Chapter 2
Big Data Analytics

Bhagya Nathali Silva, Muhammad Diyan and Kijun Han

Abstract During the last few decades, with the emergence of smart things and technological advancements of embedded devices, Big Data (BD) and Big Data Analytics (BDA) have been extensively popularized in both industrial and academic domains. The initial portion of the chapter aims to deliver a generic insight toward BD and BDA. In later sections, details that are more specific to BD and BDA are discussed. In fact, BD notion is characterized by its distinctive features such as large amounts of data, high-speed data generation, and wide variety among data type and sources. Consideration on these characteristics assists in determining potential data processing techniques. Hence, this chapter further elaborates on key BD analytical scenarios. Moreover, application of BD, BD analytical tools, and data types of BD are described, in order to enlighten the readers about this broad subject domain. Finally, the chapter concludes by identifying potential opportunities as well as challenges faced by BD and BDA.

List of Abbreviations

IoT	Internet of things
BD	Big Data
BDA	Big Data analytics
ML	Machine learning
DL	Deep learning
MR	MapReduce
TB	Terabyte
PB	Petabyte
EB	Exabyte
ZB	Zettabyte
YB	Yottabyte
RDBMS	Relational database management system
XML	Extensible Markup Language
WSN	Wireless sensor networks

M. Khan et al., *Deep Learning: Convergence to Big Data Analytics*, SpringerBriefs in Computer Science, https://doi.org/10.1007/978-981-13-3459-7_2

CPS	Cyber physical systems
MM	Multimedia
AI	Artificial intelligence
DM	Data mining
OLAP	Online analytical processing
BPM	Business performance management
NLP	Natural language processing
NER	Named-entity recognition
DBMS	Database management systems
URL	Uniform resource locator
NoSQL	Not only SQL
HDFS	Hadoop Distributed File System

2.1 Overview

The emergence of smart devices and connected networks has pioneered Internet of things (IoT) notion, boosting the data generation speed during past few decades (Khan et al. 2016, 2017; Silva et al. 2017b). As reported by International Data Corporation (IDC) in 2011, the annual data volume has increased approximately nine times within five years reaching up to 1.8ZB and further they estimated the data volume growth to be double in every other two years until 2020 (Gantz and Reinsel 2011).

Figure 2.1 concisely illustrates the evolution data growth during past few decades. Big Data (BD) notion was coined as a result of massive data deluge from a wide range of operational domains, i.e., Internet, sensor networks, managerial systems, finance systems, and user-generated data. Since initial divulgement, BD has been in the spotlight alluring both technical experts and public in general and hence been defined by many domain experts considering diversified aspects and perspectives (Silva et al. 2018b). In the beginning, BD was used as a term to define prodigious datasets. However, BD is solely not about the size or the amount of data. Doug Laney defined BD as a 3V model considering distinctive BD characteristics, namely volume, velocity, and variety (Khan et al. 2018; Laney 2001; Silva et al. 2017a). Sheer data size is indicated by volume, whereas velocity term is used to characterize expeditious data creation, and variety defines the diversity among different data sources and data types. Another widely accepted definition by IBM amended 3V model into a 4V model by introducing veracity as the fourth V of any BD model (Gandomi and Haider 2015). Veracity indicates the uncertainty of voluminous data (Khan et al. 2018).

Rapid growth of digital data has driven the research community to data-driven experiments, influencing all aspects of the social dynamics in the society (Silva et al. 2018a). Nevertheless, gathering enormous data amounts will be futile without proper knowledge discovery mechanisms. With rapid data growth, voluminous data analysis for knowledge discovery has become tedious and challenging for data engineering

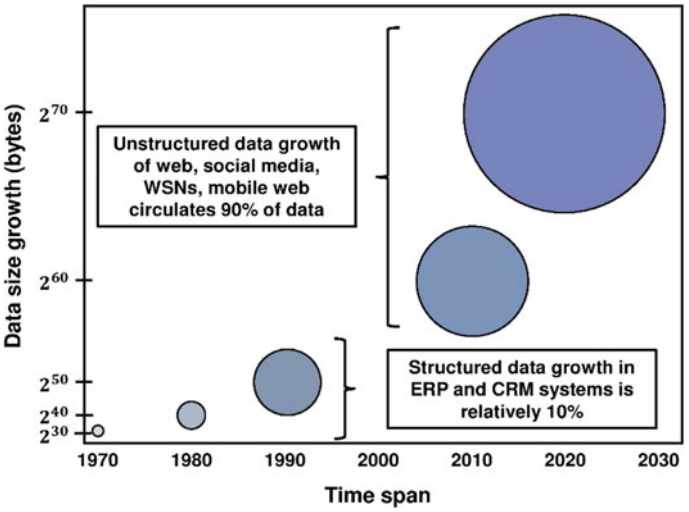

Fig. 2.1 Data growth evolution from 1970 to 2020 (predicted)

experts. Extreme variation among data sources and data types of BD has further exacerbated colossal data management tasks. Hence, BD notion was extended toward Big Data Analytics (BDA) that aims to discover valuable knowledge from large datasets enforcing intelligent decision-making. Therefore, organizations aim to possess efficient data processing mechanisms that transform high-velocity heterogeneous data into valuable information. As reported by Labrinidis and Jagadish (Labrinidis and Jagadish 2012), valuable data extraction from BD can be classified into five stages as shown in Fig. 2.2. These five stages can be bundled as data management and data analytics (Gandomi and Haider 2015). Data management involves data acquisition, storing, preparing, and retrieving for analytics. Data analytics refer to various methods, techniques, and algorithms that extract intelligence from BD. Thus, analytics can be identified as a subprocess of entire BD processing task.

Nevertheless, conventional data analyzing techniques and algorithms fail to process high-velocity data, thus creating a compelling demand to associate other related technologies capable of incorporating advance knowledge discovery mechanisms. Owing to the collaboration among multiple domains and enhanced computation facilities, knowledge discovery process has been broadened and strengthened. For example, domain experts of machine learning (ML) and data engineering have identified the potential of converging ML and BD to improve BDA performance. Contrasting to shallow learning techniques in conventional analytics, ML-converged BDA employs deep learning (DL) techniques to discover intelligence from BD.

As stated afore, BDA aims to explore hidden knowledge encapsulated in undiscovered patterns and correlations (Hu et al. 2014). Considering the demand of time, BDA is categorized into real-time analytics (streaming/online) and archived analytics (batch/off-line). Real-time analytics presumes data value is correlated with data

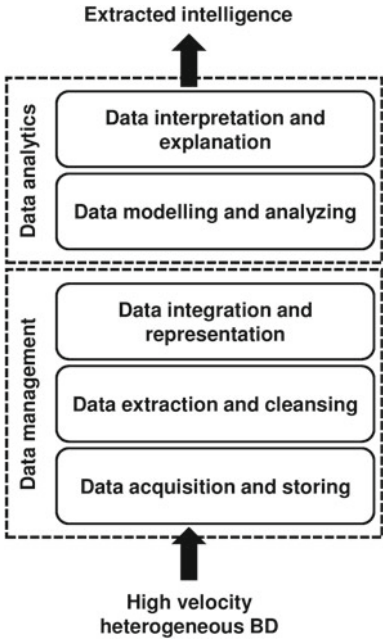

Fig. 2.2 Stages of big data processing to extract valuable knowledge from big data

freshness (Tatbul 2010). In streaming analytics, data arrives continuously as a stream at a high speed. Due to memory constraints, particularly small data portions from the stream are stored and examined to determine potential knowledge from the approximated patterns. Streaming analytics are widely used for real-time applications that require higher responsiveness with utmost accuracy. Spark, Storm, and Kafka are few dominant open-source systems, which support streaming analytics. Contrasting to online data analytics, batch processing analyzes data after storing. MapReduce (MR) is the most widely used batch processing method (Silva et al. 2017a). MR divides a large dataset into smaller portions, in order to perform parallel and distributed processing on each portion. Intermediate results obtained by small portion analysis are combined together to determine the end result.

Owing to potential applicability of discovered patterns and knowledge, many organizations and industries have welcomed and embedded BDA to the organizations' operational frameworks. Consequently, BDA problems are involved in fields varying across social administration, world economy, scientific research, etc. (Chen and Zhang 2014). On the one hand, as reported in McKinsey's report by Manyika et al. (2011), BDA has the competence to flourish global economy with its active participation in various fields. On the other hand, the McKinsey's report (Manyika et al. 2011) claims improvements in social administration can be achieved by introducing BDA functionalities like pattern identification and knowledge mapping to public services sector.

The rest of this chapter will provide descriptive information related to BD characteristics, BD analysis problems, data processing methods, opportunities and challenges, BD applications, data types of BD, and BD tools.

2.2 Characteristics of Big Data

Even though BD term initially coined around mid-1990s, the term became popular around year 2011 creating the hype about BD, owing to extensive efforts made by prominent technology organizations including IBM (Gandomi and Haider 2015). However, with wider acceptance, BD definitions have evolved rapidly inducing rather confusions. Despite of these definition variations, Laney's 3V model (Laney 2001) based on fundamental characteristics of BD was publicly accepted by industrial experts and academia. Since then, the three Vs model, namely volume, velocity, and variety, has been considered as core dimensions of BD management.

As the term implies, volume refers to the size of data. In 2012, IBM conducted a survey with 1144 participants to know about BD familiarity and found out a majority, which is over a half of the participants perceive BD as data with a size beyond one terabyte (TB) (1 TB $= 2^{40}$ Bytes). With expeditious data generation, TB era has evolved to petabyte (PB), exabyte (EB), zettabyte (ZB), and yottabyte (YT). As reported in (Beaver et al. 2010), about 16 million photographs are processed within a second at Facebook, which is approximately equal to 1 TB. According to estimated reports, Facebook stores over 260 billion photographs that roughly span over 20 PBs (1 PB $= 2^{50}$ Bytes). It is worthy to note that magnitude definitions of BD are proportional and correlate with corresponding attributes such as data type, time, and source. Hence, datasets presumed to be BD in the past might not be perceived as BD in the present with respect to current storage capacities and data acquisition methods. Similarly, BD in current era might not be able to surpass the magnitude threshold levels of BD in the future. The type of data is another concern when defining BD level thresholds. Even though data sizes can be equal for two datasets including a tabular dataset and a video dataset, the data type determines the number of records per each dataset. Hence, we can understand that defining a threshold level for data volume in order to interpret BD is impractical.

With the advancements in communication technologies, smart devices, and sensor networks, data acquisition process encountered with heterogeneous data types. Extreme heterogeneity of data in a dataset describes the variety of characteristics of BD. Consequent to technological advancements, data can be structured, semi-structured, or unstructured. Data in relational database management systems (RDBMS) and spreadsheets is identified as structured data as it complies with strict standards to support machine readability. However, only a small portion roughly about 10% belongs to structured data category. Data types with lack of structural organization, i.e., image, audio, video, and text, are categorized as unstructured data. Semi-structured data loosely spans across structured and unstructured types. Extensible Markup Language (XML) is a common semi-structured data form popular with

Web data exchange. Semi-structured data does not adhere to rigorous standards. For example, users define data tags of XML documents, in order to support machine readability. In modern world, many organizations stockpile unstructured data from sensor networks and social media. Even though data hoarding is not a relatively new concept, the novelty and beauty comes with new analytical technologies that improve business processes with stockpiled data. For example, retailers could determine customers' buying pattern and store traffic details exploiting face recognition techniques. This valuable information aids organizations to make lucrative decisions such as staff management, good placement, and appealing promotions.

Velocity implies the data generation rate, and corresponding data analyzing speed requires to make appropriate decisions. The extraordinary data generation speed consequent to the emergence of smart devices and sensor networks has drawn the attention toward real-time data analytics and data-driven decision-making. Data analytics would have to analyze thousands of streaming sources as the smart device usage escalates exponentially. Nevertheless, conventional data processing techniques fail to handle enormous data processing tasks with simultaneous data feeding. As a solution, BDA technologies start to act the role of conventional data processing techniques, with considerably very large datasets. It is noteworthy that BDA technologies are far more efficient than the conventional techniques as they discover real-time intelligence from colossal amount of raw data as batches as well as in streaming form.

In addition to these key attributes, few other characteristics of BD, i.e., veracity, variability, and value, were identified by prominent technology organizations. Veracity defined by IBM is the fourth V of 4V model. This implies uncertainty and unreliability factors of BD. Variability dimension was defined by SAS institute to indicate data flow rate variations that change from time to time with peaks and troughs. In addition to flow rate variations, this attribute encloses complexity factor arose with heterogeneous data sources. Oracle introduced value as another characteristic of BD. In general, BD value is low and not directly proportional to the volume. Nevertheless, extensive analysis on enormous data volumes is promising to improve data value.

2.3 Big Data Processing

As discussed previously, BD is less valuable in its original form. Value-added intelligence could be derived through proper data analytics. However, intelligence discovery from bulks of data is a tedious and challenging task and involves many other stages. This section divides the data processing task into three stages, namely data acquisition, preprocessing, and analyzing as shown in Fig. 2.3.

Data acquisition is the initial stage, which gathers raw data from various real-world objects, devices, and sensors. A well-designed data collection process ensures the validity of results subsequent to data processing procedures. It is worthy to note that data collection design should consider the characteristics of data generators as well as succeeding data analytic objectives. Sensors, log files, and Web crawlers are

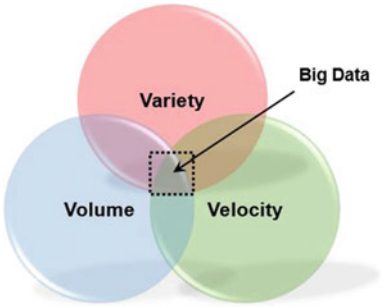

Fig. 2.3 Laney's 3V model based on fundamental characteristics of BD

some common data acquisition techniques (Hu et al. 2014). Sensors are commonly used to acquire physical measurements, e.g., temperature, humidity, and particle concentration in machine-readable digital format. Extended sensor networks can be either wired or wireless. However, during the past few decades, wireless sensor networks (WSN) have gained a significant popularity over wired sensor networks. A system consists a collection of sensors which is known as a cyber physical system (CPS) (Shi et al. 2011). Log files are another widely utilized data acquisition method that records activities in a predefined file format. For instance, Web server log files keep track on every activity performed by Web users. Except from Web servers, stock applications, financial applications, traffic management systems, and many other applications use log files for data acquisition. Web crawler is software that stores Web pages for search engines (Cho and Garcia-Molina 2002). Web crawlers perform data collection in Web-based applications. There are many other data acquisition methods other than afore-stated techniques. Descriptive explanation on different BD data types can be found in (Chen et al. 2014).

Once data is acquired, gathered data undergoes preprocessing, in order to address variabilities in noise, consistency, and redundancy resulted from heterogeneity of sources. The demand for data preprocessing has become significant with the rising cost on data transmission and storage facilities. In order to maintain the quality of stored raw data, preprocessing techniques were introduced to the BD systems. Integration, cleansing, and redundant data elimination are some common preprocessing techniques used in BD systems. Data integration combines data acquired from various sources to facilitate a collective view of data (Lenzerini 2002). Even though conventional data management incorporates data warehouse method and data federation method, "store-and-pull" characteristics of these two techniques hinder their applicability with high-velocity real-time applications. Therefore, utilizing integration techniques on streaming engines (Tatbul 2010) and search engines (Cafarella et al. 2009) would be beneficial. Data cleansing identifies uncertain, incomplete, or ambiguous data and modifies or removes them from the dataset, subsequently improving the quality of data. Considering the complexity of cleansing techniques, embedding them should be done with care to obtain data accuracy, without compromising the computational speed. Redundancy elimination techniques reduce the amount of repeated and superfluous data to improve data transmission, storage utiliza-

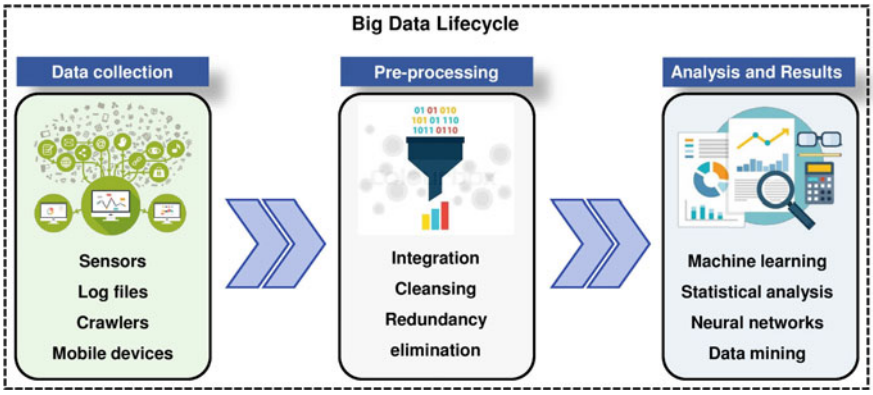

Fig. 2.4 Stages of big data life cycle with corresponding technologies

tion, reliability, and data consistency. Similar with all other approaches, redundancy elimination comes with benefits and trade-offs. Therefore, mindful usage can improve the data quality without straining the system on data compression and decompression procedures. In fact, none of the techniques assures optimal preprocessing over a wide range of large datasets. Hence, collective consideration of data characteristics, problems to be solved, and analytic objectives would be crucial for the selection of proper preprocessing mechanism (Fig. 2.4).

Data analysis is the most imperative stage of deriving intelligence from BD. The utmost goal of data analysis is to discover hidden knowledge that supports decision-making. In accordance with Hu et al., data analytics can be divided into six major research areas, i.e., structured data analytics, text analytics, Web analytics, network analytics, multimedia (MM) analytics, and mobile data analytics (Hu et al. 2014). Even though the objective of data analysis varies across different domains, some data analyzing approaches, i.e., data visualization, machine learning, neural networks, statistical analysis, pattern recognition, signal processing, and data mining, are beneficial and convenient to use in every domain (Khan et al. 2018). Data visualization methods deliver valuable information in the form of graphical illustrations. BD visualization has been studied by many interest groups, owing to the benefits in software designing and algorithm development. ML is a part of artificial intelligence (AI) that designs algorithms to learn computational behaviors with experimental data. Not only that ML enables automatic knowledge discovery and intelligent decision-making, the consolidation of ML with deep learning has pioneered deep machine learning (DML) as a novel research frontier. Statistical analysis is another common data processing method that collects, organizes, and interprets data. The statistical analysis approach identifies general relationships and valuable correlations among various data attributes. Nevertheless, lack of suitability of general statistics in BD analytics has paved the way for parallel statistical analysis and large-scale statistical algorithms. Data mining (DM) is highly favored in BD processing to extract hidden knowledge from very large datasets. In other words, DM can be expressed as

a consolidation of ML and statistics. BD mining is more challenging than conventional DM and required to be dealt with gigantic datasets. According to Wu et al. (2008), Apriori, K-means, Naïve Bayes, Cart, etc., were identified as promising DM algorithms for BD.

With the rising of BD era, experts were keen on extracting valuable data rapidly, using extensively large datasets. For that purpose, many organizations use hashing, indexing, bloom filtering, and parallel computing techniques. Hashing transforms data into index values or fixed-length numerical values to increase the speed of reading, writing, and querying, Indexing is used to reduce the cost of reading and writing by boosting the speed of insert, delete, update, and query processes. The beauty of indexing is that it can be applied on all forms of data including structured, semi-structures, and unstructured data. Bloom filter is another form of BD processing technique, which consists of a collection of hash functions along with data to facilitate lossy data compression. With increasing demands in data processing, many computational resources act together simultaneously in parallel computing. Contrasting to conventional serial computing, parallel computing segregates a computational task into several smaller tasks and assigns them to different computing resources to achieve co-processing.

2.4 Data Analysis Problems in Big Data

As stated in Sect. 1.3, BD analytical problems can be broadly categorized into six major areas, namely structured data analytics, text analytics, Web analytics, multimedia analytics, network analytics, and mobile data analytics.

Business organizations and scientific communities generate a colossal amount of structured data. Structured data analytics highly rely upon RDBMS, online analytical processing (OLAP), business performance management (BPM), and data warehousing. Deep learning has become an active player in structure BD analysis. DL-incorporated ML algorithms play a significant role in learning multilevel data representation models (Hinton 2007). Furthermore, mathematical model-based ML is occupied in sophisticated structured application domains such as energy management and fault detection (Baah et al. 2006; Moeng and Melhem 2010).

Text is a common form of data stored in documents, Web pages, emails, and social media. Considering the widespread nature, text analytics is considered to be vital than conventional structured data analytics. Text mining is another term that refers to text analytics, which discovers knowledge from unstructured textual data. Text analytical problems involve ML, statistics, DM, computation linguistics, and natural language processing (NLP) techniques. NLP techniques empower computers to comprehend text, analyze textual data, and generate textual information. Automatic knowledge extraction involves information extraction based on a specific topic or a theme. Named-entity recognition (NER) is widely utilized in such scenarios to determine data affiliated with predefined categories. Extractive summarization is another solution approach for text analytical problems, which summarizes a report

to few key sentences. Graph-based text mining is also used to categorize textual data according to a theme or a topic.

Consequent to the eruptive growth in number of Web pages, Web analytics came to the spotlight during last two decades. Knowledge discovery and intelligence extraction from Web resources is considered to be the utmost goal of Web analytics. Web analytics involve data retrieval, extraction, and data evaluation. Web analytics is further categorized into content mining, usage mining, and structure mining (Pal et al. 2002). Web content mining utilizes either database approach or information retrieval approach (Hu et al. 2014). Mining of secondary data generated by Web sessions belongs to Web usage mining. Usage mining considers user details, user registration, user queries, cookies, logs from server, browser, and proxy. Discovering graph links within a Web site or among different Web sites defines Web structure mining.

With the phenomenal popularity of social media, MM data grew rapidly along with ubiquitous accessibility. Consequently, multimedia analytics was emerged as a novel research area that aims to extract new knowledge from any form of multimedia data, i.e., audio, video, and image. MM analytics is challenging and tedious due to variability in multimedia data types. Nevertheless, MM data contains more information than structured data and textual data. Hence, experts in both academia and industry have proposed and studied many interest areas, i.e., MM annotation, retrieval, indexing, summarizing, that address key challenges of MM analytics. MM annotation assigns labels to MM files outlining the file contents. Indexing and retrieval techniques assist people to discover intending MM file easily and swiftly. MM summarization derives the most notable audio content or video content that best describes the original file.

Explosive growth in connected networks and social networks has extended conventional bibliometric network analytics to sociology network analytics (Watts 2004). Social networks consist of content data (text, image, video, etc.) and graph-like linkage data. In fact, content data and linkage data of social media are exceptionally rich from hidden knowledge. However, extreme complexity of these lucrative data has become a challenge for social network analytics. Social network analytics comprises of two major research areas, namely content data analytics and linkage data analytics. More elaborated explanations on content analytics and linkage analytics can be found in (Aggarwal and Wang 2011). As stated, social media consists of heterogeneous data types; hence, every BD analytic approach can be incorporated in social network analytics. Nevertheless, these approaches should be tailored to meet time constraints, noisiness of data, and dynamicity attributes of social networks.

Mobile analytics emerged with tremendous growth in mobile data, subsequent to advances in mobile computing. However, mobile data analytics encounter with challenges result from mobile data characteristics such as noisiness, data redundancy, location awareness, and activity awareness. In order to ensure user satisfaction, mobile analytics should facilitate intelligent decision-making on real-time basis. Owing to knowledge discovery served with mobile analytics, building of complex mobile systems that use empirical data has become more reliable and easy.

2.5 Applications of Big Data

BD applications assist small- and large-scale businesses to make intelligent decisions with the aid of voluminous data analysis (Hilbert 2016; Sagiroglu and Sinanc 2013). Internet clicks, Web server logs, activity reports, content of social media, mobile phone history records, email texts of customers, and sensor data are the main sources of data. Different interests of organizations enforce BD applications to disclose hidden information by exploring large volumes of data records. This section will cover BD applications in different domains (Clegg 2009).

2.5.1 Healthcare

Data generated in healthcare systems is not insignificant. Due to inadequate capacity and shortcomings in standardization, the healthcare industry covers BD. The usage of BDA in healthcare has revolutionized to make personalized medication easy and smart. For particular diseases, researchers are in race to provide suitable treatments in specific time intervals, analyze data pattern to recognize drug side effects and future treatments, and lower the treatment costs. Sophisticated views on health, i.e., eHealth, mHealth, and wearable devices, generate prodigious amount of data including images, records, sensor data, and patient data. Another potential application of BD in healthcare is to predict specific regions affected by some diseases by mapping health and geographical data. Consequent to accurate and timely predictions, doctors can act proactively to plan, diagnose, and provide vaccines and serums.

2.5.2 Manufacturing

A key potential benefit of applying BDA in production and manufacturing domains is to gain transparency and zero downtime. For this purpose, large amount of data is required together with latest analytical tools, in order to process data on real-time basis and to extract valuable knowledge from heaps of raw data. Following are the main areas in manufacturing and production domain, which utilize BD application for performance enhancements.

- Productivity and tolerance tracking
- Planning and Management in supply chain
- Energy efficiency
- Processes of testing and emulation for new products
- Customization for massive manufacturing

2.5.3 Government

The performance of government operations could be improved by the adaptation of BDA concepts. A particular set of data is shared among multiple applications as well as multiple governing departments to extract knowledge. Innovative applications play an important role in e-governance in multiple domains. Following are some of the areas of governance where BD and BDA play their roles for performance enhancements.

- Cyber security and intelligence
- Tax compliance
- Crime identification and its prevention measure
- Traffic management
- Weather
- Research and development

2.5.4 Internet of Things

IoT is another wide application domain of BD, which on the other hand is a source of BD creation with its enormous number of connected sensors, things, and devices. IoT gathers sensor and actuator data and then uses them in a variety of contexts.

2.6 Data Types of Big Data

A variety of data types exists in BDA, namely network data, linked data, event data, time series data, natural language data, real-time media data, structured data, and unstructured data. With the unceasing data growth, distinguishing valuable data has become a crucial demand in modern scientific era. Nevertheless, determining valuable data from ambiguous, noisy, and misinterpreted is still being a challenge to data scientists all over the world. This section provides concise descriptions about afore-stated data types.

1. Structured data: Numerical data stored in rows and columns, where every data element is defined, is known as structured data. About 10% of total data volume belongs to this type of data, which is accessible through database management systems (DBMS). Governmental departments, real estate and enterprises, organizations, and transactions generate structured data in each of their operational processes.
2. Unstructured data: Anything apart from tabular data with defined data elements belongs to unstructured data category. Data in the form of image, audio, video, and text is considered as unstructured data. This type of data contributes up to 90%

of the total data volume. In last few decades, popularity of social media has added more unstructured data, which cannot be stored and processed using traditional data storing and processing approaches. Such data can be stored in appropriate databases. NoSQL system is one of the suitable storage for unstructured data. CouchDB and MongoDB are widely used by many organizations as NoSQL repositories.

3. Geographic data: Data generated from geographic information systems (GISs), i.e., address, workplaces, buildings, transportation route, and road systems, belongs to geographic data. These data elements are easy to gather via deployed sensors and GIS. Consequent to data gathering, altering, processing, and analyzing tasks take place in order to derive knowledge from raw data. Moreover, geostatistics are occupied to monitor environmental conditions.
4. Real-time media: Streaming live media data or stored media data results in real-time media. In fact, storing and processing real-time media has been a challenge due to its streaming nature. There are number of sources that generate audio, video, and image data. To name a few, flicker, YouTube, and Vimeo are among the key real-time media generating sources. Videoconferencing is the other source of real-time media data, which facilitates seamless full-duplex communication.
5. Natural language data: People generate data in the form of verbal conversations. The level of editorial quality and level of abstraction are different of such type of data. Speech capturing devices, Internet of things, mobile phones, and fixed-line phones are the main sources of natural language data.
6. Time series: Measurements or observations on a particular indexed in time order are considered as time series. In general, time series data is observed at equal time intervals. Time series data is analyzed to extract hidden knowledge from gathered time data. Signal processing, statistics, earthquake prediction systems, pattern recognition, and many more other areas utilize time series data type.
7. Event data: Event data is generated with the correlation between external data and time series data. The utmost goal of event data type is to distinguish valuable events among innumerable amount of events taking place. The three components of an event data are action, time interval, and state. Combination of all three components creates an event data. Event data represents nested, renormalized, and schema-less characteristics.
8. Network data: Large networks such as twitter, YouTube, and Facebook generate network data. Other information networks, biological networks, and technological networks are also act as sources of networks data. Network data can be represented as one-to-one or one-to-many relationships among network nodes. A node can be a user, data item, internet device, neural cell, etc. Maintaining connection between nodes and network structures is considered to be the main challenge for network data.
9. Linked data: Linked data type includes URLs in Web technology. Computers, embedded devices, and other smart devices are then able to share and inquire semantic information. Data can be read and shared using linked data types.

2.7 Big Data Tools

Continuous development of business operations highly relies on thorough investigation of BD. Data analytics plays an important role in intelligent decision-making as well as development and well-being of the organization. However, processing sheer amounts of data with the aid of conventional data processing tools is not feasible and not efficient. Hence, various BD tools were introduced in the recent past. BD tools assist organizations as well as data scientists to derive knowledge-driven decisions efficiently and cost effectively. A variety of BDA tools are being used by the practitioners to facilitate data storage, data management, data cleansing, data mining, data visualizing, and data validation. This section briefly outlines widely used BDA tools.

1. NoSQL

In general, SQL is widely used to handle and query structured data. However, tremendous growth of unstructured data has pioneered the emergence of unstructured data analytical tools. Subsequently, not only SQL (NoSQL) has been developed to handle unstructured data types effectively. NoSQL databases do not particularly adhere to a schema when storing unstructured data. Hence, the column values of the table vary with respect to each data record (row). Due to this schema-less nature, NoSQL storages compromise consistency over speed, fault tolerance, and availability. Even though NoSQL has gained immense popularity during last few decades, challenges arise from low-level query languages and lack of standardized interfaces is yet to be addressed.

2. Cassandra

Apache Cassandra is a NoSQL, open-source, distributed database that is capable of managing very large datasets across multiple servers. The beauty of Cassandra is that it comes with no single point of failure assuring high availability under any circumstances. Hence, Cassandra is being highly favored by expert groups when scalability and availability features are crucial without compromising performance. Moreover, Cassandra facilitates data replicating across multiple clouds or data centers to ensure lower latency and fault tolerance.

3. Hadoop

Hadoop is a framework that consists of a collection of software libraries that incorporate various programming models to facilitate distributed processing of large datasets. Scalability is another benefit comes with Hadoop. Distributed storage facility offered with Hadoop is separately known as Hadoop Distributed File System (HDFS). The Hadoop does not rely on hardware to facilitate high availability. Instead, it incorporates software libraries to detect, identify, and handle points of failures at the application layer. The Hadoop framework consists of Hadoop Common (common libraries), HDFS (storage), Hadoop YARN (schedule computing resources), and Hadoop MapReduce (process large datasets).

4. Strom

Storm is a real-time open-source distributed computation system. Storm manages large amount of streaming data, which is similar to batch processing of Hadoop. The beauty of Storm is that it can be handled with any programming language. In order to ensure proper real-time analytics experience, Storm integrates the existing queueing mechanisms as well as database technologies. Streaming processing of Storm requires complex and arbitrary partitioning at each computation stage. Real-time analytics, unceasing computation, and online ML are some of the key services offered by Storm.

5. Spark

Spark is a general-purpose open-source distributed cluster-computing framework. Spark guarantees higher performance in both stream data processing and batch processing. Spark integrates SQL, GraphX, MLib, and Spark streaming components. Spark can access data from various data sources and run on various platforms such as Hadoop, Mesos, and Yarn. In comparison with Storm, Spark is considered cost effective, since same code set can be used for both real-time processing and batch processing. However, Storm has gained superiority in terms of latency with fewer restrictions.

6. Hive

Hive is a cross-platform data warehouse software operates on Hadoop. Hive enables data analysis and querying of data stored in multiple storages and file systems that are integrated to Hadoop. In order to query over distributed file systems, Hive provides SQL abstraction via HiveQL, which is a SQL-like querying language. Hence, applications do not have to implement queries using low-level Java APIs. HiveQL transparently converts queries into Spark, Apache Tez, and MapReduce. Moreover, Hive offers indexes to improve the query execution speed.

7. OpenRefine

OpenRefine is an open-source stand-alone application that was previously known as Google Refine. OpenRefine is widely used in present data-driven world to cleanse noisy raw data and to transform data from one form to another. OpenRefine is similar to typical RDBMS tables in a way as it stores rows of data under columns. However, it deviates from the conventional scenario, as no formulas are stored in cells under columns. The formulas are used in OpenRefine to transform data, and they are transformed once only.

2.8 Opportunities and Challenges

As in every other field, BD owns its benefits, advantages, and opportunities, which are followed by challenges. With prominent opportunities in one hand, on the hand BD analytics face vital challenges. This section briefly discusses the potential opportunities and identified challenges.

National Science Foundation (NSF) and National Institute of Health (NIH) of USA have recently confirmed that employing BD analytics in their decision-making process has proven the potential benefits for future endeavors (Chen and Zhang 2014). In fact, BD era has changed everyone's lifestyle with its impacts on social and economic development. Rise of BD analytics transformed people's behaviorism from reactive to proactive, subsequently enforcing precautions for possible events that might take place in future. Owing to the benefits gained through intelligent knowledge discovery, BD will become the playmaker for many organizations to attract skillful employees, while assisting in critical decision-making to compete with other competitors. With the emergence of BD analytics, Hadoop technologies have attained a remarkable popularity and success. Nevertheless, in future, these technologies might not be sufficient to manipulate high-velocity BD. Hence, focusing on sophisticated processing techniques and storage technologies such as distributed databases would create a significant breakthrough. Moreover, coexisting operation of BD analytics with other technologies such as IoT, mobile computing, and cloud computing innovates technological advancements to assist efficient data acquisition, data storage, data analysis, and data protection. Novel data representation methods are another realm of opportunities that goes together with the maturation of BD analytics, since illustration of analytical results is one of the most factors in intelligent decision-making. Extending BD research opportunities is likely to be heightened in future, as BD is considered as an extension of human brain rather than a brain substitution.

BD analytics face different challenges throughout the BD life cycle. To be specific, challenges in data collection, data storage, data processing, data analyzing, results representation may suppress the full capacity of BD analytics. Extreme complexity of BD has diminished the applicability and continuous advancements of analytical algorithms. Sluggish evolution of BD analytics id always accompanied by BD characteristics is denoted by 3V model. Large volume datasets generate countless number of output classes and free parameters that adversely affect the processing time and complexity. In order to manage these challenges, experts have extended BD analytics to make fusion with parallel computing that incorporates clusters of CPUs and GPUs. Complexity of BD analytics is highly influenced by the heterogeneity of data types. Hence, BD analytical techniques should strictly focus on mechanisms that alleviate challenges arose with extreme data variety. Integrating heterogeneous data types for machine learning processes is another key challenge encounter with variety characteristic of BD. BD analytics is further challenged by rapid data generation rates. High-velocity data requires real-time processing to serve decision-making. However, minimal works have been performed on online learning process with BD, thus requiring more developments. Experts foresee mini-batch processing supported by parallelism as promising technique to manage high-velocity BD analytics. Extreme non-stationary nature associated with high velocity is another hindrance factor for real-time BD analytics.

After the thorough literature review, we identified afore-stated opportunities and challenges attached with BD. Consequent to the unceasing technical growth, many of the experts asserted to be optimistic about the potential benefits of BD analytics. Accordingly, they claim that technological advancements would be able to overcome the obstacles arose with BD volume, velocity, and variety.

References

Aggarwal CC, Wang H (2011) Text mining in social networks. In: Social network data analytics. Springer, pp 353–378

Baah GK, Gray A, Harrold MJ (2006) On-line anomaly detection of deployed software: a statistical machine learning approach. In: Proceedings of the 3rd international workshop on Software quality assurance, 2006. ACM, pp 70–77

Beaver D, Kumar S, Li HC, Sobel J, Vajgel P (2010) Finding a needle in haystack: Facebook's photo storage. In: OSDI, 2010, pp 1–8

Cafarella MJ, Halevy A, Khoussainova N (2009) Data integration for the relational web. Proceed VLDB Endowment 2:1090–1101

Chen CP, Zhang C-Y (2014) Data-intensive applications, challenges, techniques and technologies: a survey on Big Data. Inf Sci 275:314–347

Chen M, Mao S, Liu Y (2014) Big data: a survey. Mob Netw Appl 19:171–209

Cho J, Garcia-Molina H (2002) Parallel crawlers. In: Proceedings of the 11th international conference on World Wide Web, 2002. ACM, pp 124–135

Clegg B (2009) Before the Big Bang: the prehistory of our universe. St. Martin's Press

Gandomi A, Haider M (2015) Beyond the hype: Big data concepts, methods, and analytics. Int J Inf Manage 35:137–144

Gantz J, Reinsel D (2011) Extracting value from chaos. IDC iView 1142:1–12

Hilbert M (2016) Big data for development: a review of promises and challenges. Dev Policy Rev 34:135–174

Hinton GE (2007) Learning multiple layers of representation. Trends Cogn Sci 11:428–434

Hu H, Wen Y, Chua T-S, Li X (2014) Toward scalable systems for big data analytics: a technology tutorial. IEEE Access 2:652–687

Khan M, Silva BN, Han K (2016) Internet of things based energy aware smart home control system. IEEE Access 4:7556–7566

Khan M, Silva BN, Jung C, Han K (2017) A context-aware smart home control system based on ZigBee sensor network. KSII Trans Internet Inf Syst 11:1057–1069

Khan M, Silva BN, Han K (2018) Efficiently processing big data in real-time employing deep learning algorithms. In: Deep learning innovations and their convergence with big data. IGI Global, pp 61–78

Labrinidis A, Jagadish HV (2012) Challenges and opportunities with big data. Proceed VLDB Endowment 5:2032–2033

Laney D (2001) 3D data management: controlling data volume, velocity and variety. META Group Res Note 6:1

Lenzerini M (2002) Data integration: a theoretical perspective. In: Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on principles of database systems, 2002. ACM, pp 233–246

Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Byers AH (2011) Big data: the next frontier for innovation, competition, and productivity

Moeng M, Melhem R (2010) Applying statistical machine learning to multicore voltage and frequency scaling. In: Proceedings of the 7th ACM international conference on computing frontiers, 2010. ACM, pp 277–286

Pal SK, Talwar V, Mitra P (2002) Web mining in soft computing framework: relevance, state of the art and future directions. IEEE Trans Neural Netw 13:1163–1177
Sagiroglu S, Sinanc D (2013) Big data: a review. In: 2013 international conference on collaboration technologies and systems (CTS). IEEE, pp 42–47
Shi J, Wan J, Yan H, Suo H (2011) A survey of cyber-physical systems. In: 2011 international conference on wireless communications and signal processing (WCSP). IEEE, pp 1–6
Silva BN, Khan M, Han K (2017a) Integration of Big Data analytics embedded smart city architecture with RESTful web of things for efficient service provision and energy management. Fut Gener Comput Syst
Silva BN, Khan M, Han K (2017b) Internet of things: a comprehensive review of enabling technologies, architecture, and challenges. IETE Tech Rev 1–16
Silva BN, Khan M, Han K (2018a) Towards sustainable smart cities: a review of trends, architectures, components, and open challenges in smart cities. Sustain Cities Soc 38:697–713
Silva BN et al (2018b) Urban planning and smart city decision management empowered by real-time data processing using big data analytics. Sensors 18. https://doi.org/10.3390/s18092994
Tatbul N (2010) Streaming data integration: challenges and opportunities
Watts DJ (2004) Six degrees: the science of a connected age. WW Norton & Company
Wu X et al (2008) Top 10 algorithms in data mining. Knowl Inf Syst 14:1–37