Understanding Concepts, Future Trends, and Case Studies of Big Data Technologies

Khyati Ahlawat and Amit Prakash Singh

Abstract In this paper, several promising big data technologies and their case studies are discussed in the context of their application in a real-world scenario and their potential to meet growing demands of mainstream big data customers. The prevailing big data technologies include Predictive Analytics, NoSQL Databases, Real-Time Analytics, Data Virtualization, and Knowledge Discovery in Big Data. Each of these requires one or more different big data handing platforms and tools. A brief study of big data preparation and analytics platforms for the above-mentioned technologies is also given in the paper. Big Data Analytics is rapidly becoming popular, spawning the need to comprehend its associated technologies. This paper will provide a baseline study for many researchers who wish to gain knowledge on current behavior and future trends of big data technologies. Currently, numerous platforms are available, both in the online and offline mode to handle big data, so a detailed review of all of such technologies is out of the scope of this paper. Hence, it deals with ones which are highly accepted and popular.

Keywords Big data analytics · NoSQL databases · Real-time analytics Predictive analytics · MapReduce

1 Introduction

Any large sized and complex data, which limits the capabilities of traditional data processing applications and tools to handle it, can be termed as Big Data. It has certain characteristics that are defined as several Vs by various researchers [1], however, significant ones are Volume, Variety, and Velocity. Big data can comprise of all of

K. Ahlawat (✉) · A. P. Singh
University School of Information Communication and Technology, Guru Gobind Singh
Indraprastha University, Sector 16C, Dwarka 110078, Delhi, India
e-mail: khyatiahlawat@ipu.ac.in

A. P. Singh
e-mail: amit@ipu.ac.in

C. R. Krishna et al. (eds.), *Proceedings of 2nd International Conference on Communication, Computing and Networking*, Lecture Notes in Networks and Systems 46, https://doi.org/10.1007/978-981-13-1217-5_96

these characteristics or any one of them. Volume defines substantial amount of data that challenges storage and processing capacity of currently used tools and processors. Variety refers to different types of data present for analysis like multimedia data (audio, video) and textual data, images, social media data [2], etc. Velocity on the other hand, deals with the pace at which big data is getting generated and propagated for analysis. Streaming data analysis is one example of high-velocity big data where the challenge lies in the rapid analysis of constantly generating streams of data via different sources. Similarly, researchers in [3] have introduced HACE theorem suggesting major characteristics of big data as heterogeneous and diverse data sources, autonomous with distributed control, complex, and evolving in knowledge.

Analysis of big data by incorporating its characteristics to discover and predict valuable information from it is termed as Big Data Analytics. The basic aim of analysis is to extrapolate big data, interpret it, and assist in prediction, diagnosis, and decision-making. It is one of the major research areas in parallel and distributed systems in future generation due to its wide application areas like health welfare, government and public sector, e-commerce, social networking, and natural processes [4] etc.

KDD, Knowledge Discovery, and Data Mining, is a process comprising of basic three steps, i.e. input, analysis, and output. Currently, numerous techniques are available for normal-sized data analysis. However, new techniques and algorithms are required for big data analysis to match highly complex nature of big data. Currently available tools for big data concentrate on two types of analysis, batch processing and streaming data analysis [5]. Most established batch processing tool is Apache Hadoop platform. Apache Hadoop consists of two basic components, HDFS and MapReduce Paradigm. HDFS is Hadoop-Distributed File System and is used for storing voluminous big data in master–slave fashion. MapReduce Paradigm is based on the parallel computation of big data and taking into consideration computation time, parallel processing is the most important future trend [6]. Other tools for batch processing include Apache Mahout, Dryad, Tableau, etc.

On the other hand, streaming data tools include Apache Spark, Storm, S4, Kafka, etc. It is also known as real-time processing in which ongoing data processing requires a very low latency rate for efficient analysis. Apache Storm is mostly used tool for this purpose. It is scalable, fault tolerant, and easy to use and operate. Other tools include S4, Splunk, Kafka, SAP Hana, etc. [5] has very well demonstrated the differences between batch processing and real-time tools.

One needs to decide which platform is most suitable for which type of data according to the application area it is being generated from. For this purpose, various factors are important to consider like data size, throughput optimization, training of model, scalability, etc. [7]. Rest of the paper is as follows: Sect. 2 describes various technologies in big data domain covering almost all available techniques and areas along with a discussion of some case studies. Section 3 discusses trends and forecasts for big data technologies according to Forbes. Section 4 is conclusion and future scope of this chapter.

2 Technologies in Big Data Domain

Big data, consisting of some unique features, gets generated from diverse sources in various formats like batch and streaming. It is observed that most of the big data that is generated is unstructured. Challenge lies in gaining insights into such data directly or by first converting it to structured data and then analyzing it. Storage and analysis of structured or unstructured data are always a challenge when size of data increases enormously. With the evolution of big data, techniques related to it are also developing to efficiently analyze it and extract important information from it. This section discuses top-level technologies in Fig. 1 that are currently being worked on in the big data domain. These include both for storage and analysis purpose.

Each type of big data technologies described in Fig. 1 such as Predictive Analytics, NoSQL Databases, Real-Time Analytics, Data Virtualization, and Cloud Computing are discussed next.

2.1 Predictive Analytics

Predictive analytics is an advanced level of analytics in which predictions are made for future events based upon previously available data. It is also known as Classification. For this purpose, machine learning algorithms are used to train the machine based on historic data and then it is tested for the quality of predictions it can make. Many practitioners are working toward applying existing machine learning techniques in big data domain or scaling them to map big data level.

There are multiple platforms for machine learning in big data scenario as discussed in [8] like Mahout, H2O, SAMOA, Apache Spark, etc. Mahout is the most well-known platform with a wide range of machine learning algorithms that are basically focused on classification, clustering, and collaborative filtering. It has some shortcomings also like its inefficient runtimes due to slow MapReduce principle and difficulty to set it on a Hadoop cluster. On the other hand, Apache Spark is a recent

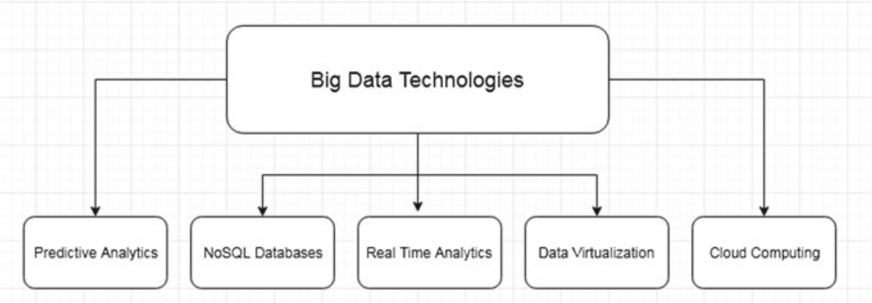

Fig. 1 Big data technologies

platform which is based upon MLlib, machine learning library. MLlib implements all algorithms of Mahout along with some regression models, feature extraction, etc., which is the lacking factor for Mahout.

H_2O is an enterprise edition with a notable feature of GUI, graphical user interface, and deep neural network toolkits. Users can work on this platform via programming or web-based interface. SAMOA stands for Scalable Advanced Massive Online Analysis and is a machine learning platform for streaming data. It is highly scalable and extensible but with low coverage of algorithms.

Apart from this, one more promising platform for machine learning in big data is R [9]. Though R provides more than 70 packages but they lack scaling to big data. Therefore, R has provided specialized packages for big data like bigrf, biglm, etc. Bigrf provides Random Forest functionality on big data while biglm deals with linear regression for data that is too large to fit in memory. R also provides an interface with Hadoop known as Radoop, which works in the form of map reduce jobs on the data stored in HDFS. Researchers have discussed and compared various machine learning platforms on big data in detail which is in [9].

In addition to this, a general-purpose framework, Petuum is introduced in [10] that systematically addresses data and works efficiently in comparatively less time.

Case Study: For instance, started as an online bookstore, Amazon has led its journey to top-level online retail store. Product recommendation is used based on customers profile and their past experiences in amazon which serves as a major big data challenge. It uses the concepts of content-based and collaborative filtering to group customers and analyzes their preferences.

2.2 NoSQL Databases

In this era of big data analytics, a variety of data in different formats are available and traditional relational databases are becoming obsolete to manage them. In their place, NoSQL databases are providing promising and efficient data storage and analysis platform to handle big data. Some popular NoSQL databases include Redis, MongoDB, Couchbase, Cassandra, HBase, etc. These can be grouped based on their characteristics in four categories as Key-value Store, Document Store, Column Family Store and Graph databases [11].

Key-value Store is where data is stored in the form of key-value pairs as hash tables where values can be easily accessed with the help of corresponding keys. In document store, semi-structured documents are stored and managed in formats like XML or JSON. Column Family Store is where data is stored in the form of rows and columns and based on similarities of columns, related columns formulate column family. Graph databases store information in the form of graph data structure and make use of graph algorithms to analyze them. It is mainly used to represent recommendation system or social networking data.

Many researchers are working in this field to analyze the capabilities and performance of NoSQL databases to handle big data. In [12], it is concluded that Redis which is a key value data store has performed better in terms of efficiency and is found well suited for loading and executing workloads. Apart from this, practitioners have also proposed a fuzzy-based methodology in [13] to store consistent fuzzy geospatial data based on certain validations and constraints for semantics. Researchers in [14] have addressed the problem of data migration in heterogeneous NoSQL databases by providing a fault-tolerant approach to migrate data. In some applications, a hybrid architecture of relational and NoSQL databases can prove useful. In [15], a data adapter system to support hybrid database model supporting both types of databases with three modes for query approach.

Case Study: EHR system as discussed in [16] replaced existing system to NoSQL database for primary data store and to prepare a local cache at each site to improve request latency and availability.

2.3 Real-Time Analytics

Real-time big data analytics is when the timely prediction of streaming data is provided. This is a very challenging task in big data scenario as storage and analysis of big data at the same time requires large in-memory and correct platform. Its applications include majorly sensor related data and social web data. Examples of real-time analytics tools are S4, Storm, and Apache Spark.

In recent times, the role of streaming analytics has been increased in the field of sentiment analysis. A distributed system based on Vertical Hoeffding Tree, a decision tree classifier for same has been presented in [17]. Apache Samoa, a tool for distributed streaming classification along with evaluation criteria to measure the performance of streaming classifiers is depicted in paper [18].

Case Study: A case study on finding the effectiveness of Storm to analyze streaming data in real time from two applications to predict trending news is given in [19].

2.4 Data Visualization

Visualization is the concept of representing the results and analysis part of big data in a graphical manner. It is becoming a concerned technology because traditional visualization approaches do not map well to big data scenario. In case of a large-scale data visualization, processing methods such as feature extraction and geometric modeling are also used. In the current scenario, researchers are actively working on searching for new visualization platforms to scale them for big data. Here, augmented reality techniques are proving to be next milestone in this scenario. In [20], practitioners have discussed utilizing virtual reality and augmented reality techniques for big data visualization. A new algorithm for visualization of big data from high dimension to a two-dimensional space is presented in [21] and found to perform noticeably well.

Similarly, in [22], high dimensional big data is visualized in three dimension using unsupervised dimension reduction techniques.

Case Study: To maintain the consistency and integrity of constantly increasing data in Qualcomm, a data virtualization solution was produced to make data available fast and in the virtual view.

2.5 Cloud Computing

Distributed computing along with Internet in an extended form can be conceptualized as cloud computing. Here, cloud refers to a network which can be accessed via the Internet anytime and anywhere to use it as any service. With the evolution of big data, storing and analyzing big data in clouds seems to be perfect solution to almost all data storage problems. Still, there exist some issues in this solution out of which security is most important. In [23], an efficient approach to handle secure storage of big data in a cloud environment is presented which is based on compressing and decompressing of data.

One more problem associated with distributed cloud computing is heavy data traffic among data centers. This leads to high communication costs in query evaluation. An online query evaluation system to deal with this problem of big data in distributed clouds is discussed in [24].

Case Study: Several case studies in this research area are discussed in [25] namely SwiftKey, Nokia, redBus, etc. SwiftKey is a language technology that uses Apache Hadoop on Amazon Cloud for scalable and multilayered model along which applies artificial intelligence component to manage constantly increasing data in prediction problems. redBus is responsible for internet bus ticketing system in India and has implemented its system using Google Query for large data analysis. On the other hand, Nokia makes use of Hadoop Distributed File System to store multi-structured data in cloud environment.

3 Trends and Forecasts for Big Data Technologies

Big data market is constantly expanding in mainstream business and there is a need to analyze which big data technologies are in demand. According to Forbes, an analysis of growing requirements of big data technologies is done and a trajectory of more than 20 techniques in big data scenario is presented as Fig. 2.

As shown in this figure, a wide range of big data technologies are shown as per their business value in coming time. Predictive analytics is the technology that is highly popular in today's business needs and the only one with maximum time for next phase. Apart from it, NoSQL databases, Stream analytics, search and knowledge discovery are some other technologies that are currently in demand. No big data technology is achieving the minimum success line on graph showing the promising

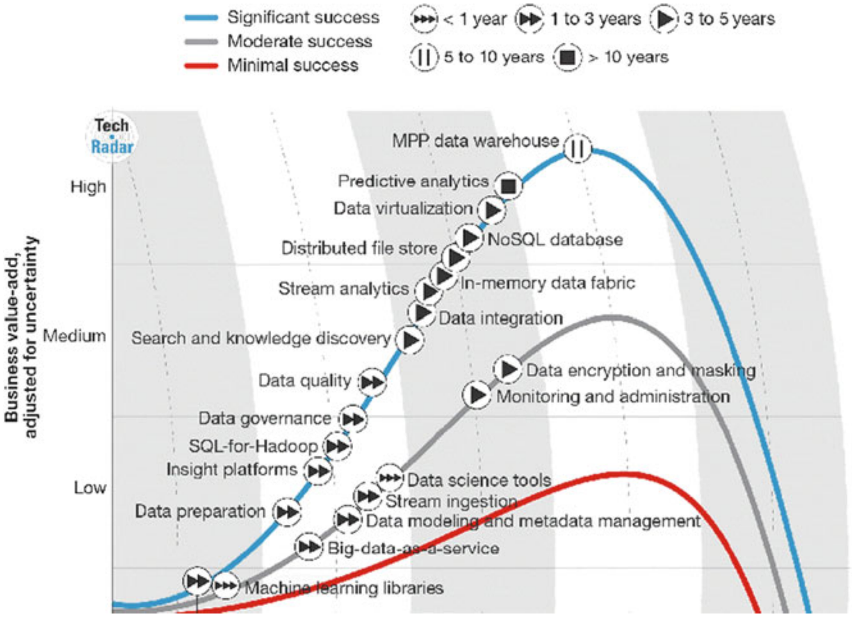

Fig. 2 Big data technologies described by Forbes

nature of the technologies in this area. Very few technologies namely data science tools and machine learning libraries are achieving moderate success and are likely to be popular in less than 1 year. Based on TechRadar methodology of Forrester, above-described technologies in this paper are most popular and likely to remain persistent for more than 5 years [26].

4 Conclusion

Big data technologies are trending in almost all present business requirements and their understanding has become a burning need in this field. In this paper, a brief study on some popular big data technologies like predictive analytics, NoSQL databases, Real-time analytics, cloud computing, and data virtualization is discussed. Prediction for future trends of popular techniques and their lifespan is also presented as described by Forbes. Sample case studies for each discussed technology are also given to understand the scope and purpose of this study.

References

1. H. Hu, Y. Wen, T.S. Chua, X. Li, Toward scalable systems for big data analytics: a technology tutorial. IEEE J. Mag. **2**, 652–687 (2014)
2. A. Gandomi, M. Haider, Beyond the hype: big data concepts, methods and analytics. Int. J. Inf. Manage. **35**(2), 137–144 (2015). Elsevier
3. X. Wu, X. Zhu et al., Data mining with big data. IEEE Trans. Knowl. Data Eng. **26**(1), 97–107 (2014)
4. K. Kambatla et al., Trends in big data analytics. J. Parallel Distrib. Comput. **74**, 2561–2573 (2014). Elsevier
5. C. Chen, C. Zhang, Data-intensive applications, challenges, techniques and technologies: a survey on big data. Inf. Sci. **275**, 314–347 (2014). Elsevier
6. C.-W. Tsai et al., Big data analytics: a survey. J. Big Data **2**(1), 21 (2015). Springer
7. D. Singh, C. Reddy, A survey on platforms for big data analytics. J. Big Data **2**(1), 8 (2014). Springer
8. S. Landset et al., A survey of open source tools for machine learning with big data in the Hadoop ecosystem. J. Big Data **2**(1), 24 (2015). Springer
9. J. Zheng, A. Dagnino, An initial study of predictive machine learning analytics on large volumes of historical data for power system applications, in *International Conference on Big Data, IEEE* (2014), pp. 952–959
10. P. Eric Xing et al., Petuum: a new platform for distributed machine learning on big data. IEEE Trans. Big Data **1**, 49–67 (2015)
11. R. Zafar et al., Big data: the NoSQL and RDBMS review, in *International Conference on Information and Communication Technology, IEEE* (2016), pp. 120–126
12. E. Tang, Y. Fan, Performance comparison between five NSQL databases, in *International Conference on Cloud Computing and Big Data, IEEE* (2016), pp. 105–109
13. B. Khalfi et al., A new methodology for storing consistent fuzzy geospatial data in big data environment, vol. 3, in *IEEE Transactions on Big Data* (2017)
14. M. Scavuzzo et al., Providing Big data applications with fault-tolerant data migration across heterogeneous NoSQL databases, in *International Workshop on BIG Data Software Engineering, ACM* (2016), pp. 26–32
15. Y.T. Liao et al., Data adapter for querying and transformation between SQL and NoSQL database. Future Gener. Comput. Sys. **65**, 111–121 (2016). Elsevier
16. J. Klein et al., Performance evaluation of NoSQL databases: a case study, in *International Workshop on Performance Analysis of Big Data Systems, ACM* (2015), pp. 5–10
17. A. Hossein, A. Rahnama, Distributed real-time sentiment analysis for big data social streams, in *International Conferences on IEEE* (2014), pp. 789–794
18. A. Bifet et al., Efficient online evaluation of big data stream classifiers, in *Conference on Knowledge Discovery and Data Mining, ACM* (2015), pp. 59–68
19. T. Chardonnens et al., Big data analytics on high velocity streams: a case study, in *International Conference on Big Data, IEEE* (2013), pp. 784–787
20. E. Olshannikova et al., Visualizing big data with augmented and virtual reality: challenges and research agenda. J. Big Data **2**(1), 22 (2015). Springer
21. B. Wu, B.M. Wilamowski, An algorithm for visualization of big data in a two-dimensional space, in *41st Annual conference of the IEEE Industrial Electronics Society, IEEE* (2015), pp. 53–58
22. Y. Xie et al., Visualization of big high dimensional data in a three dimensional space, in *3rd International Conference on Big Data Computing, Applications and Technologies, ACM* (2016), pp. 61–66

23. A. Kuma et al., Efficient and secure cloud storage for handling big data, in *International Conferences on IEEE* (2012), pp. 162–166
24. Q. Xia et al., Data locality-aware query evaluation for big data analytics in distributed clouds, in *Second International Conference on Advanced Cloud and Big Data, IEEE* (2014), pp. 1–8
25. I. Abaker et al., The rise of “big data” on cloud computing: review and open research issues. Inf. Syst. **47**, 98–115 (2015). Elsevier
26. G. Press, in *Top 10 Hot Big Data Technologies* [Online] (2013, Mar 14). Available: https://www.forbes.com/sites/gilpress/2016/03/14/top-10-hot-big-data-technologies/#1643589c65d7