Chapter 2
Big Data Analytics: The Underlying Technologies Used by Organizations for Value Generation

Bhavna Arora

Abstract The expansion of Internet and its applications globally has witnessed generation of high volume of data resulting in high volume of information. In the contemporary era of digital world, data is seen as the driving force behind the progression of business enterprises. Today, the data that is generated worldwide has grown ranging from terabytes to exabytes and petabytes, and the compounded rate of data further growing is much fast. The data generated widely has many forms and structures. The deluge of data generated, which is both valuable and challenging, along with emerging technologies and techniques that are used to handle it is referred to as the evolution and era of "Big Data". As the big data is generated from multitudinous sources, majority of this data exists in unstructured form that demands specialized processing and storage capabilities, unlike the structured data that uses storage and processing of traditional relational structures. This results in high complexity and uncertainty in data. The usage of statistical analysis, computer-based models and quantitative methods that can help the business organizations to improve insights for better operations and decision-making is referred as business analytics. To work intelligently and focus on value generation, organizations need to focus on business analytics. The analytics are a critical component of big data computing. As defined in the literature, an intelligent enterprise has the characteristics similar to human nervous system and is responsive to external stimuli. To leverage the large volume of data for driving the business enterprises, timely and accurate insights derived out of the big data are a big challenge. The technologies like Hadoop and Apache Spark assist in handling big data on both fronts. However, handling and analysis of big data are a challenge for any organization with respect to its storage and technical expertise. Business analytics is used in business organizations for value generation by data manipulation along with business intelligence and report generation. Advanced analytics are also used by business enterprises that use techniques of data mining, data optimization and predictive forecasting.

B. Arora (✉)
Department of Computer Science & IT, Central University of Jammu, Jammu, J&K, India
e-mail: bhavna.aroramakin@gmail.com

H. Chahal et al. (eds.), *Understanding the Role of Business Analytics*,
https://doi.org/10.1007/978-981-13-1334-9_2

Keywords Big Data · Data Analytics · Hadoop · V's of Big Data
Apache Spark

2.1 Introduction to Big Data

The contemporary era has witnessed very large volumes of data and the terminology and trends that have been accepted globally with these are "Big Data". The author in paper ("What Is Big Data?—Gartner IT Glossary—Big Data", n.d.) has defined big data as "Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision-making".

In Manyika et al. (2011), the author refers "Big Data" to "data set whose size is beyond the ability of typical database software tools to capture, store, manage and analyse".

The volume of data that is being stored today around the world is exploding. In the year 2000, the world witnessed storage of 8 lac petabytes of data. With the expansion of Web and its applications, the data that is being stored is growing exponentially. The data is likely to rise to 35 zettabytes by the year 2020. The data that is created is not analysed efficiently, and the insights of the data is not revealed. The data contains hidden insights that the companies can use to enhance their business perspectives. How the volume of big data impacts the human mind is very challenging. Considering the data volume that consists of multiples of terabytes may be considered as big data, but actually when it can be managed in network attached storages (NAS) or storage area network (SAN) using additional disc arrays, then it might not be considered as really Big Data. When the data exceeds this limit, i.e. about petabytes in size and can only be managed with sophisticated applications and tools, the data can be referred as "Big Data". This would require a complex distributed computing and storage grids extensively, so that this data could be managed.

However, companies used various tools and technologies to collect and store different types of big data. The analysis of these diversities of big data is challenging as the tools that are required for big data analysis are extremely complex to design and implement. The management of this data is another big challenge as the companies should have the clarity on big data adoptions. It is paramount in agreeing that such information in big data which is huge and complex has created various challenges for organizations that did not exist earlier. The large volumes of data available pose several problems for researchers, analysts and decision-makers in the industry. At times, the decision-makers in the organization tend to make their decisions without having complete facts, and others find the business intelligence along with the data analytics to be part of their visionary plans so as to enhance business competitiveness.

The evolution of big data has witnessed the explosive growth in the entire world's data that can be used to make decisions, but this can only be useful if this

can be made in timely manner. For this, powerful tools are needed that can assist in storage, extraction and analysing the data from the big data sets. The big data can also be defined as that data that cannot be processed through conventional methods of processing. To mention few varieties and sources of data that come under the big data realm are as follows (Jain 2015):

1. Black box data captures the voice and recordings of flight crew members of helicopter and other aircraft along with the information pertaining to the performance of the aircraft.
2. Data of stock holdings where the decisions made by a customer on a share or equity of different companies.
3. Data from social media websites such as Facebook and Twitter that holds views and other information posted by millions of users across the globe.
4. Transport and meteorological data sources.
5. Data retrieved by search engines from different databases.
6. Metamorphic and Census data.
7. Connection-oriented data that includes sensory data.
8. Data from cloud storages that provides computing and data on demand.

Big data is more than just more information; it represents the beginning of the end of the industry experience as a core competitive advantage (Stubbs 2014). Big data is not a philosophical fancy anymore. It is already in place in industry. Big data cannot be argued as just the latest version of "data". Today, the users are generating much more data and more types of data than before. In their work, Manyika et al. (2011) have proposed five major contributions that big data contributes to business organizations:

- transparency creation by making big data more accessible and ready to use in timely manner for value generation
- performance improvement by enabling experimentation
- population segmentation by tailoring products and services that meet specific needs
- decision-making support
- innovative business models, products and services

Big data and business analytics work hand in glove. Without data, analysis cannot be done. Without business analytics, big data is just noise. Big data bears the potential of making things efficient and is capable of generating returns. These returns include benefits to internal value such as productivity or external value like revenue generation. It offers exceptional insights along with predictive capabilities for those who are able to leverage it.

2.2 The V's of the Big Data

The contemporaneous era is witnessing production of data at astronomical rates. To analyse this data that is constantly varying, new tools and technologies are continuously being developed by experts that will be able to handle the complexities of large volumes of data. The future trend is that the big data is going to grow more rather than decrease, as more and more data generating applications are growing. Big data can be characterized by volume, value, variety, velocity (Philip Chen and Zhang 2014), veracity and variability, and each of these parameters can be defined as under:

2.2.1 Volume

Today, the large volume of data is generated because nowadays organizations collect and process data from a diverse range of sources such as application generated logs, machine-generated data, email data, weather and geographic information systems (GIS) data, survey data, reports, social media data. Big data analytics have the capability to compute gigantic volume of information. Data volumes have reached levels to terabytes (TB) or petabytes (PB). As an example, the financial industry produces voluminous data in terms of market data, quotes and financial trading. The New York Stock Exchange creates about more than one terabyte of data per day ("Want to make big bucks in stock market? Use Big Data Analytics", n.d.) and if this volume is calculated over the month, year and so on, the volume of data is immense. About 10 billion photographs (Beaver 2008) were hosted by Facebook creating about one petabyte of data storage in the year 2008. Another site Ancestry.com, stores around four petabytes of data (Bertolucci 2013). Even the Internet archive stores about two petabytes of data, and it is accelerated at a rate of about twenty terabytes per month ("1. Meet Hadoop—Hadoop: The Definitive Guide, 3rd Edition [Book]", n.d.). These big data structures comprising of high volumes of data tends to impose limitation on the storage and processing capabilities. It also imposes limitations to the database structures, and hence, the database modelling gets complicated as the data grows. In his work (Brock and Khan 2017), the author analyses that the huge amount of data poses challenges to underlying storage infrastructure which in turns calls for systems with scalability and capability for distributed querying. High computational power and parallel processing are required for analysing big data as the traditional database techniques are not able to cope with big data, as the size of data sets has surpassed the capabilities of computation and storage.

2.2.2 Variety

The traditional systems heavily rely on underlying structured data whose dimensions are considered as accuracy, completeness, relevance and timeliness. The inputs to such systems need to be entered judiciously and meticulously so that the output that is produced in the form of reports is meaningful and useful. The big data is heterogeneous. In such environments, the system needs to use techniques for data cleaning so as to eliminate the garbage data in source. Data collected from various sources like applications, stock data, emails, geographical data, weather data, social media application data is difficult to handle as it comes from a variety of sources, and it is virtually impossible to convert this heterogeneous data to a conventional structured form for processing. In order to process such data, special techniques and technologies are used that can understand and go beyond the traditional processing of the relational structured data. Big data solutions need different types of processing tools to process heterogeneous data.

2.2.3 Velocity

The rate at which the data flows in the system and its environment is termed as velocity. With the Internet and mobile data coming in the lives of the consumer, the era witnesses high rate of data flow as the consumers carry with their devices, a huge volume of streaming source of data that consist of geo-located images and audios. Studies reveal that in the year 2013, about five exabyte data were generated in the world every 10 min. Today, this figure has risen exponentially risen, and the data is being generated every minute. However, the importance of velocity of big data follows the similar rate of increase as in the case of volume of the data. For example, the business Walmart creates about 2.5 petabytes per hour (Brock and Khan 2017). One of the main challenges to the velocity is the communication networks. Since the big data processing demands real-time processing, the processing capabilities for inflow of the data streams in the networks are also a big challenge.

2.2.4 Veracity

The reliability and trustworthiness of data are termed as veracity of data. It also refers to the quality of the data. The point to focus on is "how accurate is all this data?" As an example, consider the tweets in Twitter posts. These posts contain hashtags, typos, abbreviations, etc. The data that is to be considered should be reliable, accurate and trustworthy. Manipulation and analysis of such data need to be qualitative and trustworthy to get correct insights from it. For real-time applications, to provide correct and reliable data at times the applications may produce nearest best results in the cases where the data that is being analysed in real-time applications fails to deliver in a particular moment.

2.2.5 Value

Value refers to the worthiness of the data being extracted. On one hand, if the organization has voluminous amount of data but unless it can be made useful for the organization, it is worthless. Even though there is an explicit association between data and insights, it does not definitely mean that there is value in big data. The original data received might have low value as compared to its volume. By analysing a large volume of data appropriately, high value and pronounced insights from the data can be obtained. The significance of embarking on initiatives of big data is to understand the costs of analysing and reaping the benefits during the process of collecting and analysing data. Thus, it ensures that the data that is reaped is monetized and the organization is benefitted to the maximum.

2.2.6 Variability

The variation in the flow rates of data is referred to as variability of data. The velocity of big data is inconsistent and has periodic troughs and peaks. As the big data is generated from myriad resources, complexity also has to be analysed. Complexity arises after collecting data from different sources as it has to connect, clean, match and transform the data (Fig. 2.1).

2.3 Big Data Classification

As the data sources of big data are numerous, based on the types of data, big data can be classified broadly into three categories—unstructured, semi-structured and structured.

2.3.1 Structured Data

The data that is stored in the databases in an orderly manner is referred to as the structured data. Statistics reveal that structured data only constitutes about

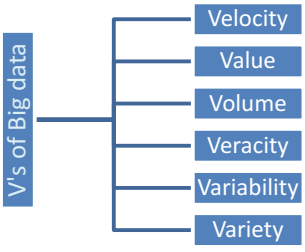

Fig. 2.1 6V's of Big Data

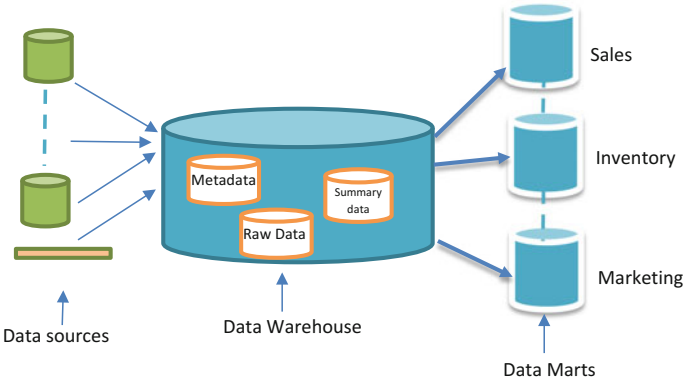

Fig. 2.2 Overview of data warehouse and data marts

one-fourth of the data that the organizations use. This data can be used in programming and querying structures from databases. The sources of such data can be from machine input or humans. Data that is generated by machines includes data from GPS-like devices or sensory devices like the medical equipment, Web interfaces and logs. Data that is included by human systems includes database records which involve human intervention in generating data. The two most popular approaches that can be used to manage large data sets in a structured way are data warehouses and its subsets, i.e. data marts. A data warehouse can be seen as an assemblage of data which is isolated from the operational systems and the decision-making process in any organization. It is a huge repository of historic data. The data is compiled and assembled from various resources so as to provide timely and accurate information. The data in a data warehouse is the extracted information from various functional units of an organization. Before the data is integrated into the warehouse, it undergoes a series of processes. These preprocesses are data cleaning, data transformation and data cataloguing. After the preprocessing, the data is available for higher-level online data mining functions. These warehouses are controlled by a centralized unit. The subset of data warehouse is a data mart. The data marts focus on specific functional area only. The warehouse and data mart both primarily vary in their scope and usage area. The structured data forms only a small subset of the Big data that is ready for analysis (Fig. 2.2).

2.3.2 Unstructured Data

Unlike the traditional row–column database structure, the unstructured data has no clear formats in storage. Such data constitutes about 80% of the total data that is included in big data. Till few years back, such data has been stored and analysed manually as it was quite difficult to analyse this data. The unstructured data can

comprise the machine-generated or the human-generated data. Typical examples of unstructured machine-generated data include satellite images, data captured from radars, whereas the unstructured human-generated data is spread across the entire globe which includes data from the Web content, social media and mobile data. Users tend to upload data on Facebook, Instagram, audios and videos, etc.; they all are a part of the massive unstructured data contributed by the user. The largest unstructured data component is the video data that constitutes big data.

2.3.3 Semi-structured Data

A very thin line exists between the unstructured and semi-structured data. Unless clearly defined, the semi-structured data can appear as unstructured. Even though that the information is not typically arranged in traditional database structured formats, but in order to process this data, some properties that make the data processing easier and convenient to process are contained in it.

2.4 Data at Rest and Data in Motion

Technically speaking, in order to gain insights from big data, right technology for various processes like collecting, managing and analysing is required. The predictive purpose for the same is quite critical. Since the data is from a variety of sources and is of different types, there is a requirement of different computing platforms that support in providing meaningful insights. In order to determine the technology and processes that are required to glean the insights from the big data, it is imperative to understand the difference between data at rest and data in motion.

2.4.1 Data at Rest

In data handling systems, data at rest refers to data that is stored in stable systems. The data at rest comprises data compiled and stored in structures like spreadsheets, databases, warehouses, archives and backups, mobile data. There can be different points where data is analysed and where its action is taken. This can occur at two separate times. For example, present month's business activities can be predicted based on last month's sales data by a retailer. The action is making strategic decisions, and the activity is the sales of the previous month. Market campaigns and strategies can be planned accordingly based on the variables like customer behaviour, sale schemes. Looking at this data analysis, the business can take advantage and such decisions can impact the sales in the stores, while the customer would be benefitted with the sale schemes the store offers.

2.4.2 Data in Motion

There is a difference in the analytics for data in motion, even though the process of data collection is similar to that of data at rest. Unlike data at rest, analytics can occur at the same time when the event occurs, i.e. in real time. The data in motion refers to data that is moving from one place to another. In such cases, many different networks can be used, e.g. sending email on Internet. Many nodes are connected to same network, and the transfer of the email has to go through multiple nodes in a network. Security issues arise for data in motion, and the data needs to be protected. Another example would be locating clients and their choices at various outlets of a water park. Latency also becomes a key concern as the lag in processing may affect the business results by missing an opportunity. This type of data is also referred to as data in transit or data in flight. Data at rest and data in motion can provide quiet meaningful insights for business analysis. It is important that appropriate processing methods and infrastructure may be used and deployed in order to obtain the perfect analysis of data.

2.5 Data Analytics

The process and techniques used for examining the data with the aim and purpose for inferring and to draw conclusions about that information are data analytics. It finds its usage in business organizations to help them make better decisions. Data analytics can be distinguished from data mining with respect to the purpose, scope and focus of the analysis. In order to mine data from the huge set of data that is in question, sophisticated software is used that rely on algorithms and are capable of working on large data sets. They can uncover undiscovered patterns and hence are capable of establishing the hidden relationships. This process is referred to as data mining. Appropriate analysis of big data can help a company to achieve cost reductions and dramatic growth. So the business houses should not wait too long to exploit the potential of analytics. Big data analytics focus on inferences, i.e. the deriving to conclusion based on known facts. The analysis can be categorized as under ("Data mining versus data analysis and analytics—Fraud and fraud detection —Academic library—free online college e textbooks", n.d.):

- Exploratory data analysis (EDA)—It is the preliminary stage where the data is explored and new features are discovered.
- Confirmatory data analysis (CDA)—Existing hypotheses are proven true or false.
- Qualitative data analysis (QDA)—It is analysis of the quality parameters of data to draw conclusions from non-quantitative and non-numerical data like pictures, audios, words or text, videos.

Big data analytics finds its significance in the cases of audits when the information systems of business organizations' along with other operations, procedures and processes are under reference. Data analysis also helps to determine if the systems under reference can effectively protect data while operating efficiently and also helps the organization accomplish overall goals. Business intelligence defines analytics from various perspectives. In call centre applications, it can be defined from online analytical processing (OLAP) to customer relationship management analytics (CRM). CRM analytics includes all process that analyses data about customers and presents it to facilitate and streamline for better business decisions of the organization.

2.5.1 Types of Business Analytics

The key sub-processes defined in big data are data management and data analytics. The process of data management looks after the acquiring, storing, retrieval and preparation for data analytics. The underlying technologies for data analytics are acquisition, annotation, aggregation, etc. (Saravanakumar and Nandini 2017). For analysing various types of structured, unstructured and semi-structured data, the following analytics are used (Saravanakumar and Nandini 2017):

- **Text Analytics**: It is also known as text mining from the textual data. It refers to the extraction of high-quality information from textual data by using statistical patterns. It includes machine learning and statistical analysis of text data using techniques such as information extraction (IE), summarization text, question answering and sentiment analysis. Tools for text analytics are SAS text analytics, IBM text analytics, SAP text analytics, etc.
- **Audio Analytics**: To analyse the unstructured audio data, speech or audio analytics is used. In audio analytics, information is extracted from natural language, i.e. languages spoken by humans. The most popular application for audio analytics is the call centres which have data for million hours and can be used to improve the customer experience and to enhance the business turnover. For audio analytics, two approaches are used—transcript-based approach and phonetic-based approach. The tools that are used for audio analytics are Marsyas, Vamp, SoundRuler and WaveSurfer, etc.
- **Video Analytics**: To monitor, detect and analyse data from video streams is referred to video analytics. This includes determining meaningful data from temporal and spatial events. The key applications where video analytics help are retail stores, health centres, transportation, securities, etc. Video analytics is also called video content analysis. This technology uses CCTV and surveillance cameras for detecting breaches, recognizing suspicious activities, etc. The tools used are Ooyala, Vidyard, Vimeo Analytics, etc.
- **Social Media Analytics**: The social media data consists of information that is gathered from websites such as Facebook, Twitter and blogs. The data needs to

be analysed by business houses for decision-making by studying behaviours and pattern of the user. The user opinions are extracted and analysed. The analytics can be content based and structured based for social media analytics. Tools used are ViralWoot, Collecto, SumAll, Tailwind, Beevolve, etc.

- **Predictive Analytics**: Historical and present data is analysed, and based on this, data prediction is made. It expresses reliability that what might happen in the future. This method is based on statistical methods. The various tools to perform predictive analytics are splunk, medalogux, etc.

2.6 Big Data Paradigm in Business Organizations

Since past few years, various business enterprises and other organizations are storing a large amount of data in large databases in data warehouses and data marts. However, data was analysed with data-mining algorithms to extract insights. Nowadays, the data stored is no longer homogenous in nature, but on contrary, it is a compilation from a variety of sources. The data in the traditional systems was organized and structured in rows and columns as it was largely generated from transactions. On the contrary, nowadays, the stored data is unstructured and generated from a variety of sources like audio–videos, photographs, text messages, maps generated from GPS devices, data from emails, social media sites, etc. All these data when stored in digital media is unstructured as there cannot be a common structure that can be defined for such data. Another key characteristic of such data is its real-time accessibility. Data can be retrieved about activities and events in real time and will also influence its outcomes. It is only possible if we have an organization that is designed to operate in real time. The process design should able to analyse and use real-time data. It should be able to produce instant insights and process those insights to support real-time decisions. The "real-time" factor affects the organizations to take timely and appropriate action. Summarizing, in order to gain maximum benefit out of big data, the organizations must work in real time.

2.6.1 Business Analytics: The Organizational Transformation

Business analytics can be defined as the use of the data-driven insights to generate value in real time. It is done by understanding the business relevancy, organizational insights, performance and value measurements (Stubbs 2014). The data-driven insights include data manipulation, reporting and business intelligence and advanced analytics. The advance analytics is that form of analytics that help provide answers to questions like what happened, what will happen, why it happened and what best possible one could do (Stubbs 2014). The advanced analytics include data-driven insights that include data mining, optimization and forecasting.

It can also be defined as the deep analysis of data or content by using appropriate technologies, tools and different techniques that are typically beyond those that are used with the traditional systems. The business intelligence (BI) may be used to uncover patterns that assist in discovering deeper insights, along with generating recommendations and making predictions. The techniques of advanced analytic may include text and data mining, machine learning and pattern matching, visualization, semantic and sentiment analysis, forecasting, network and cluster analysis, multivariate statistics, graph analysis, simulation, complex event processing, neural networks ("Business Intelligence—BI—Gartner IT Glossary", n.d.).

The output of business analytics is seen as value generation in an organization. This could be internal or external (Stubbs 2014). Internal value is from the perspective of teams that are within an organization. The outside or external value is seen from outside the organization. The organization needs to create these values through its key resources, i.e. people, processes, data and the technologies. A series of activities that can be linked to achieve an outcome is defined as a process. The processes can be strongly or weakly defined. A series of specific steps that is repeatable and may be automated is strongly defined process. On the contrary, an undefined process that relies on the capability of the personnel for execution of the process to complete it successfully is a weakly defined process. To generate new assets, various tools and technologies are applied and are consolidated to a common analytical platform. The key to business analytics is facilitating change, not driving towards better outcomes. There is a major paradigm shift in the way organizations execute their operational, tactical and strategic objectives as an outcome of business analytics.

2.6.2 The Intelligent Enterprise

Irrespective of how the people act and react to situations in the organizations, most of the organizations can be seen united under a common objective. A truly intelligent enterprise operates like our nervous system (Stubbs 2014) and possesses properties of agility, adaptability, flexibility and is appropriately responsive to external stimuli. There are different levels that are described as progressive for any organization. These are the approaches usually opted by organizations for building capability. The first level is the unstructured mode; the second level is the structured mode. The real process and mode start at level three, and from here, the system starts its "best practices" from the theory of "things working". When the gap between these two is closed or reduced, the business system becomes an intelligence enterprise.

Level 1: The Unstructured Mode In this system, everyone is working hard without a plan or clarity of work. Quality is hard to measure at this level as whatever is archived is not because of design and planning but just because of the efforts of motivated individuals. Technologically, the analysts use tools that are

basic desktop-centric and devote considerable amount of time to try to manage, source and exchange data within the tools that are semi-compatible. The data fragmentation is done at this level, and each of the team that works in this environment creates their own data repository. They tend to restart from beginning every time they work on a new project. However, difficulty may arise as the processes are manual, undefined, and may require substantial efforts to execute.

Level 2: Structured System Structured systems are the next level of the unstructured system. It is when the system follows higher-order patterns, but the system behaves randomly around a broader pattern. The organizations at this state try to balance local choice while considering global requirements. Constraints are set, functional and divisional strategies are established, and the entire organization tries to comply with them. Unconscious ignorance is one of the key barriers to success at this level. Data at this level exists in tabular format on networks that allow sharing of the data as well. Simple and common tools for desktop processing are used. Processes are weakly defined, and the skills are used of the employees across various processes.

Level 3–5: Towards an Intelligence Enterprise The level next to the structured system is the intelligent enterprise. This is the time when the organizations recognize the business analytics as integrated soul of the organization. The level at which the intelligent enterprises reach this point is totally process-centric. The three levels at which the intelligent enterprises work are at the team, department and enterprise levels. Understanding that business analytics is a journey and must be incorporated in all its functions and processes which is the key to success for the organization.

2.6.3 Intelligent Analysis

The pyramid hierarchy that exists in organizations suffers a number of challenges. The top management, i.e. where the strategic decisions are to be made, is far off than the actual scene where the actions are done. This time lag leads to delay in the decisions. The strategic decisions have a significant impact on various other levels of organizations. The decisions range from resource allocation to affecting the impact of the organization's competitiveness in the marketplace. Then is the level of tactical management, where the mid-level managers operate and affect the key operations like marketing, accounting, production. The focus is not on the entire organization and has lower resource implications as compared to the level above. The transaction processing system that is used by the operational management handles the structured data. It provides operational-level support. Figure 2.3 shows different levels of management and their data usage in business organizations.

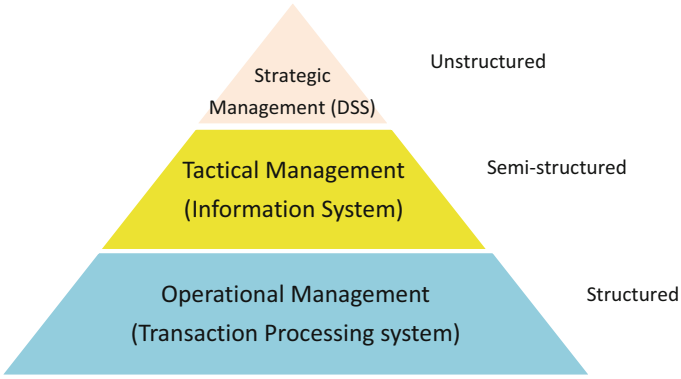

Fig. 2.3 Management system pyramid (Management systems with type of data handled)

2.7 Technologies for Data Analytics

Analytics refer to the discovery of meaningful patterns of data, e.g. data related to sales, transaction, revenue. Most information system deploy traditional database tools for relational databases such as structured query language (SQL). Business houses need data experts that have broader and deeper analytical skills that can provide support to challenges like data management, real-time analysis, real-time predictive analysis, data management issues including security and privacy (Miller 2014). The data engineer responsible for the data extraction and analysis has thorough and clear understanding of traditional relational databases along with non-traditional and NoSQL databases like Hadoop. These engineers are capable of integrating data from variety of data sources and are able to design data-driven services. They work in coordination with the scientist that works on handling of data.

In order to cope with the trends of big data, a variety of tools, techniques, methods, and technologies have been developed in recent years. When data derives in huge magnitudes, then the companies cannot rely on the in-house storage and processing anymore, as the "traditional" technology that focuses only around the central databases is no longer appropriate to handle it.

To determine what is needed and what fits in well, the requirements for big data processing need to be reviewed (Vossen 2014). These requirements can be characterized as follows:

- High processing capabilities
- High storage capabilities
- Scalability and support for distributed processing
- Fault-tolerant processing capabilities
- Support for parallel programming and processing paradigms
- Appropriate platform and execution environments.

2.7.1 *Hadoop—The Underlying Technology for Big Data Analytics*

The Apache Hadoop is Java-based software platform that supports data-intensive distributed applications (Philip Chen and Zhang 2014). It has been designed to avoid the low performance and the complexity encountered when processing and analysing Big data using traditional technologies (Oussous et al. 2017). The Hadoop platform is used for distributing computing and spreads the data and its processing across a number of servers. The paradigm that is used by Hadoop is the MapReduce (Fig. 2.4).

The kernel, Hadoop distributed file system (HDFS) MapReduce along with add-on projects that include Apache Hive and Apache HBase constitute the Apache Hadoop. The model of the MapReduce is used for programming and execution. It is also capable of processing and generating large volume of data sets. The underlying algorithm that is used by MapReduce is divide and conquer. The divide and conquer algorithm works by breaking a complex high-level problem into several sub-problems recursively. The sub-problems are then allocated to a cluster of working notes which solve these problems separately and in parallel. Later, the solutions are merged to give a solution to the problem in question. The Map step and the Reduce step are the two steps that are used to implement this algorithm. The two nodes that the Hadoop works on are master nodes and worker nodes. The role of the master nodes is to take the input and divide it into smaller sub-problems which are further distributed to worker nodes. Finally, the master node collects the solutions to all of the allocated sub-problems and aggregates them to produce output in Reduce step.

By default, Hadoop uses Hadoop distributed file system (HDFS). Hadoop also has the capability of working on other file systems as well. The HDFS uses the storage cluster arrays to hold the actual data. The data is dumped in HDFS, and it can be analysed within Hadoop, or it can export the data to other tools for performing analysis. The patterns used by Hadoop have three stages:

- LOAD—data into HDFS
- OPERATE—Map and Reduce sub-operations
- RETRIEVE—retrieve results from HDFS

The entire process is a batch operation. This is most suitable for analytical or non-interactive tasks. Hadoop cannot be termed as a data warehouse solution, neither it is a database, but it can support the analytical processes of the data. As an example, Facebook is the most popular application that follows the patterns of Hadoop. A database like MySQL stores the data, and this data is then replicated in Hadoop for computations and analysis (Fig. 2.5).

Fig. 2.4 Hadoop system

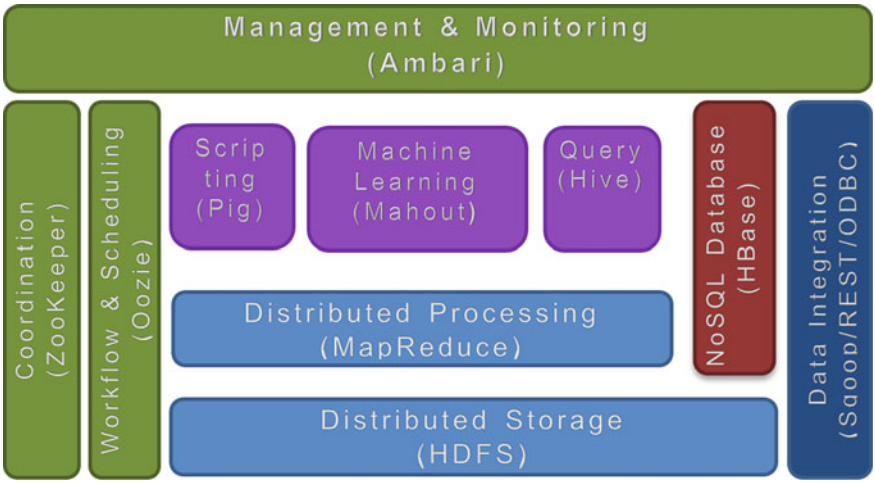

Fig. 2.5 Hadoop Ecosystem (Al-Barznji and Atanassov 2017)

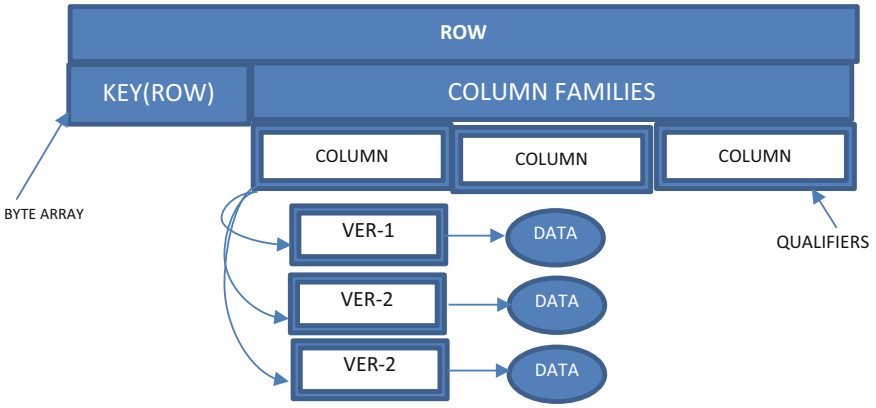

Fig. 2.6 Structure of HBase (Haines 2014)

HBase It is the Hadoop database, a NoSQL database that runs on Hadoop. It runs on the HD file system (HDFS) and provides scalability and real-time data access. This is provided as a key-value store along with the analytic capabilities of MapReduce. As the HBase is not a traditional relational database structure, it uses different methodology to model data. A four-dimensional data model is proposed for HBase. Each dimension is defined as under (Haines 2014), and the following four coordinates define each cell (Fig. 2.6):

- **Row Key**: Each of the rows in HBase has a unique key termed as row key. It is a byte array without a data type.

- **Column Family**: Data in the rows are structured into column families with every row having the same set of column families. HBase stores column families in its own data files. Any changes to be made to column families are difficult to incorporate; therefore, they need to be cautiously defined.
- **Column Qualifier**: The actual columns are referred as column qualifiers. Spread across different rows, the same column families do not require the same column qualifiers.
- **Version**: A configurable number of versions can be associated with each column. Data can be accessed for a specific version for a qualifier.

2.7.2 Hadoop in Business Organizations

The usage of Hadoop in business organizations can be understood in this example. The production server of a company stores a data set that deals with the structure and the business dealings of the company. This server facilitates the copying of the data set to an analytics engine. This can be a Hadoop cluster and can assist in the analytics of the data set. In order to prepare data and to copy it to be ready for analytics, three major processes are used. These are extract, transform and load (ETL) processes. The traditional ETL process consumes a lot of network resources along with the processing power and the bandwidth. It is noted that about 80% of the time is consumed by ETL process from each of the analytical job. Accordingly, traditional ETL may lead to excessive resource consumption and/or prolonged processing times in connection with analytics jobs.

Many organizations expand the data generated by the internal sources such as sales and services with the external demographics and social media, using the Hadoop-based analytics (Hortonworks 2013). These focus on the following key issues:

- How to identify new customer segments
- How to personalize offers
- How to reduce the customer roil

The Hadoop-based analytics can also help in businesses by reducing maintenance costs and improve asset utilization in asset-intensive industries, such as utilities, oil and gas, and industrial manufacturing. The machine-generated data along with the internally generated service data and external data can be integrated and used for predictive analytics. These businesses can manage the maintenance intervals as desired by the companies.

The Hadoop-based analytics find its application across broad spectrum of industries today. Applications like retail management use these for site selection, brand analysis, loyalty programs, market-based analysis along with sentiment analysis of the products. The financial service-providing organizations leverage Hadoop for fraud detection and risk assessment. Hadoop-based analytics are used

by government agencies for applications that relate to law enforcement, public transportation, national security, health and public safety (Hortonworks 2013).

2.7.3 *Apache SPARK*

It is an open-source framework that is available for processing of big data. It was inititally developed in the AMP Lab at U.C. Berkeley in 2009 and later was later open sourced in 2010 as an Apache project. The concept behind Spark is to provide a memory abstraction which allows efficient sharing of data. The data is shared across different stages of a map-reduce job. It also provides in-memory data sharing. It provides a comprehensive, unified framework which can manage big data processing requirements for data sets that are from myriad sources. These sources could be real-time data streaming or batch processing, online Web-sourced data, etc. Hadoop cluster application is very fast and can execute up to a hundred times faster in memory and ten times faster when running on disc (Shanahan and Dai 2015). Deployment of Spark can be done in different ways. It can provide native bindings for Java, Python, Scala and R programming languages. It also supports streaming data, SQL, machine learning along with graph processing. Apache Spark provides the potential and power of big data along with support for real-time analytics to the business organizations.

Spark Ecosystem Apache Spark is an open source cluster computing system. It consists of libraries and framework ecosystems for advanced data analytics. Apache Spark is a powerful and easy to use tool and is more productive as compared to the MapReduce. It also provides in-memory, faster runtimes and support for distributed computing. Some other libraries in addition to the Spark Core API library are also part of the Spark Ecosystem. These are capable of providing advanced capabilities for big data analysis as shown in Fig. 2.7 (Hightower and Maalouli 2015; Penchikala 2015).

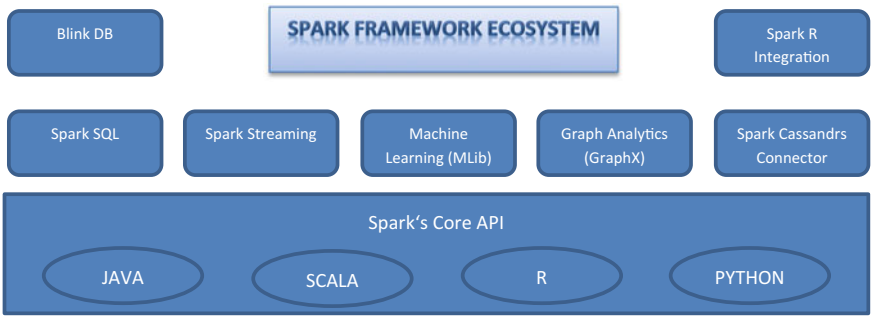

Fig. 2.7 Apache Spark Ecosystem (Hightower and Maalouli 2015; Penchikala 2015)

The base is the core engine on which the entire ecosystem is built. The API has support for Scala, Java, Python and R programming languages. Various libraries which provide additional computational power to spark are as follows:

- Spark Streaming can be used to process real-time streaming data which is based on the micro-batch style (i.e. splitting the input data into small batches) of computing and processing. To process the real-time data stream, DStream is used, which is a series of resilient distributed data sets (RDDs) (Zaharia et al. 2016).
- Spark data sets can be uncovered by Spark SQL which runs over the JDBC API. With the help of the traditional business intelligence along with the help of the visualization tools, Spark SQL allows running of SQL-like queries on big data. It also helps the users in data extraction from different formats (like Parquet, JSON or a database), transforming it and then finally exposing using to handle ad hoc queries.
- The machine learning library of Spark is the MLlib. It consists of both supervised and unsupervised machine learning algorithms, which include data classification, data clustering, linear and logistics regression, collaborative filtering, dimensionality reduction of data and optimization primitives("Big Data Processing with Apache Spark—Part 1: Introduction", n.d.).
- To compute graphs, the Spark GraphX component of Spark API is used. GraphX extends the Spark resilient distributed data set (RDD). It introduces the resilient distributed property graph which is a directed multigraph along with the properties that are associated with every edge and every vertex. GraphX also includes a collection of graph algorithms for simplifying graph analysis.

There are also adapters for integration with other products like Cassandra (Spark Cassandra Connector) and R (SparkR). The Cassandra connector allows Apache Spark to access data that is stored in a Cassandra database for data analysis.

2.8 Challenges of Big Data

With the growing size of Big data and the use of analytics, many challenges are uncovered. They are presented as under (Saravanakumar and Nandini 2017):

- **Size–Volume**
  New technologies have been proposed to facilitate the user and allow to store and query large data sets. However, the volume of data that is generated today is enormous; hence, new techniques with new algorithm along with new technology platform and ability to understand the data structure and business values is essential. To handle such challenges, "Data Scientists" with multidisciplinary expertise are required.

- **Acceptance of Big Data**
  It involves client motivation so as to acknowledge big data as a medium or channel for accepting and adopting new procedures and system. The acceptance of the same is time-consuming because to understand the big data and analytics is a tough task.
- **Understanding Analytics**
  Understanding the analytics to reduce the size and improve the business value is a major challenge. This happens as the objects that have to be modelled are of different nature than the contemporary. They are huge, complex and distributed. To handle this challenge, modelling and simulation techniques are needed which should be simple, robust, distributed and parallel computing.
- **Capturing Data**
  The data that constitutes the big data is of different types. They can be unstructured, semi-structured or unstructured. To capture such data for analysis is also quite challenging for the business organization.
- **Data Curation**
  In the big data era, data curation indicates processes and activities that are related to the organization along with the integration of data that is collected from a variety of sources. The data curation has become important as the software processes very high volume of complex data. It also includes data annotation, publication and presentation. Technically, data curation indicates the process of extracting of relevant information from large data set of interest.
- **Data Visualization**
  There is a vibrant problem of data visualization. As the big data is enormous, the users when access complex information and handle associated tasks, there is a vibrant difficulty faced by the users. In order to face these challenges, system software need to be used judiciously.
- **Performance and Scalability**
  The two main concerns and challenges of storage and processing enormous volume of big data systems are performance and scalability. To accomplish this, process analysis can be used so as to improve the performance and scalability.
- **Distributed Storage**
  Big data storage depends on distributed storages as the huge volume of data can only be stored and accessed in distributed platforms. These storages have to be handled technically and intelligently so that the data is available on the go. Handling the high volume and high velocity of big data is also a big challenge. Security of the data in motion also poses a great amount of trial.
- **Content Validation**
  Another major challenge that the data of the Internet face is the content validation. A large number of data sources like blogs, social networking sites, tweets, comments have information that is difficult to validate. Automated validation may be performed by using the machine learning algorithms to extract and validate the Web content.

2.9 Conclusion and Future Trends

In the past decade, the data exploded and became bigger and bigger ranging from terabytes to petabytes to exabytes. The business intelligence has revolutionized in past few years. Cloud technology has gained the maximum acceptance. Business houses rely on this structure for their data storage. However, the data can be stored in big reservoirs termed as data lakes. Unlike the data that the traditional databases use, the big data comprises unstructured, semi-structured and structured data that is generated from a large number and a variety of data sources. The big data is said to be at rest when it is stored in a stable structure, whereas the data in motion is when the data is in transit and has not reached the repository. The missionary data structure storages took a backseat and big data provided an actionable and insightful data presented with visualizations and interactive business dashboards.

Business intelligence was in full boom in year 2017. Trends that are present in the year 2017 will continue in 2018, but additional trends in analytics will be seen. The adopted strategies for analytics will be increasingly customizable. The question for business organizations would be somewhat like "*What is the best solution that is available?*" for business and what opportunities can be explored. The expected analytics and business intelligence trends for 2018 include (Lebied 2017) use of artificial intelligence for business intelligence; use of analytics tools—predictive and prescriptive; data quality management; the multicloud strategy deployment; data governance; natural language processing; security concerns; chief data officer —roles and responsibility embedded and collaborative business intelligence.

References

Al-Barznji, K., & Atanassov, A. (2017). Collaborative filtering techniques for generating recommendations on big data. In *International Conference Automatics and Informatics* (pp. 225–228). Sofia, Bulgaria.

Beaver, D. (2008). *10 billion photos*. Retrieved from https://www.facebook.com/notes/facebook-engineering/10-billion-photos/30695603919/.

Bertolucci, J. (2013). *How ancestry.com manages generations of big data—Information week*. Retrieved from https://www.informationweek.com/big-data/big-data-analytics/how-ancestrycom-manages-generations-of-big-data/d/d-id/1112975?.

Big Data Processing with Apache Spark—Part 1: Introduction. (n.d.). Retrieved from https://www.infoq.com/articles/apache-spark-introduction.

Brock, V. & Khan, H. U. (2017). Big data analytics: Does organizational factor matters impact technology acceptance? *Journal of Big Data, 4*(1). https://doi.org/10.1186/s40537-017-0081-8.

Business Intelligence—BI—Gartner IT glossary. (n.d.). Retrieved from https://www.gartner.com/it-glossary/business-intelligence-bi/.

Data mining versus data analysis and analytics—Fraud and fraud detection—Academic library—free online college e textbooks. (n.d.). Retrieved from https://ebrary.net/13380/business_finance/data_mining_versus_data_analysis_analytics.

Haines, S. (2014). *Introduction to HBase, the NoSQL database for Hadoop | Introduction to HBase | InformIT*. Retrieved from http://www.informit.com/articles/article.aspx?p=2253412.

Hightower, R., & Maalouli, F. (2015). *Introduction to big data analytics w/ Apache Spark Pt. 1—DZone Big Data*. Retrieved from https://dzone.com/articles/introduction-to-bigdata-analytics-with-apache-spar-6.

Hortonworks, A. (2013). *The business analyst's guide to Hadoop get ready, get set, and go: A three-step guide to implementing Hadoop-based analytics*. Retrieved from https://hortonworks.com/wp-content/uploads/2013/01/Alteryx-Hadoop-Whitepaper-Final1.pdf.

Jain, V. K. (2015). *Big Data and Hadoop*. Khanna Publishers. Retrieved from https://Books.Google.Co.In/Books?Id=I6nodqaaqbaj&Printsec=Frontcover&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false.

Lebied, M. (2017). *Top 10 analytics & business intelligence trends for 2018*. Retrieved from https://www.datapine.com/blog/business-intelligence-trends/.

Manyika, J., Chui, M., Brad, B., Bughin, J., Dobbs, R., Roxburgh, C., & Hung, A. (2011). *McKinsey Global Institute The McKinsey Global Institute*. Retrieved from https://www.mckinsey.com/~/media/McKinsey/Business-Functions/McKinsey-Digital/Our-Insights/Big-data-The-next-frontier-for-innovation/MGI_big_data_exec_summary.ashx.

Meet Hadoop—Hadoop: The definitive guide (3rd ed) (n.d.). Retrieved from https://www.safaribooksonline.com/library/view/hadoop-the-definitive/9781449328917/ch01.html.

Miller, S. (2014). Collaborative approaches needed to close the big data skills gap. *Journal of Organization Design, 3*(1), 26. https://doi.org/10.7146/jod.9823.

Oussous, A., Benjelloun, F. Z., Ait Lahcen, A., & Belfkih, S. (2017). Big data technologies: A survey. *Journal of King Saud University—Computer and Information Sciences*. Retrieved from https://doi.org/10.1016/j.jksuci.2017.06.001.

Penchikala, S. (2015). *Big data processing with apache spark—Part 1: Introduction*. Retrieved from https://www.infoq.com/articles/apache-spark-introduction.

Philip Chen, C. L., & Zhang, C. Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on big data. *Information Sciences, 275,* 314–347. https://doi.org/10.1016/j.ins.2014.01.015.

Saravanakumar, R., & Nandini, C. (2017). A survey on the concepts and challenges of big data: Beyond the hype. *Advances in Computational Sciences and Technology*, *10*(5), 875–884. http://www.ripublication.com.

Shanahan, J. G., & Dai, L. (2015). Large scale distributed data science using apache spark. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining—KDD '15* (pp. 2323–2324). New York, New York, USA: ACM Press. https://doi.org/10.1145/2783258.2789993.

Stubbs, E. (2014). In *Wiley Big Data Series: Big data, big innovation : Enabling competitive differentiation through business analytics*.

Vossen, G. (2014). Big data as the new enabler in business and other intelligence. *Vietnam Journal of Computer Science*, *1*(1), 3–14. Retrieved from https://doi.org/10.1007/s40595-013-0001-6.

Want to make big bucks in stock market? Use big data analytics. Retrieved from https://analyticsindiamag.com/want-make-big-bucks-use-big-data-analytics/.

What is big data?—Gartner IT glossary. Retrieved from https://www.gartner.com/it-glossary/big-data.

Zaharia, M., Franklin, M. J., Ghodsi, A., Gonzalez, J., Shenker, S., Stoica, I., et al. (2016). Apache Spark: A unified engine for big data processing. *Communications of the ACM, 59*(11), 56–65. https://doi.org/10.1145/2934664.