# Trends and Future Perspective Challenges in Big Data

**Muhammad Naeem, Tauseef Jamal, Jorge Diaz-Martinez, Shariq Aziz Butt, Nicolo Montesano, Muhammad Imran Tariq, Emiro De-la-Hoz-Franco, and Ethel De-La-Hoz-Valdiris**

**Abstract** We are living in an era of big data, where the process of generating data is continuously been taking place with each coming second. Data that is more varied and extremely complex in structure (unstructured/semi-structured) with problems of indexing, sorting, searching, analyzing and visualizing are major challenges of today's organizations. Big data is always defined by its 5-v characteristics which are Volume, Velocity, Veracity, Variety, and Value. Almost each data model comprising big data is dependent on these 5-v characteristics. A large number of researches have been done on velocity and volume, but the complete and efficient solution for the variety is still not available in the markets. Traditional solutions provided by DBMS generally use multidimensional data type. However, many new data types cannot be compatible with these traditional systems. Big Data is a general problem affecting different fields, whether it is business, economic, social security or scientific research. To analyze huge data sets in order to get insights and find patterns in data is called big data analytics. Big data analytics is

M. Naeem
Friendly Health Technologies, San Ramon, CA 94583, USA

T. Jamal
Pakistan Institute of Engineering and Applied Sciences, Islamabad, Pakistan

J. Diaz-Martinez · E. De-la-Hoz-Franco · E. De-La-Hoz-Valdiris
Universidad De La Costa, Barranquilla, CUC, Colombia
e-mail: Diaz5@cuc.edu.co

E. De-la-Hoz-Franco
e-mail: edelahoz@cuc.edu.co

E. De-La-Hoz-Valdiris
e-mail: edelahoz3@cuc.edu.co

S. A. Butt (✉)
The University of Lahore, Lahore, Pakistan

N. Montesano
Iinformatica Srls, Trapani, Italy
e-mail: nicolo@iinformatica.it

M. I. Tariq
The Superior University, Lahore, Pakistan

J.-S. Pan et al. (eds.), *Advances in Intelligent Data Analysis and Applications*,
Smart Innovation, Systems and Technologies 253,
https://doi.org/10.1007/978-981-16-5036-9_30

the need of every corporate and state of the art organization to look forward and make useful decisions. This paper comprises of discussion on current issues, opportunities, trends, and challenges of big data aimed to discuss variety in more detail. An efficient solution for the big data variety problem will be discussed.

**Keywords** Big data · Big data challenges · Big data approaches

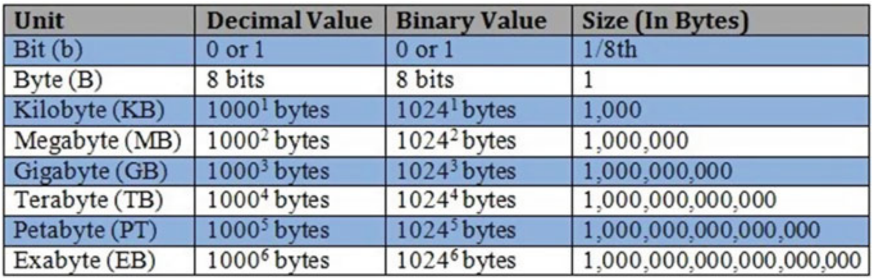
## 1 Introduction

Data is generating exponentially from different sources that require special mechanisms to store, process and analyze [1]. This huge amount of data generation is alarming situation for scientist and large organizations like Google, Microsoft, YouTube, IBM, Twitter etc. [2]. Roger Magoulas first introduces big data in 2005 by defining huge volume of data this is not possible to process for traditional database systems and applications due to its huge amount and complexity in structure. Madden defines big data as data that is extremely big, extremely hard and extremely fast to process by existing infrastructure.

According to MarTech internet consists of 2.7 zettabytes of data until 2017 and in 2010 it will grow to 44 zettabytes by Forbes report 2015 [3]. Some data generation sources are following below:

- 4+ billion Active internet users.
- 2.5 quintillion bytes of data are generated every day.
- 223+ emails sent (mostly spam emails).
- 5.5+ billion Searches on google.
- 5.9+ billion Viewed every day on youtube.
- 69 million pictures uploaded on Instagram.
- 272 million calls on Skype each day.
- 100,900 hacked websites.

500 GB of data held by each personal computer today, so to store 20 all worlds' data we require 20 billion PCs [4]. Multimedia poses a huge weight on internet bandwidth and will grow 70% in the next 5 years [5]. There are currently more than 6 billion mobile subscriptions and in the year 2020, it is expected that almost 50 billion devices will join network and internet [6]. This section consists of the following sections, Sect. 2 comprises of characteristics and challenges of big data, Sect. 3 comprises of literature review, Sect. 4 comprises of observations and Sect. 5 concludes the work (Fig. 1).



| Unit | Decimal Value | Binary Value | Size (In Bytes) |
|---|---|---|---|
| Bit (b) | 0 or 1 | 0 or 1 | 1/8th |
| Byte (B) | 8 bits | 8 bits | 1 |
| Kilobyte (KB) | $1000^1$ bytes | $1024^1$ bytes | 1,000 |
| Megabyte (MB) | $1000^2$ bytes | $1024^2$ bytes | 1,000,000 |
| Gigabyte (GB) | $1000^3$ bytes | $1024^3$ bytes | 1,000,000,000 |
| Terabyte (TB) | $1000^4$ bytes | $1024^4$ bytes | 1,000,000,000,000 |
| Petabyte (PT) | $1000^5$ bytes | $1024^5$ bytes | 1,000,000,000,000,000 |
| Exabyte (EB) | $1000^6$ bytes | $1024^6$ bytes | 1,000,000,000,000,000,000 |

**Fig. 1** Data size chart

## 2 Characteristic of Big Data

Big data has five basic challenges defined by 5-v characteristics [7]:

### 2.1 *Data Volume*

Data volume is a large amount to data generated by every second from Social media, sensors, cell phones, credit cards, etc. Almost 90% of data in the world is generated by the last two years. This large amount of data is too large so that we cannot process and analyze all the world's data at the same time. However, storage of that data is possible by big data tools and technologies.

### 2.2 *Data Value*

Data value means the worth of the data being generated, without analyzing and getting insight this huge amount of data is useless. This is not necessary that big data has a value in fact; there is a direct link between data and insight.

### 2.3 *Data Veracity*

Veracity refers to data quality, data trustworthiness, data accuracy. It also refers to the trustworthiness of data source from where we extract the data. Solving the problem like data duplication, inconsistencies, abnormalities are included in veracity.

### 2.4 Data Velocity

Data velocity refers to the rate at which data is growing. Emails, tweets on Twitter, likes, and comments on Facebook, photos on Instagram are increasing at light speed.

### 2.5 Data Variety

Data Variety refers to a different types of data. Data generated is of different types and complex structures. Data being different in types and complex in structure is very difficult to store, process and analyze. Example: A telecom organization extracts data from its multiple sources and loads it in a data warehouse. Now the extracted data is of multiple data type example log files, pdf, CSV, png, dat, etc. It is very difficult for an organization to perform ETL and to get insight into that data (Fig. 2).

## 3 Challenges of Big Data

### 3.1 Inadequate Understanding and Adoption of Big Data

Most of the times organizations are not aware about big data: what big data is, what is its benefits and challenges. Without a good understanding of tools and

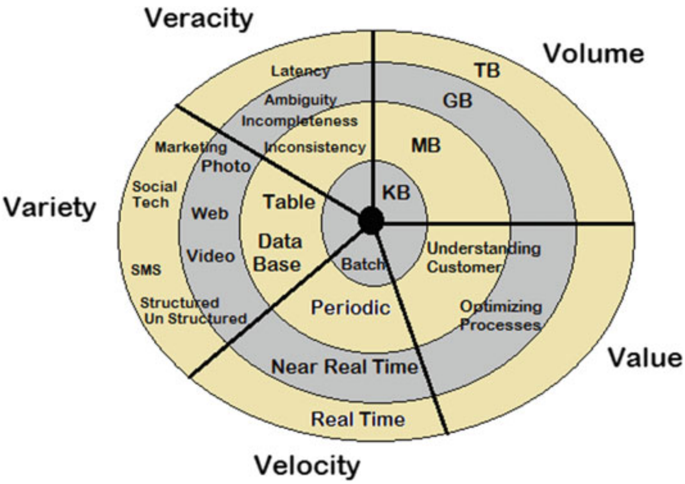


**Fig. 2** Big data characteristics

technologies of big data, organizations can waste their resources, time and they are unable to transform their current infrastructure to big data. Solution is top management in coordination with planning experts should adopt the big data first, after proper planning management should organize trainings and seminars to aware their employees about big data technology and then they should transform their legacy IT infrastructure to big data.

### 3.2 Bewilder with Big Data Technologies

In market today there are a lot of big data tools and technologies available, so employees can easily get confused that which technology to use whether spark or hadoop, which is used for storage whether Cassandra or HBase. IF the organization is new to big data they should first seek towards big data professional for consulting and then with coordination from experts they should choose required technology stack.

### 3.3 Complication of Managing Data Quality

By transforming traditional infrastructure to big data, organization will face the problem of data integration because data is coming from several heterogeneous sources in complex and different format. A telecom organization needs to be analyzing data coming from call centers, websites, mobile apps etc. Each source has its data in different data type and to store, process and analyze that data is major challenge. Solution is organization has to be design proper big data model through which they can manage big data and increase its quality. Find similarity between data from one truth point and merge the data having same entities.

### 3.4 Data Privacy and Security

This issue has major significance as amount of data increase risk also increases. This issue is very sensitive to organizations having technical and legal significance. Today big data technologies are evolving but their security features are still neglected putting big data organizations on high risk. Data Security includes personal information protection, product information protection, commercial and financial data protection, intellectual property protection. Most of the medium level organizations cannot afford big data infrastructure, so they have to rely on cloud services from third party, Involvement of third party increase high risk of data security. Solution to big data security problem is to consider security first at the time of designing the big data architecture because if one neglect security features in design phase it will bite at later stage.

### 3.5 Big Data Analytics Challenge

Big Data is growing with its huge analytical challenges. Big data analytics is the process of transforming raw data into useful insights for better decision and strategy making. Because amount of data under discussion is very large and unstructured analytics requires expert levels skills. So it requires skillful data engineering teams that make data ready to be analyzed by data scientists.

## 4 Big Data Techniques

Big Data requires high value techniques to perform analytics and get hidden trends in data. Big data techniques are composed of many theoretical concepts based on statistics, mathematics and computer programming. Applications based on big data are used to perform various task examples: stock price prediction, load and price forecasting in marts, weather forecasting, marketing campaigns, pattern recognition etc. These techniques are widely studied in academia and a lot of work is going on to improve their performance.

### 4.1 Optimization Methods

It is used to solve many big data problems examples: quantitative problems in Biology, Physics, Mathematics, Statistics, and Economics etc. These methods are also useful for solving global optimization problems such as simulated annealing and genetic algorithms. Stochastic optimization problems, many nature inspired algorithms such as Ant colony optimization, particle swarm optimization, fire fly optimization algorithms, evolutionary programming are useful and requires specific optimization methods. However, these problems require large memory and time to execute. Some big data problems require real time optimization such as data reduction etc. [8].

### 4.2 Statistical Methods

These are used in various big data problems. Statistics is the science to collect the data, analyze data and find correlations between different objectives. Statistics sometimes provides us numerical explanations. However, traditional statistical techniques are alone not enough to optimize big data problems that are why many researchers proposed these techniques in extended form or proposed new techniques for data science problems [9]. Statistical learning and statistical computing are two hot fields for research; according to survey performed in [10] this research includes implementation of statistical algorithms in scale and parallel form.

### *4.3 Data Mining*

It includes extraction of useful patterns and information from huge sets of raw data by using its famous techniques such as clustering, association rule mining, classification and regression, anomaly detection. As compared to data mining big data mining algorithms are more challenging to implement. Data mining use machine learning algorithms to efficiently perform its task and use statistics for special task such as optimization, data distribution, objective function etc. Clustering is one of the famous algorithms of data mining which makes group of data having similar in nature. Now if we want to make clusters of big data its extended versions will be used such as K-Mean, hierarchical clustering [11, 12]. Another purpose of big data clustering is parallel and distributed implementations [13]. Another famous example is discriminant analysis, scientist are developing algorithms for improving results of discriminant analysis [14]. Another famous example is computational biology and bio-informatics which now a days are using data mining algorithms for a large scale gene database comparisons [15].

### *4.4 Machine Learning*

It is important aspect of artificial intelligence and data science, its purpose is to develop algorithms that learn from experience or empirical data without being explicitly programmed. Machine learning enables the systems using its algorithms to discover knowledge and intelligently make decisions automatically. When we talk about big data, these algorithms both supervised and un-supervised have to be scaled up to deal with big data. Deep learning is another hot trend in artificial intelligence [16]. Additionally, there are many tools in market to scale up machine learning such as map-reduce, IBM parallel machine learning toolbox. Example, SVM is most fundamental algorithm of machine learning which is used for classification and regression. Most of the time scalability problems affect its performance. Scientists develop parallel SVM for reducing time as well as memory consumptions [17].

### 4.5 Visualization Approaches

These are used to create tables, different types of charts, animations, diagrams and other type of displays to understand data [18]. Big data visualization techniques in big data concern are not easy because of 5-v problem but like other statistical and machine learning techniques, visualization approaches are also enhanced to cope with big data challenges [19]. Scientists use various feature reduction techniques to reduce the size of data so that they can easy represent data. Selecting the correct representation of data is very difficult and important [20].

### 4.6 Social Network Analysis

It is one of the hot trends now days connecting billions of peoples around the world. The main purpose of SNA is to find trend among peoples through social networks such as Facebook, Twitter, and Instagram. For example, a business organization uses SNA to find current trends for marketing purpose and increase their sale. Social network visualizations, graph query and mining are included in SNA [21]. The main obstacle regarding SNA is vastness of big data, analyzing the network consisting billions of peoples is computationally very expensive and requires large memory and time consumption (Fig. 3).

## 5 Overcoming Variety Challenges of Big Data

Big data is everywhere around us, from bank transactions to shopping marts, from medical data to scientific data, from space data to stock exchange data. To analyze and get useful information from this data, it is necessary to integrate data which is one of the major tasks in big data and analytics.

### 5.1 Extract Transform and Load

In large organizations having their own big data infrastructure, the overall process starts from extracting the data from operational data sources. The extracted data is in different formats such as log files, video files, signaling data, PDF files, CSV files, etc. It requires a large transformation to be analyzed by data analysts and data scientists. Data transformations include removing duplicates, removing redundancies, handling the missing values, removing noise and anomalies and finally transform the data from legacy format to data warehouse format. After extraction and transformation finally, data is loaded into the data warehouse then data analysts

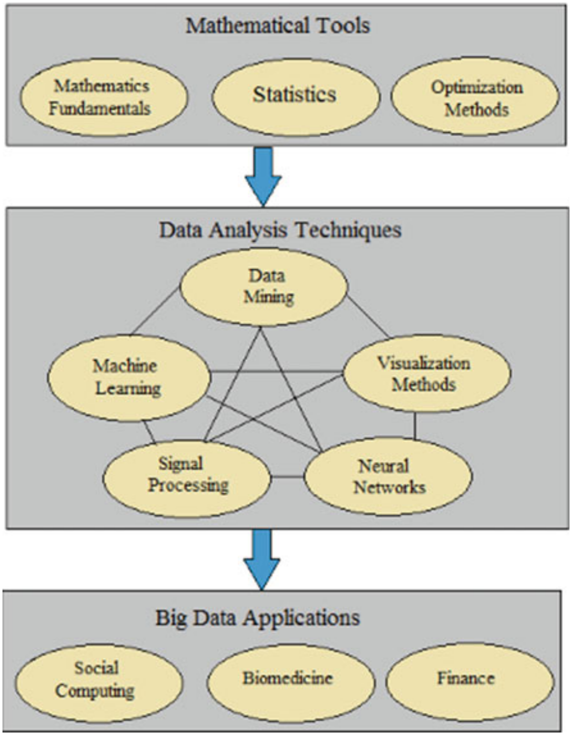


**Fig. 3** Big data techniques

again extract data from the data warehouse for ad-hoc query reporting. The above whole process of extracting, transforming and loading data is called the ETL process. The solution and it is one of the major processes in data warehousing. ETL takes more than 50% time of the overall data warehousing process (Fig. 4).

The difference between operational databases and data warehouse is, the database saves transactional data which is updated on a daily bases that's why the database is called online transactional processing (OLTP) and data in the data warehouse is subject-oriented, time-variant, integrated non-volatile data collection which is useful for decision making purpose. Data warehouse stores analytical data that's why it is called online analytical processing (OLAP). An operational database is used by the end-user while the data warehouse is used by data engineers. Operational database concerns with simple queries while OLAP involves large and complex queries such as aggregation of data. Data in the operational database is relatively small such as daily transactions while data in OLAP is data warehouse is extremely large.

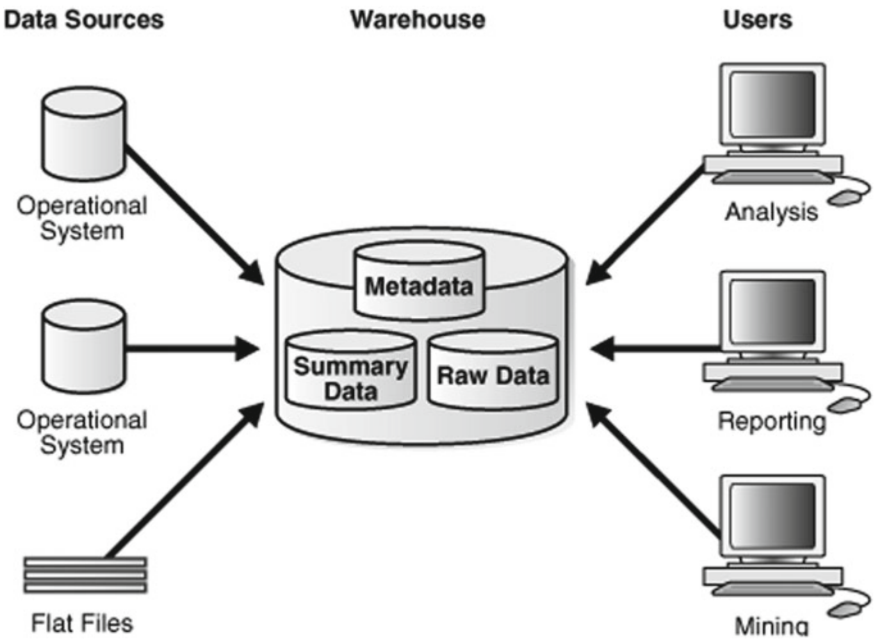


**Fig. 4** Basic architecture of DWH

## 5.2 Big Data Types

Integrating the data comes up with a major problem that data which is to be integrated is extremely huge and highly unstructured. Unstructured data means data has no proper format, consists of missing values, duplicates, data redundancies and cannot be processed by traditional tools. On the other side is structured data that is in the proper format (rows and columns) and easily processed and searchable by legacy infrastructure (Fig. 5).

## 5.3 Big Data Architecture

Before going into details it is necessary to understand how big data works in real-life scenarios [22]. The thing that makes the big data really big is the fact that it depends on extracting the data from different sources. Therefore, the main part of big data architecture is the Application Programming Interface (API). Additionally, other main important and an operational component of big data architecture is a columnar database. Columnar is the relational database that efficiently stores data in columns and rows resulting in faster performance. When it is the case of geographic data, a special type of database called spatial database use to store and retrieve

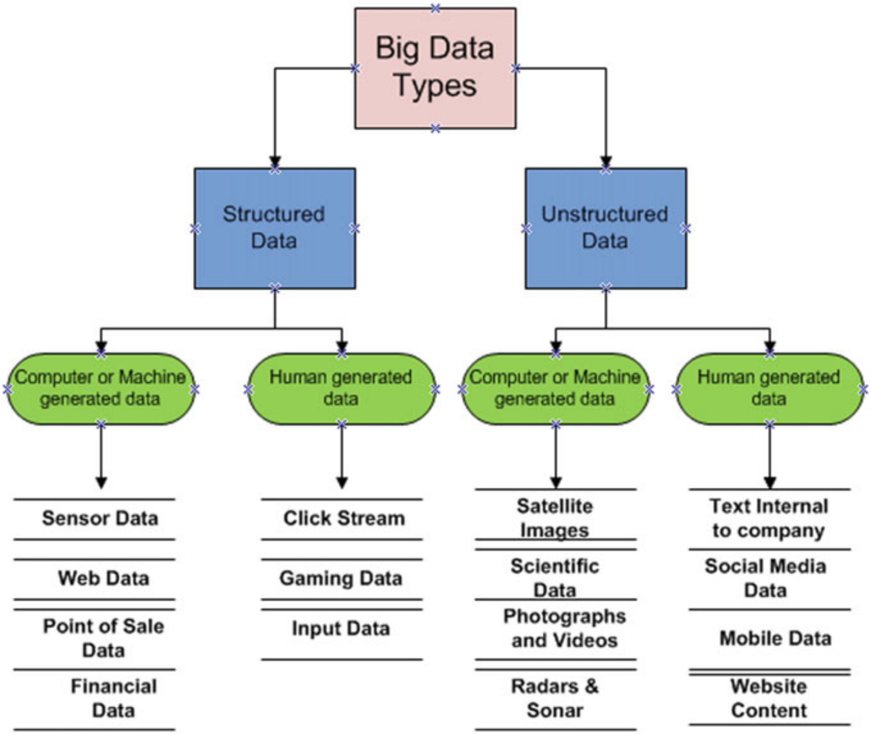


**Fig. 5** Big data types

space data from satellites directly [23]. Physical architecture is the basic and core part of big data infrastructure. To support an unpredictable and huge volume of data, robust infrastructure is required that is totally different from traditional IT infrastructure. Big data infrastructure is to be placed in a distributed way. Distributed means database servers are located in multiple locations connected by the network [24]. As big data is becoming important to industries, its privacy and security are also becoming crucial. For example, a medical organization uses big data applications to store the patient's data and maintaining the shift needs of patients. This data is very critical to that medical organization and needs to be protected in order to maintain the privacy of patients [25]. A company must ensure that its data sources are authentic and data coming from the sources are authentic and provides us a broader picture to understand key business needs. Traditionally the data was used by business organizations as an asset of the company and appended with enterprise applications generally. Now the hierarchy of these applications is changing and these companies are developing applications for a specific purpose and to take advantage of unique characteristics of big data [26].

# 6 Data Variety Possible Solutions

It is obvious that most of the data we get from multiple sources such as weblogs, sensors, multimedia, social media, e-mails, streaming, transactional data is highly unstructured. The real difficulty arises due to the data having different data types. It requires highly skillful teams, big data tools, and highly scalable infrastructure to process data having different varieties of nature. Processing of data having different data types can be made possible with state of the art OLAP tools. These tools enable the connection between data and information to gather data logically and experts can get top numbers in speed with low fall back time to process a huge quantity of data. All the data that is to be processed give to OLAP tools as input, OLAP tools process all the data whether it is relevant or not. This is one of the major disadvantages of OLAP tools [27].

## *6.1 Apache Hadoop*

It is one of the open-source software used to store and process big data with a very short interval of time with high performance that cannot be made possible by legacy tools. Hadoop uses the Map-Reduce algorithm to run its applications that can be processed in parallel. Hadoop provides the functionality to develop applications that perform complete statistical analysis on big data. It processes the data in a distributed way with a cluster of computers using a simple programming model. It is designed to start from a single server and can be extended to thousands of machines each has its own memory and storage [28]. Hadoop has many core components such as HDFS, Map-Reduce, Yarn, and Hadoop common. Hadoop common is multiple java libraries that are necessary for other Hadoop modules. Hadoop Yarn is used for scheduling of jobs and resource management in clusters:

**Hadoop Distributed File System** is a GFS based file system having distributed architecture and designed to run on commodity hardware. Although it has many similar functionalities as compared to previous distributed systems. The key difference is HDFS has low cost hardware and designed to provide high throughput to application data and useful for applications having large datasets.

**Map-Reduce** is a programming model and works on the basis of parallel computing and use to write applications that efficiently process a large volume of data, on huge clusters of hardware. Analytical capabilities on analyzing the huge volumes of data are provided by Map-Reduce. Map-Reduce is designed to replace the centralized storage infrastructure used by old organizations. Centralized servers are not suitable for processing huge data sets. Moreover, centralized systems are not able to process multiple files simultaneously and create a huge bottleneck. Map-Reduce solves this problem by dividing the tasks into different parts and assign them to various separate nodes for processing and finally integrate the result in one place.

### *6.2 Sap Hana*

SAP HANA is deployable on the cloud or as an on-premise machine. It is for the most part utilized for continuous examination and best appropriate for creating and sending constant applications. The center segment of this continuous stage is the SAP-HANA database, which is entirely unexpected from other regular databases in the market. Due to its multicore architecture in a distributed system environment, this revolutionary platform can easily overcome the variety and other types of relevant challenges in big data [29, 30].

### *6.3 Physical Infrastructure*

The perfect IT infrastructure can easily transform big data implementation after evaluating and making all the requirements about criteria:

- Performance: The performance of the system and cost both are important modules of any flawless infrastructure and both are directly proportional to each other. The cost will increase as the system performance will also increase.
- Availability: We need expensive infrastructure to be run all the time if we want system availability for 24 h. This is a very difficult task to perform as it requires proper datacenter to run high-speed servers all the time.
- Scalability: We have to make our big data infrastructure scalable by designing our infrastructure in such a way that its computing power and storage capacity can be extendable.

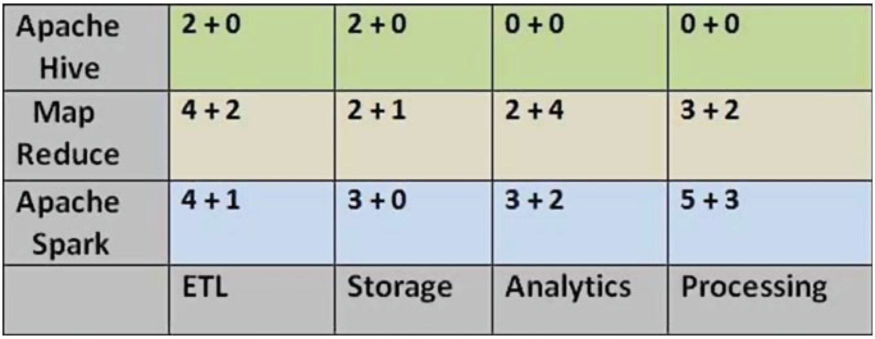
Data variety challenges can be solved by various tools discussed such as ETL tools, data visualizations, OLAP tools, having scalable and robust infrastructure, data reduction algorithms, highly skilled full teams and big data algorithms, etc. It is hard to say that anyone from above is alone required to solve a challenge or more than one tool is required or there exist some algorithms that can synchronize the data in a uniform format. It depends on the overall scenario. We cannot make a generalized solution for a variety of problems.

## 7 Literature Work

Literature finds a lot of work has been done to define big data and provides solutions to challenges concern with it. In [29] a detailed survey has been conducted on big data in which C.L. Philip Chen and Chun-Yang Zhang discussed intensive data applications, techniques, challenges, and technologies and concluded that humans are itself the creators of big data and dealing with big data is not a single domain problem, it requires multidisciplinary skills and methods to solve efficiently big data problems. Hiba Jasim Hadi and Ammar Hameed Shnain in [2] conducted a study on big data and its 5-v characteristics in which they discussed in detail that big data is very new technology to us and it requires technical as well as business skills to understand and make it useful in our daily life. In [31] dynamic resource allocation based on data characteristics (5Vs) for data streams has been discussed by Navroop Kaur and Sandeep K. Sood in which they proposed a predictor that predicts veracity, velocity, volume, variety, the variability of data in the context of big data streams in real time. A similar study was conducted in [32] in which Seref SAGIROGLU reviewed the various aspects of big data such as methods, advantages, challenges and security concerns. Another big data review has been conducted in [33] in which authors discussed the basic concepts of big data, an increase in data, the demand for big data and the role of big data in today's enterprises. Moreover, to enhance the efficiency of management of big data authors proposed data life cycles. In [7] applications of the big data have been discussed by authors Satan and Mishra and Vijay Dhote and they emphasize the business applications affected by huge volumes and complex structure of big data. Rui Mao, Honglong Xu, Wenbo Wu, Jianqiang Li introduced the concept of data abstraction in [34]. This article emphasizes on the big data variety challenges and state of the art indexing in metric space and worked on the pivot space mode. In [35] Silva Robak and Bogdan Franczyk discussed the big data utilization research problems in the design and management of the supply chains. The main aim of this article is to get maximum throughput from big data unique characteristics using advanced data science techniques in the field of logistics and supply chains. A big data analytics framework has been proposed for smart cities in [36] by Ahmed M and Shahat Osman. They reviewed the big data analytics characteristics applied in smart cities and design principles that will be used to design the data analytics framework. In [37] Gema Bello-Orgaz discussed the recent challenges and new achievements in social big data and concluded the currently available most relevant solutions for knowledge management in social media [38–40]. Conducted a survey on deep learning for big data analytics [41, 42]. Reviewed a survey on big data for health care by SAFA BAHRI1, NESRINE ZOGHLAMI1 in which they described the technologies and techniques of big data in healthcare organizations. Christina Orphanidou in [43] reviewed the big data applications in physiological signal data.

## 8 Observation

Big data is everywhere around us and the hot trend of today's information technology markets. As we know data is growing at an alarming rate, our traditional infrastructures are becoming useless and we need new, scalable, robust and optimized infrastructure to deal with big data. Big data is not a stand-alone problem or technology which cannot be solved by a single method or tool, it requires teams of highly skilled experts, robust physical infrastructure, mathematical and statistical models, distributed frameworks to process and analyze the huge volume of data.



| | | | | |
|---|---|---|---|---|
| Apache Hive | 2 + 0 | 2 + 0 | 0 + 0 | 0 + 0 |
| Map Reduce | 4 + 2 | 2 + 1 | 2 + 4 | 3 + 2 |
| Apache Spark | 4 + 1 | 3 + 0 | 3 + 2 | 5 + 3 |
| | ETL | Storage | Analytics | Processing |

**Fig. 6** Technologies and related research

Today big data cannot be ignored by organizations. However, it can be delayed for some period of time. For small organizations needs to depend on third-party cloud providers which raise serious privacy and security concerns means the rate of growing big data is proportional to the growing rate of security issue. Big data analytics in its initial stage is an efficient technique to get insights into data and find useful hidden patterns. However, still, tools and techniques are not enough for solving real-life big data problems completely. Scientific researchers and industry experts are still in the phase of designing more advanced tools and methods for real-life big data problems (Fig. 6).

## 9 Conclusion

This article emphasizes on challenges concerned with big data variety. Big data variety problems are not managed by traditional approaches, and then we started big data techniques, tools, and infrastructure to overcome big data variety problems. The challenges of big data resulted from more than 50 years of technology evolution means it is not a standalone technology or problem. In this paper, we reviewed current state, trends and future perspectives of Variety challenges in Big Data. We also reviewed big data characteristics, tools, techniques, and architecture. Solutions to overcome various challenges in big data are also discussed.

## References

1. Jagadish, H.V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J.M., Ramakrishnan, R., Shahabi, C.: Big data and technical challenges. Commun. ACM **57**(7), 86–94 (2014)
2. Fan, J., Han, F, Liu, H.: Challenges of big data analysis. Nat. Sci. Rev. **1**(2), 293–314 (2014)

3. Rabl, T., Gmez-Villamor, S., Sadoghi, M., Munts-Mulero, V., Jacobsen, H.-A., Mankovskii, S.: Solving big data challenges for enterprise application performance management. Proc. VLDB Endowment **5**(12), 1724–1735 (2012)
4. http://hpccsystems.com/. Last Accessed on 18 May 2019
5. Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A.H.: Big data: the next frontier for innovation, competition, and productivity (2011)
6. Gerhardt, B., Griffin, K., Klemann, R.: Unlocking value in the fragmented world of big data analytics. Cisco Int. Bus. Sol. Group **7** (2012)
7. Luo, J., Wu, M., Gopukumar, D., Zhao, Y.: Big data application in biomedical research and health care: a literature review. Biomed. Inf. Insights **8**, BII-S31559 (2016)
8. Shan, S., Gary Wang, G.: Survey of modeling and optimization strategies to solve high-dimensional design problems with computationally-expensive black-box functions. Struct. Multi. Optim. **41**(2), 219–241 (2010)
9. Di Ciaccio, A., Coli, M., Ibanez, J.M.A. (eds.): Advanced Statistical Methods for the Analysis of Large Data-Sets. Springer Science Business Media (2012)
10. Pbay, P., Thompson, D., Bennett, J., Mascarenhas, A.: Design and performance of a scalable, parallel statistics toolkit. In: 2011 IEEE International Symposium on Parallel and Distributed Processing Workshops and Phd Forum, pp. 1475–1484. IEEE (2011)
11. Zhou, J., Chen, C.L.P., Chen, L, Li, H.-X.: A collaborative fuzzy clustering algorithm in distributed network environments. IEEE Trans. Fuzzy Syst. **22**(6), 1443–1456 (2013)
12. Pentaho, B.I.: Getting Started with Pentaho Business Analytics. Pentaho Corporation (2012)
13. Ranka, S., Sahni, S.: Clustering on a hypercube multicomputer. IEEE Trans. Parallel Distrib. Syst. **2**(2), 129–137 (1991)
14. Cai, D., He, X., Han, J.: SRDA: an efficient algorithm for large-scale discriminant analysis. IEEE Trans. Knowl. Data Eng. **20**(1), 1–12 (2007)
15. Bertone, P., Gerstein, M.: Integrative data mining: the new direction in bioinformatics. IEEE Eng. Med. Biol. Mag. **20**(4), 33–40 (2001)
16. Hinton, G.E.: Learning multiple layers of representation. Trends Cogn. Sci. **11**(10), 428–434 (2007)
17. Bekkerman, R., Bilenko, M., Langford, J. (eds.): Scaling up Machine Learning: Parallel and Distributed Approaches. Cambridge University Press (2011)
18. Simoff, S., Bhlen, M.H., Mazeika, A. (eds.): Visual Data Mining: Theory, Techniques and Tools for Visual Analytics, vol. 4404. Springer Science and Business Media (2008)
19. Keim, D.A., Panse, C., Sips, M., North, S.C.: Visual data mining in large geospatial point sets. IEEE Comput. Graphics Appl. **24**(5), 36–44 (2004)
20. Thompson, D., Levine, J.A., Bennett, J.C., Bremer, P.T., Gyulassy, A., Pascucci, V., Pbay, P. P.: Analysis of large-scale scalar data using hixels. In: 2011 IEEE Symposium on Large Data Analysis and Visualization, pp. 23–30. IEEE (2011)
21. Kolari, P., Joshi, A.: Web mining: research and practice. Comput. Sci. Eng. **6**(4), 49–53 (2004)
22. Acharjya, D.P., Ahmed, K.: A survey on big data analytics: challenges, open research issues, and tools. Int. J. Adv. Comput. Sci. Appl. **7**(2), 511–518 (2016)
23. Che, D., Safran, M., Peng, Z.: From big data to big data mining: challenges, issues, and opportunities. In: International Conference on Database Systems for Advanced Applications, pp. 1–15. Springer, Berlin, Heidelberg (2013)
24. Michael, K., Miller, K.W.: Big data: new opportunities and new challenges [guest editors' introduction]. Computer **46**(6), 22–24 (2013)
25. Bologa, A.R., Bologa, R., Florea, A.: Big data and specific analysis methods for insurance fraud detection. Database Syst. J. **4**(4), 30–39 (2013)
26. Chung, P.T., Chung, S.H.: On data integration and data mining for developing business intelligence. In: IEEE Long Island Systems, Applications, and Technology Conference (LISAT), pp. 1–6. IEEE (2013)
27. Pattnaik, K., Mishra, B.S.P.: Introduction to big data analysis. In: Techniques and Environments for Big Data Analysis, pp. 1–20. Springer, Cham (2016)

28. Wu, X., Zhu, X., Wu, G.Q., Ding, W.: Data mining with big data. IEEE Trans. Knowl. Data Eng. **26**(1), 97–107 (2014)
29. Chen, C.P., Zhang, C.Y.: Data-intensive applications, challenges, techniques, and technologies: a survey on Big Data. Inf. Sci. **275**, 314–347 (2014)
30. Tariq, M.I., Tayyaba, S., Ashraf, M.W., Rasheed, H.: Risk based NIST effectiveness analysis for cloud security. Bahria University J. Inf. Commun. Technol. (BUJICT) **10**(Special Is) (2017)
31. Kaur, N., Sood, S.K.: Dynamic resource allocation for big data streams based on data characteristics (5 V s). Int. J. Netw. Manage. **27**(4), e1978 (2017)
32. Sagiroglu, S., Sinanc, D.: Big data: a review. In: International Conference on Collaboration Technologies and Systems (CTS), pp. 42–47. IEEE (2013)
33. Khan, N., Yaqoob, I., Hashem, I.A.T., Inayat, Z., Ali, M., Kamaleldin, W., Alam, M., Shiraz, M., Gani, A.: Big data: survey, technologies, opportunities, and challenges. Sci. World J. (2014)
34. Mao, R., Xu, H., Wu, W., Li, J., Li, Y., Lu, M.: Overcoming the challenge of variety: big data abstraction, the next evolution of data management for AAL communication systems. IEEE Commun. Mag. **53**(1), 42–47 (2015)
35. Robak, S., Franczyk, B., Robak, M.: Research Problems Associated with Big Data Utilization in Logistics and Supply Chains Design and Management. In: FedCSIS Position Papers, pp. 245–249 (2014)
36. Osman, A.M.S.: A novel big data analytics framework for smart cities. Futur. Gener. Comput. Syst. **91**, 620–633 (2019)
37. Butt, S.A., Jamal, T., Azad, M.A., Ali, A., Safa, N.S.: A multivariant secure framework for smart mobile health application. Trans. Emerg. Telecommun. Technol. e3684 (2019)
38. Butt, S.A., Jamal, T.: IoT smart health security threats. In: 19th International Conference on Computational Science and Its Applications (ICCSA) IEEE, At Pittsburgh, Russia. https://doi.org/10.1109/ICCSA.000-8 (2019)
39. Jamal, T., Butt, S.A.: Cooperative cloudlet for pervasive networks. Proc. Asia Pacific J. Multi. Res. **5**(3), 42–26 (2017)
40. Tariq, M.I.: Agent based information security framework for hybrid cloud computing. KSII Trans. Internet Inf. Syst. **13**(1) (2019)
41. De-La-Hoz-Franco, E., Ariza-Colpas, P., Quero, J.M., Espinilla, M.: Sensor-based datasets for human activity recognition–a systematic review of literature. IEEE Access **6**, 59192–59210 (2018)
42. Jamal, T., Butt, S.A.: Malicious node analysis in MANETS. Int. J. Inf. Technol. 1–9 (2018)
43. De la Hoz, E., de la Hoz, E., Ortiz, A., Ortega, J., Martínez-Álvarez, A.: Feature selection by multi-objective optimisation: application to network anomaly detection by hierarchical self-organising maps. Knowl.-Based Syst. **71**, 322–338 (2014)