1

Introduction to the World of Big Data

CHAPTER OBJECTIVE

This chapter deals with the introduction to big data, defining what actually big data means. The limitations of the traditional database, which led to the evolution of Big Data, are explained, and insight into big data key concepts is delivered. A comparative study is made between big data and traditional database giving a clear picture of the drawbacks of the traditional database and advantages of big data. The three Vs of big data (volume, velocity, and variety) that distinguish it from the traditional database are explained. With the evolution of big data, we are no longer limited to the structured data. The different types of human- and machine-generated data—that is, structured, semi-structured, and unstructured—that can be handled by big data are explained. The various sources contributing to this massive volume of data are given a clear picture. The chapter expands to show the various stages of big data life cycle starting from data generation, acquisition, preprocessing, integration, cleaning, transformation, analysis, and visualization to make business decisions. This chapter sheds light on various challenges of big data due to its heterogeneity, volume, velocity, and more.

1.1 Understanding Big Data

With the rapid growth of Internet users, there is an exponential growth in the data being generated. The data is generated from millions of messages we send and communicate via WhatsApp, Facebook, or Twitter, from the trillions of photos taken, and hours and hours of videos getting uploaded in YouTube every single minute. According to a recent survey 2.5 quintillion (2 500 000 000 000 000 000, or 2.5×10^{18}) bytes of data are generated every day. This enormous amount of data generated is referred to as "big data." Big data does not only mean that the data sets are too large, it is a blanket term for the data that are too large in size, complex in nature, which may be structured or

Big Data: Concepts, Technology, and Architecture, First Edition. Balamurugan Balusamy, Nandhini Abirami. R, Seifedine Kadry, and Amir H. Gandomi.

unstructured, and arriving at high velocity as well. Of the data available today, 80 percent has been generated in the last few years. The growth of big data is fueled by the fact that more data are generated on every corner of the world that needs to be captured.

Capturing this massive data gives only meager value unless this IT value is transformed into business value. Managing the data and analyzing them have always been beneficial to the organizations; on the other hand, converting these data into valuable business insights has always been the greatest challenge. Data scientists were struggling to find pragmatic techniques to analyze the captured data. The data has to be managed at appropriate speed and time to derive valuable insight from it. These data are so complex that it became difficult to process it using traditional database management systems, which triggered the evolution of the big data era. Additionally, there were constraints on the amount of data that traditional databases could handle. With the increase in the size of data either there was a decrease in performance and increase in latency or it was expensive to add additional memory units. All these limitations have been overcome with the evolution of big data technologies that lets us capture, store, process, and analyze the data in a distributed environment. Examples of Big data technologies are Hadoop, a framework for all big data process, Hadoop Distributed File System (HDFS) for distributed cluster storage, and MapReduce for processing.

1.2 Evolution of Big Data

The first documentary appearance of big data was in a paper in 1997 by NASA scientists narrating the problems faced in visualizing large data sets, which were a captivating challenge for the data scientists. The data sets were large enough, taxing more memory resources. This problem is termed big data. Big data, the broader concept, was first put forward by a noted consultancy: McKinsey. The three dimensions of big data, namely, volume, velocity, and variety, were defined by analyst Doug Laney. The processing life cycle of big data can be categorized into acquisition, preprocessing, storage and management, privacy and security, analyzing, and visualization.

The broader term big data encompasses everything that includes web data, such as click stream data, health data of patients, genomic data from biologic research, and so forth.

Figure 1.1 shows the evolution of big data. The growth of the data over the years is massive. It was just 600 MB in the 1950s but has grown by 2010 up to 100 petabytes, which is equal to 100 000 000 000 MB.

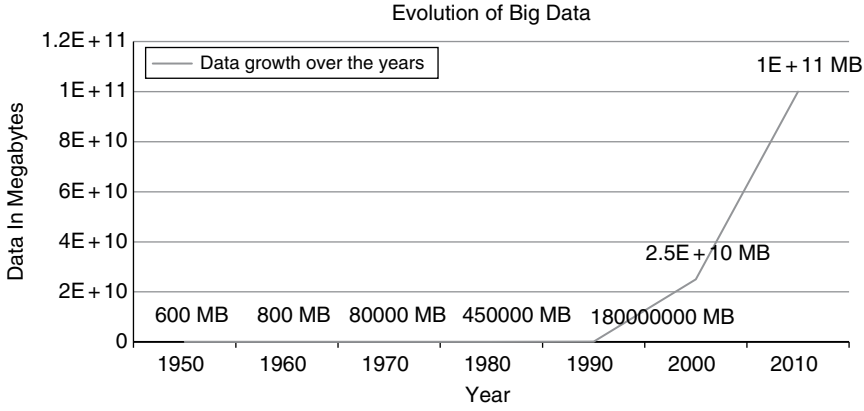

Figure 1.1 Evolution of Big Data.

1.3 Failure of Traditional Database in Handling Big Data

The Relational Database Management Systems (RDBMS) was the most prevalent data storage medium until recently to store the data generated by the organizations. A large number of vendors provide database systems. These RDBMS were devised to store the data that were beyond the storage capacity of a single computer. The inception of a new technology is always due to limitations in the older technologies and the necessity to overcome them. Below are the limitations of traditional database in handling big data.

- Exponential increase in data volume, which scales in terabytes and petabytes, has turned out to become a challenge to the RDBMS in handling such a massive volume of data.
- To address this issue, the RDBMS increased the number of processors and added more memory units, which in turn increased the cost.
- Almost 80% of the data fetched were of semi-structured and unstructured format, which RDBMS could not deal with.
- RDBMS could not capture the data coming in at high velocity.

Table 1.1 shows the differences in the attributes of RDBMS and big data.

1.3.1 Data Mining vs. Big Data

Table 1.2 shows a comparison between data mining and big data.

Table 1.1 Differences in the attributes of big data and RDBMS.

ATTRIBUTES	RDBMS	BIG DATA
Data volume	gigabytes to terabytes	petabytes to zettabytes
Organization	centralized	distributed
Data type	structured	unstructured and semi-structured
Hardware type	high-end model	commodity hardware
Updates	read/write many times	write once, read many times
Schema	static	dynamic

Table 1.2 Data Mining vs. Big Data.

S. No.	Data mining	Big data
1)	Data mining is the process of discovering the underlying knowledge from the data sets.	Big data refers to massive volume of data characterized by volume, velocity, and variety.
2)	Structured data retrieved from spread sheets, relational databases, etc.	Structured, unstructured, or semi-structured data retrieved from non-relational databases, such as NoSQl.
3)	Data mining is capable of processing large data sets, but the data processing costs are high.	Big data tools and technologies are capable of storing and processing large volumes of data at a comparatively lower cost.
4)	Data mining can process only data sets that range from gigabytes to terabytes.	Big data technology is capable of storing and processing data that range from petabytes to zettabytes.

1.4 3 Vs of Big Data

Big data is distinguished by its exceptional characteristics with various dimensions. Figure 1.2 illustrates various dimensions of big data. The first of its dimensions is the size of the data. Data size grows partially because the cluster storage with commodity hardware has made it cost effective. Commodity hardware is a low cost, low performance, and low specification functional hardware with no distinctive features. This is referred by the term “volume” in big data technology. The second dimension is the variety, which describes its heterogeneity to accept all the data types, be it structured, unstructured, or a mix of both. The third dimension is velocity, which relates to the rate at which the data is generated and being processed to derive the desired value out of the raw unprocessed data.

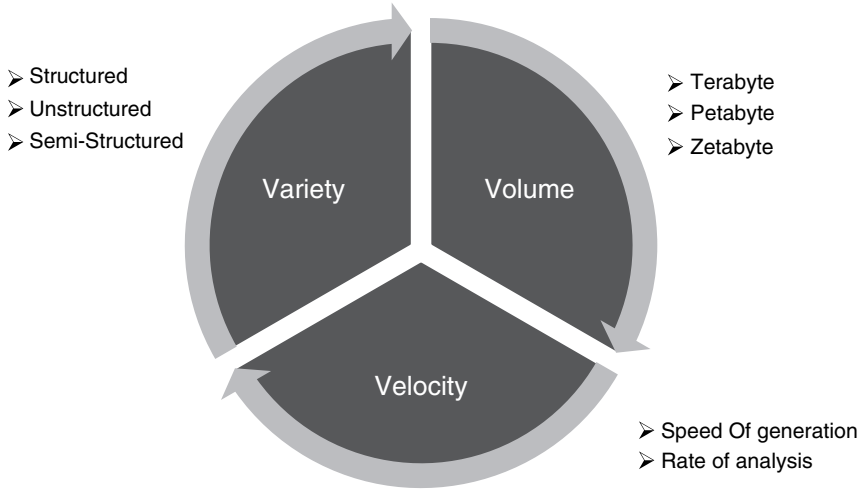

Figure 1.2 3 Vs of big data.

The complexities of the data captured pose a new opportunity as well as a challenge for today's information technology era.

1.4.1 Volume

Data generated and processed by big data are continuously growing at an ever increasing pace. Volume grows exponentially owing to the fact that business enterprises are continuously capturing the data to make better and bigger business solutions. Big data volume measures from terabytes to zettabytes (1024 GB = 1 terabyte; 1024 TB = 1 petabyte; 1024 PB = 1 exabyte; 1024 EB = 1 zettabyte; 1024 ZB = 1 yottabyte). Capturing this massive data is cited as an extraordinary opportunity to achieve finer customer service and better business advantage. This ever increasing data volume demands highly scalable and reliable storage. The major sources contributing to this tremendous growth in the volume are social media, point of sale (POS) transactions, online banking, GPS sensors, and sensors in vehicles. Facebook generates approximately 500 terabytes of data per day. Every time a link on a website is clicked, an item is purchased online, a video is uploaded in YouTube, data are generated.

1.4.2 Velocity

With the dramatic increase in the volume of data, the speed at which the data is generated also surged up. The term "velocity" not only refers to the speed at which data are generated, it also refers to the rate at which data is processed and

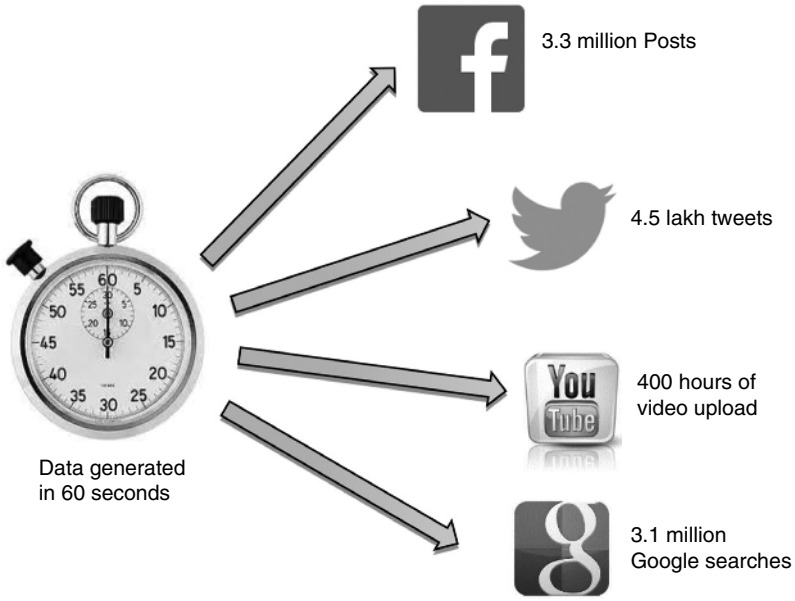

Figure 1.3 High-velocity data sets generated online in 60 seconds.

analyzed. In the big data era, a massive amount of data is generated at high velocity, and sometimes these data arrive so fast that it becomes difficult to capture them, and yet the data needs to be analyzed. Figure 1.3 illustrates the data generated with high velocity in 60 seconds: 3.3 million Facebook posts, 450 thousand tweets, 400 hours of video upload, and 3.1 million Google searches.

1.4.3 Variety

Variety refers to the format of data supported by big data. Data arrives in structured, semi-structured, and unstructured format. Structured data refers to the data processed by traditional database management systems where the data are organized in tables, such as employee details, bank customer details. Semi-structured data is a combination of structured and unstructured data, such as XML. XML data is semi-structured since it does not fit the formal data model (table) associated with traditional database; rather, it contains tags to organize fields within the data. Unstructured data refers to data with no definite structure, such as e-mail messages, photos, and web pages. The data that arrive from Facebook, Twitter feeds, sensors of vehicles, and black boxes of airplanes are all

Figure 1.4 Big data—data variety.

unstructured, which the traditional database cannot process, and here is when big data comes into the picture. Figure 1.4 represents the different data types.

1.5 Sources of Big Data

Multiple disparate data sources are responsible for the tremendous increase in the volume of big data. Much of the growth in data can be attributed to the digitization of almost anything and everything in the globe. Paying E-bills, online shopping, communication through social media, e-mail transactions in various organizations, a digital representation of the organizational data, and so forth, are some of the examples of this digitization around the globe.

- **Sensors:** Sensors that contribute to the large volume of big data are listed below.
  - *Accelerometer sensors* installed in mobile devices to sense the vibrations and other movements.
  - *Proximity Sensors* used in public places to detect the presence of objects without physical contact with the objects.
  - Sensors in vehicles and medical devices.
- **Health care:** The major sources of big data in health care are:
  - *Electronic Health Records* (EHRs) collect and display patient information such as past medical history, prescriptions by the medical practitioners, and laboratory test results.
  - *Patient portals* permit patients to access their personal medical records saved in EHRs.
  - *Clinical data repository* aggregates individual patient records from various clinical sources and consolidates them to give a unified view of patient history.
- **Black box:** Data are generated by the black box in airplanes, helicopters, and jets. The black box captures the activities of flight, flight crew announcements, and aircraft performance information.

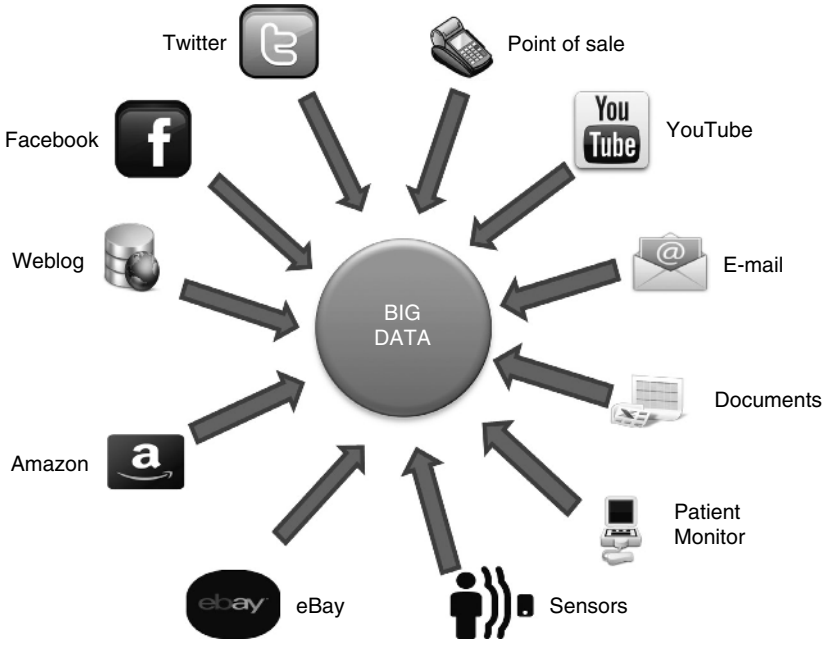

Figure 1.5 Sources of big data.

- **Web data:** Data generated on clicking a link on a website is captured by the online retailers. This is perform click stream analysis to analyze customer interest and buying patterns to generate recommendations based on the customer interests and to post relevant advertisements to the consumers.
- **Organizational data:** E-mail transactions and documents that are generated within the organizations together contribute to the organizational data.

Figure 1.5 illustrates the data generated by various sources that were discussed above.

1.6 Different Types of Data

Data may be machine generated or human generated. Human-generated data refers to the data generated as an outcome of interactions of humans with the machines. E-mails, documents, Facebook posts are some of the human-generated data. Machine-generated data refers to the data generated by computer applications or hardware devices without active human intervention. Data from sensors, disaster warning systems, weather forecasting systems, and satellite data are some of the machine-generated data. Figure 1.6 represents the data generated by a

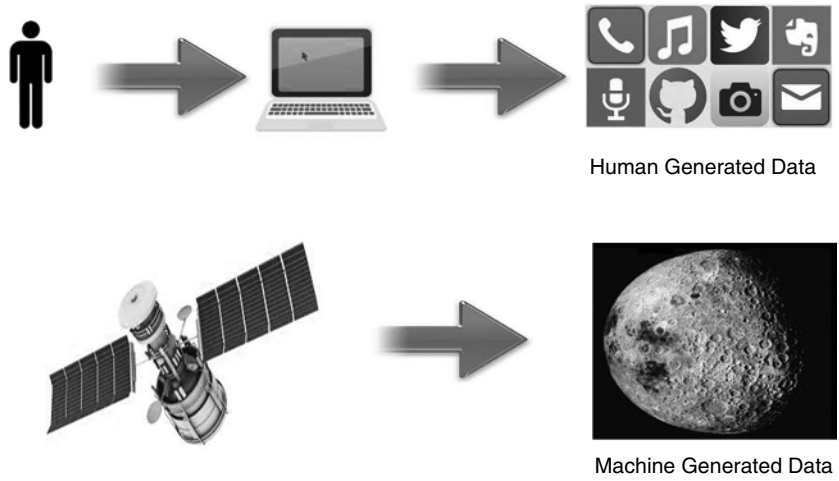

Figure 1.6 Human- and machine-generated data.

human in various social media, e-mails sent, and pictures that were taken by them and machine data generated by the satellite.

The machine-generated and human-generated data can be represented by the following primitive types of big data:

- Structured data
- Unstructured data
- Semi-structured data

1.6.1 Structured Data

Data that can be stored in a relational database in table format with rows and columns is called structured data. Structured data often generated by business enterprises exhibits a high degree of organization and can easily be processed using data mining tools and can be queried and retrieved using the primary key field. Examples of structured data include employee details and financial transactions. Figure 1.7 shows an example of structured data, employee details table with EmployeeID as the key.

1.6.2 Unstructured Data

Data that are raw, unorganized, and do not fit into the relational database systems are called unstructured data. Nearly 80% of the data generated are unstructured. Examples of unstructured data include video, audio, images, e-mails, text files,

Employee ID	Employee Name	Sex	Salary
334332	Daniel	Male	$2300
334333	John	Male	$2000
338332	Michael	Male	$2800
339232	Diana	Female	$1800
337891	Joseph	Male	$3800
339876	Agnes	Female	$4000

Figure 1.7 Structured data—employee details of an organization.

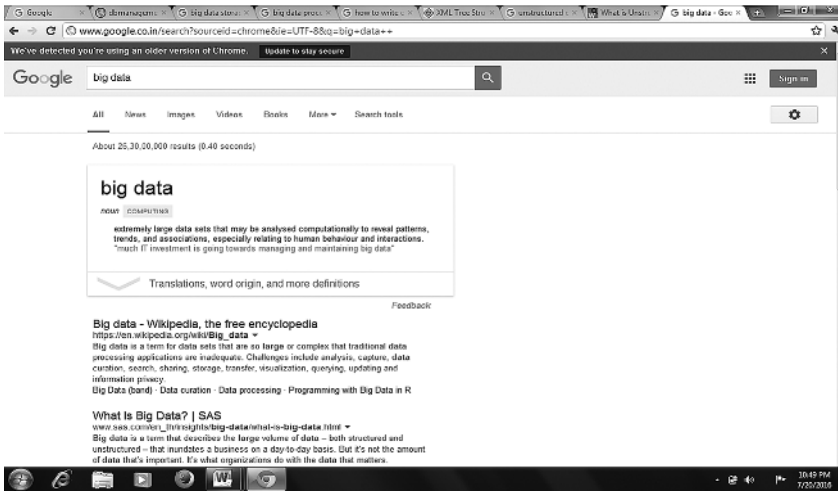

Figure 1.8 Unstructured data—the result of a Google search.

and social media posts. Unstructured data usually reside on either text files or binary files. Data that reside in binary files do not have any identifiable internal structure, for example, audio, video, and images. Data that reside in text files are e-mails, social media posts, pdf files, and word processing documents. Figure 1.8 shows unstructured data, the result of a Google search.

1.6.3 Semi-Structured Data

Semi-structured data are those that have a structure but do not fit into the relational database. Semi-structured data are organized, which makes it easier to analyze when compared to unstructured data. JSON and XML are examples of semi-structured data. Figure 1.9 is an XML file that represents the details of an employee in an organization.

```
<?xml version = "1.0"?>
<Company>
<Employee>
        <EmployeeId>339876</EmployeeId>
        <FirstName>Joseph</FirstName>
        <LastName>Agnes</LastName>
        <Sex>Female</Sex>
        <Salary>$4000<Salary>
</Employee>
</Company>
```

Figure 1.9 XML file with employee details.

1.7 Big Data Infrastructure

The core components of big data technologies are the tools and technologies that provide the capacity to store, process, and analyze the data. The method of storing the data in tables was no longer supportive with the evolution of data with 3Vs, namely volume, velocity, and variety. The robust RBDMS was no longer cost effective. The scaling of RDBMS to store and process huge amount of data became expensive. This led to the emergence of new technology, which was highly scalable at very low cost.

The key technologies include

- Hadoop
- HDFS
- MapReduce

Hadoop – Apache Hadoop, written in Java, is open-source framework that supports processing of large data sets. It can store a large volume of structured, semi-structured, and unstructured data in a distributed file system and process them in parallel. It is a highly scalable and cost-effective storage platform. Scalability of Hadoop refers to its capability to sustain its performance even under highly increasing loads by adding more nodes. Hadoop files are written once and read many times. The contents of the files cannot be changed. A large number of computers interconnected working together as a single system is called a cluster. Hadoop clusters are designed to store and analyze the massive amount of disparate data in distributed computing environments in a cost effective manner.

Hadoop Distributed File system – HDFS is designed to store large data sets with streaming access pattern running on low-cost commodity hardware. It does not require highly reliable, expensive hardware. The data set is generated from multiple sources, stored in an HDFS file system in a write-once, read-many-times pattern, and analyses are performed on the data set to extract knowledge from it.

MapReduce – MapReduce is the batch-processing programming model for the Hadoop framework, which adopts a divide-and-conquer principle. It is highly scalable, reliable, and fault tolerant, capable of processing input data with any format in parallel and distributed computing environments supporting only batch workloads. Its performance reduces the processing time significantly compared to the traditional batch-processing paradigm, as the traditional approach was to move the data from the storage platform to the processing platform, whereas the MapReduce processing paradigm resides in the framework where the data actually resides.

1.8 Big Data Life Cycle

Big data yields big benefits, starting from innovative business ideas to unconventional ways to treat diseases, overcoming the challenges. The challenges arise because so much of the data is collected by the technology today. Big data technologies are capable of capturing and analyzing them effectively. Big data infrastructure involves new computing models with the capability to process both distributed and parallel computations with highly scalable storage and performance. Some of the big data components include Hadoop (framework), HDFS (storage), and MapReduce (processing).

Figure 1.10 illustrates the big data life cycle. Data arriving at high velocity from multiple sources with different data formats are captured. The captured data is stored in a storage platform such as HDFS and NoSQL and then preprocessed to make the data suitable for analysis. The preprocessed data stored in the storage platform is then passed to the analytics layer, where the data is processed using big data tools such as MapReduce and YARN and analysis is performed on the processed data to uncover hidden knowledge from it. Analytics and machine learning are important concepts in the life cycle of big data. Text analytics is a type of analysis performed on unstructured textual data. With the growth of social media and e-mail transactions, the importance of text analytics has surged up. Predictive analysis on consumer behavior and consumer interest analysis are all performed on the text data extracted from various online sources such as social media, online retailing websites, and much more. Machine learning has made text analytics possible. The analyzed data is visually represented by visualization tools such as Tableau to make it easily understandable by the end user to make decisions.

1.8.1 Big Data Generation

The first phase of the life cycle of big data is the data generation. The scale of data generated from diversified sources is gradually expanding. Sources of this large volume of data were discussed under the Section 1.5, “Sources of Big Data.”

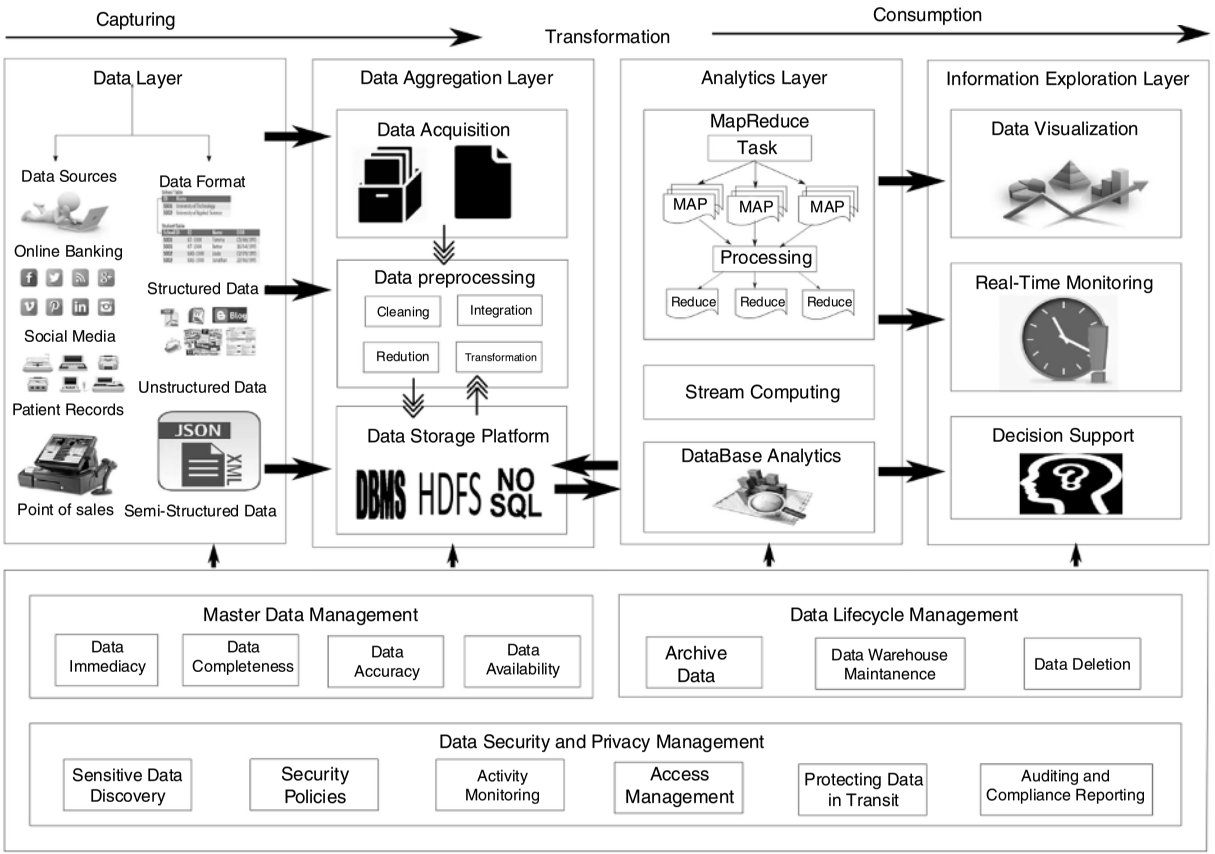

Figure 1.10 Big data life cycle.

1.8.2 Data Aggregation

The data aggregation phase of the big data life cycle involves collecting the raw data, transmitting the data to the storage platform, and preprocessing them. Data acquisition in the big data world means acquiring the high-volume data arriving at an ever-increasing pace. The raw data thus collected is transmitted to a proper storage infrastructure to support processing and various analytical applications. Preprocessing involves data cleansing, data integration, data transformation, and data reduction to make the data reliable, error free, consistent, and accurate. The data gathered may have redundancies, which occupy the storage space and increase the storage cost and can be handled by data preprocessing. Also, much of the data gathered may not be related to the analysis objective, and hence it needs to be compressed while being preprocessed. Hence, efficient data preprocessing is indispensable for cost-effective and efficient data storage. The preprocessed data are then transmitted for various purposes such as data modeling and data analytics.

1.8.3 Data Preprocessing

Data preprocessing is an important process performed on raw data to transform it into an understandable format and provide access to consistent and accurate data. The data generated from multiple sources are erroneous, incomplete, and inconsistent because of their massive volume and heterogeneous sources, and it is meaningless to store useless and dirty data. Additionally, some analytical applications have a crucial requirement for quality data. Hence, for effective, efficient, and accurate data analysis, systematic data preprocessing is essential. The quality of the source data is affected by various factors. For instance, the data may have errors such as a salary field having a negative value (e.g., salary = −2000), which arises because of transmission errors or typos or intentional wrong data entry by users who do not wish to disclose their personal information. Incompleteness implies that the field lacks the attributes of interest (e.g., Education = “”), which may come from a not applicable field or software errors. Inconsistency in the data refers to the discrepancies in the data, say date of birth and age may be inconsistent. Inconsistencies in data arise when the data collected are from different sources, because of inconsistencies in naming conventions between different countries and inconsistencies in the input format (e.g., date field DD/MM when interpreted as MM/DD). Data sources often have redundant data in different forms, and hence duplicates in the data also have to be removed in data preprocessing to make the data meaningful and error free. There are several steps involved in data preprocessing:

1) Data integration
2) Data cleaning
3) Data reduction
4) Data transformation

1.8.3.1 Data Integration

Data integration involves combining data from different sources to give the end users a unified data view. Several challenges are faced while integrating data; as an example, while extracting data from the profile of a person, the first name and family name may be interchanged in a certain culture, so in such cases integration may happen incorrectly. Data redundancies often occur while integrating data from multiple sources. Figure 1.11 illustrates that diversified sources such as organizations, smartphones, personal computers, satellites, and sensors generate disparate data such as e-mails, employee details, WhatsApp chat messages, social media posts, online transactions, satellite images, and sensory data. These different types of structured, unstructured, and semi-structured data have to be integrated and presented as unified data for data cleansing, data modeling, data warehousing, and to extract, transform, and load (ETL) the data.

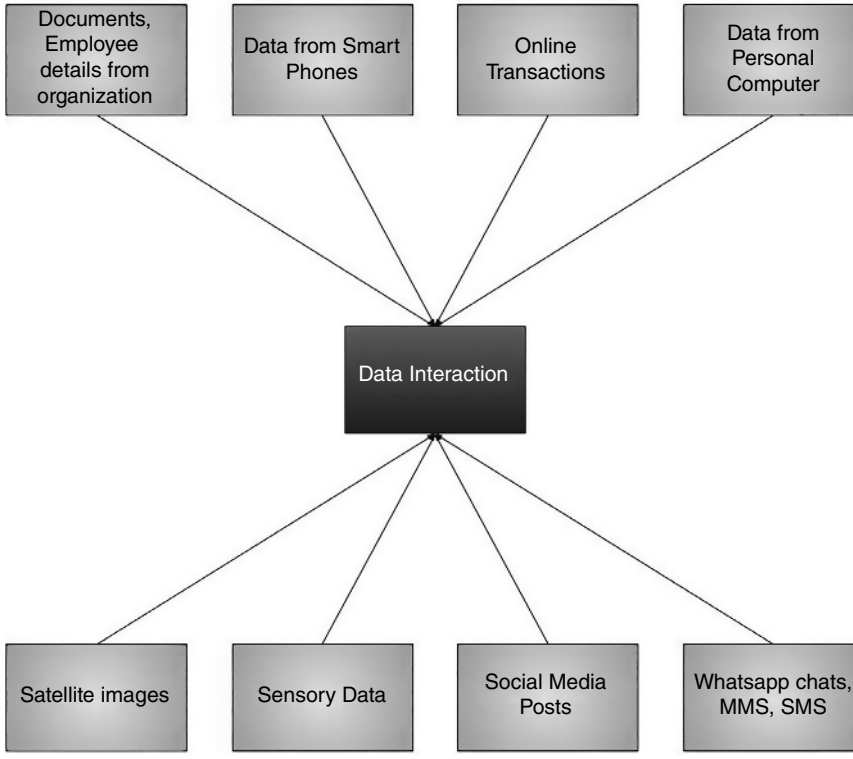

Figure 1.11 Data integration.

1.8.3.2 Data Cleaning

The data-cleaning process fills in the missing values, corrects the errors and inconsistencies, and removes redundancy in the data to improve the data quality. The larger the heterogeneity of the data sources, the higher the degree of dirtiness. Consequently, more cleaning steps may be involved. Data cleaning involves several steps such as spotting or identifying the error, correcting the error or deleting the erroneous data, and documenting the error type. To detect the type of error and inconsistency present in the data, a detailed analysis of the data is required. Data redundancy is the data repetition, which increases storage cost and transmission expenses and decreases data accuracy and reliability. The various techniques involved in handling data redundancy are redundancy detection and data compression. Missing values can be filled in manually, but it is tedious, time-consuming, and not appropriate for the massive volume of data. A global constant can be used to fill in all the missing values, but this method creates issues while integrating the data; hence, it is not a foolproof method. Noisy data can be handled by four methods, namely, regression, clustering, binning, and manual inspection.

1.8.3.3 Data Reduction

Data processing on massive data volume may take a long time, making data analysis either infeasible or impractical. Data reduction is the concept of reducing the volume of data or reducing the dimension of the data, that is, the number of attributes. Data reduction techniques are adopted to analyze the data in reduced format without losing the integrity of the actual data and yet yield quality outputs. Data reduction techniques include data compression, dimensionality reduction, and numerosity reduction. Data compression techniques are applied to obtain the compressed or reduced representation of the actual data. If the original data is retrieved back from the data that is being compressed without any loss of information, then it is called lossless data reduction. On the other hand, if the data retrieval is only partial, then it is called lossy data reduction. Dimensionality reduction is the reduction of a number of attributes, and the techniques include wavelet transforms where the original data is projected into a smaller space and attribute subset selection, a method which involves removal of irrelevant or redundant attributes. Numerosity reduction is a technique adopted to reduce the volume by choosing smaller alternative data. Numerosity reduction is implemented using parametric and nonparametric methods. In parametric methods instead of storing the actual data, only the parameters are stored. Nonparametric methods stores reduced representations of the original data.

1.8.3.4 Data Transformation

Data transformation refers to transforming or consolidating the data into an appropriate format and converting them into logical and meaningful information for data management and analysis. The real challenge in data transformation

comes into the picture when fields in one system do not match the fields in another system. Before data transformation, data cleaning and manipulation takes place. Organizations are collecting a massive amount of data, and the volume of the data is increasing rapidly. The data captured are transformed using ETL tools.

Data transformation involves the following strategies:

Smoothing, which removes noise from the data by incorporating binning, clustering, and regression techniques.

Aggregation, which applies summary or aggregation on the data to give a consolidated data. (E.g., daily profit of an organization may be aggregated to give consolidated monthly or yearly turnover.)

Generalization, which is normally viewed as climbing up the hierarchy where the attributes are generalized to a higher level overlooking the attributes at a lower level. (E.g., street name may be generalized as city name or a higher level hierarchy, namely the country name).

Discretization, which is a technique where raw values in the data (e.g., age) are replaced by conceptual labels (e.g., teen, adult, senior) or interval labels (e.g., 0–9, 10–19, etc.)

1.8.4 Big Data Analytics

Businesses are recognizing the unrevealed potential value of this massive data and putting forward the tools and technologies to capitalize on the opportunity. The key to deriving business value from big data is the potential use of analytics. Collecting, storing, and preprocessing the data creates a little value. It has to be analyzed and the end users must make decisions out of the results to derive business value from the data. Big data analytics is a fusion of big data technologies and analytic tools.

Analytics is not a new concept: many analytic techniques, namely, regression analysis and machine learning, have existed for many years. Intertwining big data technologies with data from new sources and data analytic techniques is a newly evolved concept. The different types of analytics are descriptive analytics, predictive analytics, and prescriptive analytics.

1.8.5 Visualizing Big Data

Visualization makes the life cycle of big data complete assisting the end users to gain insights from the data. From executives to call center employees, everyone wants to extract knowledge from the data collected to assist them in making better decisions. Regardless of the volume of data, one of the best methods to discern relationships and make crucial decisions is to adopt advanced data analysis and visualization tools. Line graphs, bar charts, scatterplots, bubble plots, and pie charts are conventional data visualization techniques. Line graphs are

used to depict the relationship between one variable and another. Bar charts are used to compare the values of data belonging to different categories represented by horizontal or vertical bars, whose heights represent the actual values. Scatterplots are used to show the relationship between two variables (X and Y). A bubble plot is a variation of a scatterplot where the relationships between X and Y are displayed in addition to the data value associated with the size of the bubble. Pie charts are used where parts of a whole phenomenon are to be compared.

1.9 Big Data Technology

With the advancement in technology, the ways the data are generated, captured, processed, and analyzed are changing. The efficiency in processing and analyzing the data has improved with the advancement in technology. Thus, technology plays a great role in the entire process of gathering the data to analyzing them and extracting the key insights from the data.

Apache Hadoop is an open-source platform that is one of the most important technologies of big data. Hadoop is a framework for storing and processing the data. Hadoop was originally created by Doug Cutting and Mike Cafarella, a graduate student from the University of Washington. They jointly worked with the goal of indexing the entire web, and the project is called "Nutch." The concept of MapReduce and GFS were integrated into Nutch, which led to the evolution of Hadoop. The word "Hadoop" is the name of the toy elephant of Doug's son. The core components of Hadoop are HDFS, Hadoop common, which is a collection of common utilities that support other Hadoop modules, and MapReduce.

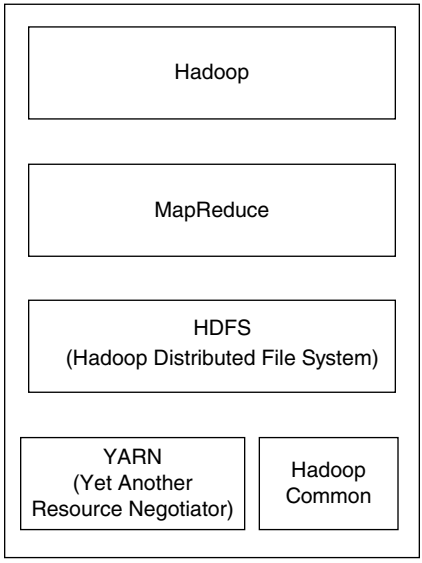

Figure 1.12 Hadoop core components.

Apache Hadoop is an open-source framework for distributed storage and for processing large data sets. Hadoop can store petabytes of structured, semi-structured, or unstructured data at low cost. The low cost is due to the cluster of commodity hardware on which Hadoop runs.

Figure 1.12 shows the core components of Hadoop. A brief overview

about Hadoop, MapReduce, and HDFS was given under Section 1.7, "Big Data Infrastructure." Now, let us see a brief overview of YARN and Hadoop common.

YARN – YARN is the acronym for Yet Another Resource Negotiator and is an open-source framework for distributed processing. It is the key feature of Hadoop version 2.0 of the Apache software foundation. In Hadoop 1.0 MapReduce was the only component to process the data in distributed environments. Limitations of classical MapReduce have led to the evolution of YARN. The cluster resource management of MapReduce in Hadoop 1.0 was taken over by YARN in Hadoop 2.0. This has lightened up the task of MapReduce and enables it to focus on the data processing part. YARN enables Hadoop to run jobs other than MapReduce jobs as well.

Hadoop common – Hadoop common is a collection of common utilities, which supports other Hadoop modules. It is considered as the core module of Hadoop as it offers essential services. Hadoop common has the scripts and Java Archive (JAR) files that are required to start Hadoop.

1.9.1 Challenges Faced by Big Data Technology

Indeed, we are facing a lot of challenges when it comes to dealing with the data. Some data are structured that could be stored in traditional databases, while some are videos, pictures, and documents, which may be unstructured or semi-structured, generated by sensors, social media, satellite, business transactions, and much more. Though these data can be managed independently, the real challenge is how to make sense by integrating disparate data from diversified sources.

- Heterogeneity and incompleteness
- Volume and velocity of the data
- Data storage
- Data privacy

1.9.2 Heterogeneity and Incompleteness

The data types of big data are heterogeneous in nature as the data is integrated from multiple sources and hence has to be carefully structured and presented as homogenous data before big data analysis. The data gathered may be incomplete, making the analysis much more complicated. Consider an example of a patient online health record with his name, occupation, birth data, medical ailment, laboratory test results, and previous medical history. If one or more of the above details are missing in multiple records, the analysis cannot be performed as it may not turn out to be valuable. In some scenarios a NULL value may be inserted in the place of missing values, and the analysis may be performed if that particular value

does not have a great impact on the analysis and if the rest of the available values are sufficient to produce a valuable outcome.

1.9.3 Volume and Velocity of the Data

Managing the massive and ever increasing volume of big data is the biggest concern in the big data era. In the past, the increase in the data volume was handled by appending additional memory units and computer resources. But the data volume was increasing exponentially, which could not be handled by traditional existing database storage models. The larger the volume of data, the longer the time consumed for processing and analysis.

The challenge faced with velocity does not only mean rate at which data arrives from multiple sources but also the rate at which data has to be processed and analyzed in the case of real-time analysis. For example, in the case of credit card transactions, if fraudulent activity is suspected, the transaction has to be declined in real time.

1.9.4 Data Storage

The volume of data contributed by social media, mobile Internet, online retailers, and so forth, is massive and was beyond the handling capacity of traditional databases. This requires a storage mechanism that is highly scalable to meet the increasing demand. The storage mechanism should be capable of accommodating the growing data, which is complex in nature. When the data volume is previously known, the storage capacity required is predetermined. But in case of streaming data, the required storage capacity is not predetermined. Hence, a storage mechanism capable of accommodating this streaming data is required. Data storage should be reliable and fault tolerant as well.

Data stored has to be retrieved at a later point in time. This data may be purchase history of a customer, previous releases of a magazine, employee details of a company, twitter feeds, images captured by a satellite, patient records in a hospital, financial transactions of a bank customer, and so forth. When a business analyst has to evaluate the improvement of sales of a company, she has to compare the sales of the current year with the previous year. Hence, data has to be stored and retrieved to perform the analysis.

1.9.5 Data Privacy

Privacy of the data is yet another concern growing with the increase in data volume. Inappropriate access to personal data, EHRs, and financial transactions is a social problem affecting the privacy of the users to a great extent. The data has to

be shared limiting the extent of data disclosure and ensuring that the data shared is sufficient to extract business knowledge from it. Whom access to the data should be granted to, limit of access to the data, and when the data can be accessed should be predetermined to ensure that the data is protected. Hence, there should be a deliberate access control to the data in various stages of the big data life cycle, namely data collection, storage, and management and analysis. The research on big data cannot be performed without the actual data, and consequently the issue of data openness and sharing is crucial. Data sharing is tightly coupled with data privacy and security. Big data service providers hand over huge data to the professionals for analysis, which may affect data privacy. Financial transactions contain the details of business processes and credit card details. Such kind of sensitive information should be protected well before delivering the data for analysis.

1.10 Big Data Applications

- Banking and Securities – Credit/debit card fraud detection, warning for securities fraud, credit risk reporting, customer data analytics.
- Healthcare sector – Storing the patient data and analyzing the data to detect various medical ailments at an early stage.
- Marketing – Analyzing customer purchase history to reach the right customers in order market their newly launched products.
- Web analysis – Social media data, data from search engines, and so forth, are analyzed to broadcast advertisements based on their interests.
- Call center analytics – Big data technology is used to identify the recurring problems and staff behavior patterns by capturing and processing the call content.
- Agriculture–Sensors are used by biotechnology firms to optimize crop efficiency. Big data technology is used in analyzing the sensor data.
- Smartphones—Facial recognition feature of smart phones is used to unlock their phones, retrieve information about a person with the information previously stored in their smartphones.

1.11 Big Data Use Cases

1.11.1 Health Care

To cope up with the massive flood of information generated at a high velocity, medical institutions are looking around for a breakthrough to handle this digital flood to aid them to enhance their health care services and create a successful

business model. Health care executives believe adopting innovative business technologies will reduce the cost incurred by the patients for health care and help them provide finer quality medical services. But the challenges in integrating patient data that are so large and complex growing at a faster rate hampers their efforts in improving clinical performance and converting the assets to business value.

Hadoop, the framework of big data, plays a major role in health care making big data storage and processing less expensive and highly available, giving more insight to the doctors. It has become possible with the advent of big data technologies that doctors can monitor the health of the patients who reside in a place that is remote from the hospital by making the patients wear watch-like devices. The devices will send reports of the health of the patients, and when any issue arises or if patients' health deteriorates, it automatically alerts the doctor.

With the development of health care information technology, the patient data can be electronically captured, stored, and moved across the universe, and health care can be provided with increased efficiency in diagnosing and treating the patient and tremendously improved quality of service. Health care in recent trend is evidence based, which means analyzing the patient's healthcare records from heterogeneous sources such as EHR, clinical text, biomedical signals, sensing data, biomedical images, and genomic data and inferring the patient's health from the analysis. The biggest challenge in health care is to store, access, organize, validate, and analyze this massive and complex data; also the challenge is even bigger for processing the data generated at an ever increasing speed. The need for real-time and computationally intensive analysis of patient data generated from ICU is also increasing. Big data technologies have evolved as a solution for the critical issues in health care, which provides real-time solutions and deploy advanced health care facilities. The major benefits of big data in health care are preventing disease, identifying modifiable risk factors, and preventing the ailment from becoming very serious, and its major applications are medical decision supporting, administrator decision support, personal health management, and public epidemic alert.

Big data gathered from heterogeneous sources are utilized to analyze the data and find patterns which can be the solution to cure the ailment and prevent its occurrence in the future.

1.11.2 Telecom

Big data promotes growth and increases profitability across telecom by optimizing the quality of service. It analyzes the network traffic, analyzes the call data in real-time to detect any fraudulent behavior, allows call center representatives to modify subscribers plan immediately on request, utilizes the insight gained by analyzing

the customer behavior and usage to evolve new plans and services to increase profitability, that is, provide personalized service based on consumer interest.

Telecom operators could analyze the customer preferences and behaviors to enable the recommendation engine to match plans to their price preferences and offer better add-ons. Operators lower the costs to retain the existing customers and identify cross-selling opportunities to improve or maintain the average revenue per customer and reduce churn. Big data analytics can further be used to improve the customer care services. Automated procedures can be imposed based on the understanding of customers' repetitive calls to solve specific issues to provide faster resolution. Delivering better customer service compared to its competitors can be a key strategy in attracting customers to their brand. Big data technology optimizes business strategy by setting new business models and higher business targets. Analyzing the sales history of products and services that previously existed allows the operators to predict the outcome or revenue of new services or products to be launched.

Network performance, the operator's major concern, can be improved with big data analytics by identifying the underlying issue and performing real-time troubleshooting to fix the issue. Marketing and sales, the major domain of telecom, utilize big data technology to analyze and improve the marketing strategy and increase the sales to increase revenue.

1.11.3 Financial Services

Financial services utilize big data technology in credit risk, wealth management, banking, and foreign exchange to name a few. Risk management is of high priority for a finance organization, and big data is used to manage various types of risks associated with the financial sector. Some of the risks involved in financial organizations are liquidity risk, operational risk, interest rate risk, the impact of natural calamities, the risk of losing valuable customers due to existing competition, and uncertain financial markets. Big data technologies derive solutions in real time resulting in better risk management.

Issuing loans to organizations and individuals is the major sector of business for a financial institution. Issuing loans is primarily done on the basis of creditworthiness of an organization or individual. Big data technology is now being used to find the credit worthiness based on latest business deals of an organization, partnership organizations, and new products that are to be launched. In the case of individuals, the credit worthiness is determined based on their social activity, their interest, and purchasing behavior.

Financial institutions are exposed to fraudulent activities by consumers, which cause heavy losses. Predictive analytics tools of big data are used to identify new patterns of fraud and prevent them. Data from multiples sources such as shopping

patterns and previous transactions are correlated to detect and prevent credit card fraud by utilizing in-memory technology to analyze terabytes of streaming data to detect fraud in real time.

Big data solutions are used in financial institutions call center operations to predict and resolve customer issues before they affect the customer; also, the customers can resolve the issues via self-service giving them more control. This is to go beyond customer expectations and provide better financial services. Investment guidance is also provided to consumers where wealth management advisors are used to help out consumers for making investments. Now with big data solutions these advisors are armed with insights from the data gathered from multiple sources.

Customer retention is becoming important in the competitive markets, where financial institutions might cut down the rate of interest or offer better products to attract customers. Big data solutions assist the financial institutions to retain the customers by monitoring the customer activity and identify loss of interest in financial institutions personalized offers or if customers liked any of the competitors' products on social media.

Chapter 1 Refresher

1 Big Data is _____.
- **A** Structured
- **B** Semi-structured
- **C** Unstructured
- **D** All of the above

Answer: d
Explanation: Big Data is a blanket term for the data that are too large in size, complex in nature, and which may be structured, unstructured, or semi-structured and arriving at high velocity as well.

2 The hardware used in big data is _____.
- **A** High-performance PCs
- **B** Low-cost commodity hardware
- **C** Dumb terminal
- **D** None of the above

Answer: b
Explanation: Big data uses low-cost commodity hardware to make cost-effective solutions.

3 What does commodity hardware in the big data world mean?

- **A** Very cheap hardware
- **B** Industry-standard hardware
- **C** Discarded hardware
- **D** Low specifications industry-grade hardware

Answer: d

Explanation: Commodity hardware is a low-cost, low performance, and low specification functional hardware with no distinctive features.

4 What does the term "velocity" in big data mean?

- **A** Speed of input data generation
- **B** Speed of individual machine processors
- **C** Speed of ONLY storing data
- **D** Speed of storing and processing data

Answer: d

5 What are the data types of big data?

- **A** Structured data
- **B** Unstructured data
- **C** Semi-structured data
- **D** All of the above

Answer: d

Explanation: Machine-generated and human-generated data can be represented by the following primitive types of big data

- Structured data
- Unstructured data
- Semi-Structured data

6 JSON and XML are examples of _____.

- **A** Structured data
- **B** Unstructured data
- **C** Semi-structured data
- **D** None of the above

Answer: c

Explanation: Semi-structured data are that which have a structure but do not fit into the relational database. Semi-structured data are organized, which makes it easier for analysis when compared to unstructured data. JSON and XML are examples of semi-structured data.

7 _____ is the process that corrects the errors and inconsistencies.

A Data cleaning
B Data Integration
C Data transformation
D Data reduction

Answer: a
Explanation: The data-cleaning process fills in the missing values, corrects the errors and inconsistencies, and removes redundancy in the data to improve the data quality.

8 _____ is the process of transforming data into an appropriate format that is acceptable by the big data database.

A Data cleaning
B Data Integration
C Data transformation
D Data reduction

Answer: c
Explanation: Data transformation refers to transforming or consolidating the data into an appropriate format that is acceptable by the big data database and converting them into logical and meaningful information for data management and analysis.

9 _____ is the process of combining data from different sources to give the end users a unified data view.

A Data cleaning
B Data integration
C Data transformation
D Data reduction

Answer: b

10 _____ is the process of collecting the raw data, transmitting the data to a storage platform, and preprocessing them.

A Data cleaning
B Data integration
C Data aggregation
D Data reduction

Answer: c

Conceptual Short Questions with Answers

1 What is big data?

Big data is a blanket term for the data that are too large in size, complex in nature, which may be structured or unstructured, and arriving at high velocity as well.

2 What are the drawbacks of traditional database that led to the evolution of big data?

Below are the limitations of traditional databases, which has led to the emergence of big data.

- Exponential increase in data volume, which scales in terabytes and petabytes, has turned out to become a challenge to the RDBMS in handling such a massive volume of data.
- To address this issue, the RDBMS increased the number of processors and added more memory units, which in turn increased the cost.
- Almost 80% of the data fetched were of semi-structured and unstructured format, which RDBMS could not deal with.
- RDBMS could not capture the data coming in at high velocity.

3 What are the factors that explain the tremendous increase in the data volume?

Multiple disparate data sources are responsible for the tremendous increase in the volume of big data. Much of the growth in data can be attributed to the digitization of almost anything and everything in the globe. Paying e-bills, online shopping, communication through social media, e-mail transactions in various organizations, a digital representation of the organizational data, and so forth, are some of the examples of this digitization around the globe.

4 What are the different data types of big data?

Machine-generated and human-generated data can be represented by the following primitive types of big data

- Structured data
- Unstructured data
- Semi-Structured data

5 What is semi-structured data?

Semi-structured data are that which have a structure but does not fit into the relational database. Semi-structured data are organized, which makes it easier for analysis when compared to unstructured data. JSON and XML are examples of semi-structured data.

6 What does the three Vs of big data mean?

Volume–Size of the data

1) Velocity–Rate at which the data is generated and is being processed
2) Variety–Heterogeneity of data: structured, unstructured, and semi-structured

7 What is commodity hardware?

Commodity hardware is a low-cost, low-performance, and low-specification functional hardware with no distinctive features. Hadoop can run on commodity hardware and does not require any high-end hardware or supercomputers to execute its jobs.

8 What is data aggregation?

The data aggregation phase of the big data life cycle involves collecting the raw data, transmitting the data to a storage platform, and preprocessing them. Data acquisition in the big data world means acquiring the high-volume data arriving at an ever increasing pace.

9 What is data preprocessing?

Data preprocessing is an important process performed on raw data to transform it into an understandable format and provide access to a consistent and an accurate data. The data generated from multiple sources are erroneous, incomplete, and inconsistent because of their massive volume and heterogeneous sources, and it is pointless to store useless and dirty data. Additionally, some analytical applications have a crucial requirement for quality data. Hence, for effective, efficient, and accurate data analysis, systematic data preprocessing is essential.

10 What is data integration?

Data integration involves combining data from different sources to give the end users a unified data view.

11 What is data cleaning?

The data-cleaning process fills in the missing values, corrects the errors and inconsistencies, and removes redundancy in the data to improve the data quality. The larger the heterogeneity of the data sources, the higher the degree of dirtiness. Consequently, more cleaning steps may be involved.

12 What is data reduction?

Data processing on massive data volume may take a long time, making data analysis either infeasible or impractical. Data reduction is the concept of reducing the volume of data or reducing the dimension of the data, that is, the number of attributes. Data reduction techniques are adopted to analyze the data in reduced format without losing the integrity of the actual data and yet yield quality outputs.

13 What is data transformation?

Data transformation refers to transforming or consolidating the data into an appropriate format that is acceptable by the big data database and converting them into logical and meaningful information for data management and analysis.

Frequently Asked Interview Questions

1 Give some examples of big data.

Facebook is generating approximately 500 terabytes of data per day, about 10 terabytes of sensor data are generated every 30 minutes by airlines, the New York Stock Exchange is generating approximately 1 terabyte of data per day. These are examples of big data.

2 How is big data analysis useful for organizations?

Big data analytics is useful for the organizations to make better decisions, find new business opportunities, compete against business rivals, improve performance and efficiency, and reduce cost by using advanced data analytics techniques.