhney, D.: Technology integration in indian schools using a value-stream based framework (2016)
13. Bassi, L.: Industry 4.0: hope, hype or revolution? (2017)
14. Muralidhar Somisetty.: Big data analytics in 5G (2018)
15. Khurpade, J.M., Rao, D., Sanghavi, P.D.: A survey on IoT and 5G network (2018)
16. Gregory, I., Landryová, L., Soldán, P.: The monetization of highly automated systems in SMEs (2011)
17. Roukounaki, A., Efremidis, S., Soldatos, J., Neises, J., Walloschke, T., Kefalakis, N.: Scalable and configurable end-to-end collection and analysis of IoT security data towards end-to-end security in IoT systems (2019)
18. Singh, M., Ghutla, B., Jnr, R.L., Mohammed, A.F., Rashid, M.A.: Walmart's sales data analysis—a big data analytics perspective (2017)
19. Dean J., Ghemawat, S.: MapReduce: simplified data processing on large clusters. Assoc. Comput. Mach. (2008)
20. Riyaz, P.A., Surekha, V.: Leveraging map reduce with hadoop for weather data analytics. OSR J. Comput. Eng. (2015)
21. Sharma, M., Chauhan, V., Kishore, K.: A review: MapReduce and spark for big data analysis. In: 5th International Conference on Recent Innovations in Science. 5: Engineering and Management (2016)
22. Nivargi, V.: Big data: from batch processing to interactive analysis, [Online]. Available: Accessed Sept 2017 (2013)
23. Singh, M., Bhatia, V., Bhatia, R.: Big data analytics solution to healthcare (2017)
24. Mian, M., Teredesai, A., Hazel, D., Pokuri, S., Uppala, K.: Work in progress—In memory analysis for healthcare big data (2014)
25. Rahman, F., Slepian, M., Mitra, A.: A novel big-data processing framework for healthcare applications big-data-healthcare-in-a-box (2016)
26. Li, L., Bagheri, S., Goote, H., Hasan, A., Hazard, G.: Risk adjustment of patient expenditures: a big data analytics approach (2013)
27. Koppad, S.H., Kumar, A.: Application of big data analytics in healthcare system to predict COPD (2016)
28. Dragan, I., Zota, R.: Collecting Facebook data for big data research (2017)
29. Bronson, N., Lento, T., Wiener, J.L.: Open data challenges at Facebook (2015)
30. singh Bhathal, G., Dhiman, A.S.: Big data solution: improvised distributions framework of hadoop (2018)
31. Bhandarkar, M.: MapReduce programming with Apache hadoop (2010)
32. Verel, S., Collard, P., Clergue, M.: Scuba search: when selection meets innovation (2004)
33. Wang, K., Bian, Z., Chen, Q., Wang, R., Xu, G.: Simulating hive cluster for deployment planning, evaluation and optimization (2014)
34. Fuad, A., Erwin, A., Ipung, H.P.: Processing performance on apache pig, apache hive and MySQL cluster (2014)
35. Yan, L.X.Y.: Machine learning for astronomical big data processing (2017)
36. Zhang, Z., Barbary, K., Nothaft, F.A., Sparks, E., Zahn, O., Franklin, M.J., Patterson, D.A., Perlmutter, S.: Scientific computing meets big data technology: an astronomy use case (2015)
37. Keka, I., Çiço, B.: Data visualization as helping technique for data analysis, trend detection and correlation of variables using R programming language (2019)

38. Rubel, O., Weber, G.H., Huang, M.Y., Bethel, E.W., Biggin, M.D., Fowlkes, C.C., Hendriks, C.L., Keranen, S.V., Eisen, M.B., Knowles, D.W., Malik, J., Hagen, H., Hamann, B.: Integrating data clustering and visualization for the analysis of 3D gene expression data (2010)
39. Charalampopoulos, I., Anagnostopoulos, I.: A comparable study employing weka clustering/classification algorithms for web page classification (2011)
40. Jenitha, G., Vennila, V.: Comparing the partitional and density based clustering algorithms by using weka tool (2014)
41. Saad, S., Ishtiyaque, M., Malik, H.: Selection of most relevant input parameters using weka for artificial neural network based concrete compressive strength prediction model (2016)
42. VanderPlas, J., Connolly, A.J., Ivezić, Ž, Gray, A.: Introduction to AstroML: machine learning for astrophysics (2012)