BIG DATA ANALYTICS

G. Sudha Sadasivam
Professor and Head
Department of Computer Science and Engineering
PSG College of Technology, Tamil Nadu

R. Thirumahal
Assistant Professor
Department of Computer Science and Engineering
PSG College of Technology, Tamil Nadu

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide. Oxford is a registered trade mark of
Oxford University Press in the UK and in certain other countries.

Published in India by
Oxford University Press
22 Workspace, 2nd Floor, 1/22 Asaf Ali Road, New Delhi 110002

First published in 2020

ISBN-13: 978-0-19-949722-5
ISBN-10: 0-19-949722-2

Typeset in Garamond
by B2K-BYTES 2 KNOWLEDGE, Tamil Nadu
Printed in India by

Cover design/illustration: Alok Rawat

For product information and current price, please visit www.india.oup.com

Preface

Data proliferation from weblogs, texts, videos, images, and sensors has made it difficult to organize, manage, analyse, disseminate, and draw decisions from data. Large-scale computing infrastructure along with reliable frameworks is required to handle such data-intensive computing tasks. Further, analytical models are required to make effective decisions from raw data to improve business. Big data analytics facilitates businesses to make such decisions in an efficient and agile manner. As data enhances business value, it is considered a precious resource analogous to crude oil. Big data analytics deals with effective usage of infrastructure to handle and derive decisions from voluminous, real-time, and heterogeneous data. Hence, the book aims at exposing its readers to the frameworks on which analytics can be performed as well as the techniques for data analytics on such frameworks.

This book on big data analytics is the fruition of expert knowledge gathered from industrial and academic experts as well as research done in big data analytics lab at PSG College of Technology, Tamil Nadu. This book provides a comprehensive treatment to the subject as required for undergraduate and postgraduate students of computer science and engineering, information technology, and other related disciplines. We have tried to make the language and the approach in the book simple and lucid with use cases so that the students can easily understand the concepts.

About the Book

The book aims to strike a balance between theory and practice in the domain of big data and analytics. It seeks to translate theory behind big data into principles and practices for a data analyst. Working of various algorithms has been illustrated with examples. The reader need not have prior knowledge of data mining concepts. Relevant codes are provided for installation and usage of various tools for big data analytics like Hadoop, MongoDB, Neo4j, Spark, and R. The text is organized into 16 chapters based on big data and NoSQL systems (Chapters 1–4), frameworks for handling big data (Chapters 5, 14, 15), theory and methods of big data analytics (Chapters 6–13), and infrastructure for big data (Chapter 16).

Features of the Book

The following are the salient features of the book:

- Easy-to-understand language and approach
- Chapter outlines and learning outcomes listed at the beginning of each chapter
- Illustrative discussion on big data frameworks and infrastructure
- Lucid algorithms for data analytics on big data frameworks and tools
- A number of solved numerical examples to supplement the text
- Practice exercises and codes for various case studies on Hadoop, R, Spark, MongoDB, Storm, and Neo4j
- Interview questions highlighted as boxed items in each chapter
- Point-wise summary at the end of each chapter to enable quick revision
- Chapter-end exercises comprising objective-type questions with answers, critical thinking questions, descriptive type questions, and numerical exercises.

Organization of the Book

The book is divided into 16 chapters:

Chapter 1 on Introduction to Big Data Analytics introduces its readers to the basic concepts of big data technology and its relationship to various disciplines related to data science. It helps them gain an understanding of the need and usage of big data technology in various organizations. The architecture of big data systems with a use case in health care system is discussed in this chapter.

Chapter 2 on Data Analytics Life Cycle enlists the issues in applying the traditional software development life cycle (SDLC) for data-centric projects. It discusses in detail the common techniques used to develop data-centric project along with the challenges. A detailed discussion on the various stages of a big data project is then provided followed by a case study.

Chapter 3 is on Introduction to R. R is a simple and easy-to-use tool for data analysis. This chapter elaborates on the features of R tool and R programming. Data visualization in R and data analytics using R Analytic Tool To Learn Easily (Rattle) are illustrated with examples. The chapter also explores the various functions in R for data exploration.

Chapter 4 is on NoSQL. The birth of NoSQL is attributed to the need to handle voluminous data in various formats with good performance. This chapter starts with an introduction to such schema-less data stores encompassing its need, characteristics, advantages, and limitations. The underlying principles of NoSQL stores including normalization and softschema is described with illustrations. CAP theorem and BASE properties of schemaless data stores are illustrated. Polyglot persistence with four categories of NoSQL stores—key-value, column family, document, and graph—is illustrated with an e-commerce application.

Chapter 5 on Hadoop provides a lucid explanation on Hadoop, a key-value store. This chapter illustrates the architecture of Hadoop, and its file system along with MapReduce programming paradigm. Installing Hadoop, and writing and executing MapReduce programs are elucidated with examples. A detailed discussion on the components of Hadoop ecosystem including YARN, HBase, Hive, Pig, Sqoop, Flume, and Zookeeper is provided.

Chapter 6 on Preprocessing presents the basic statistical measures for data summarization including mean, weighted mean, median, and mode for measuring the central tendency of data. It includes range, quartiles, interquartile range, variance, and standard deviation for measuring the dispersion of data. This chapter discusses the types of hypothetical tests used to infer the result of a hypothesis performed on sample data from a larger population. It describes how to extract features using principal component analysis (PCA). The chapter also details the process of t-test, ANNOVA test, and PCA using R.

Chapter 7 on Association Rule Mining starts with the basic data mining method—association rule mining—which is used to find association between frequent itemsets in a given data. It then discusses how apriori algorithm is used to find the frequent itemset(s) in a given data set. Finally, the generation of association rules from the identified frequent itemset(s) is provided.

Chapter 8 on Clustering provides a detailed discussion on common clustering approaches with examples. It presents partition-based technique, hierarchical methods like agglomerative (bottom-up) approach or divisive (top-down) approach, and BIRCH. Partition-based and hierarchical-based clustering in R are also illustrated. A MapReduce version of *k*-means algorithm is also explained.

Chapter 9 on Regression introduces regression and causal inference in data analysis. It explains systematic prediction using supervised models. Simple linear regression, multiple regression, and a discriminative classifier support vector machine (SVM) are discussed. All the methods are demonstrated in R and the MapReduce version of linear regression is also explained.

Chapter 10 on Classification introduces systematic prediction using supervised models. It explains how to apply decision tree algorithm and Naïve Bayes algorithm for classifying data sets and to predict the class of an unknown tuple. Illustrated decision tree classification and Naïve-based classification using R and Hadoop are covered in this chapter.

Chapter 11 on Time Series Analysis discusses forecasting models such as autocorrelation function, autoregression model, moving average model, and the ARMA and ARIMA models. It elaborates on finding the coefficients and forecasts using the Solver tool of Excel. The procedure to build the ARIMA model and forecasting the future value using NumXL tool are also discussed.

Chapter 12 on Text Analysis describes the processes and applications of text analysis, topic modelling, and sentimental analysis. The processes of text analysis, sentimental analysis, and topic modelling are demonstrated with examples using R.

Chapter 13 is on Mining Data Streams. Processing data in real time involves sampling, filtering, and summarization. Techniques for the same are discussed in this chapter. Data streaming using Kafka and Spark installed in R is illustrated with use cases.

Chapter 14 on NoSQL Databases—Neo4j and MongoDB—goes a step forward toward introducing the concepts of NoSQL databases such as graph databases and document databases. It provides the data models of graph database (Neo4j) and document database (MongoDB). Neo4j with cypher query language and the commands and methods of MongoDB queries are further illustrated.

Chapter 15 is on Big Data Tools—Spark and Storm. Spark is an in-memory data processing engine that is highly suitable for iterative processing that is used in machine learning. Storm is a distributed processing framework that can handle real-time data. The chapter explicates the architecture of Spark and Storm. Machine learning using Sparklyr—the layer that connects Spark to R—is illustrated with examples. Installation and usage of Storm libraries in R is also explained with examples.

Chapter 16 on Big Data Infrastructure discusses the architecture of database systems for big data along with sharding and replication. Data virtualization enables agile and effective data management. This chapter elucidates the use of data virtualization for big data applications.

The book ends with an Appendix that consists of answers to all 'Interview Questions' included in the chapters.

Acknowledgements

We are thankful to everybody who have motivated and guided us in preparing the manuscript. I bow my head before the almighty God who blessed us with health and confidence to undertake and complete the manuscript successfully.

We express our sincere thanks to the Principal and the management of PSG College of Technology, for their encouragement and support. We would like to thank Mr Chidambaran Kollengode, Head, Analytics Platform and Applications, LinkedIn, for collaborating in setting up the big data analytics lab and inspiring us to work in the area of big data analytics.

We are deeply indebted to our parents for raising us with values. We thank our family who always stood beside us and encouraged us to complete the manuscript by sharing our routine domestic chores. Their love and blessings are a perpetual source of inspiration to us.

Last but not the least, we would like to thank the editorial team at Oxford University Press, India for their help and support.

Comments and suggestions for the improvement of the book can be sent to us at sudhasadhasivam@yahoo.com and trk1193@gmail.com.

G. Sudha Sadasivam
R. Thirumahal

The publisher and the author would like to thank the following reviewers for their feedback:

- Amit Ganatra (Charotar University of Science and Technology, Gujarat)
- Rahul Katarya (Delhi Technological University, Delhi)
- Sachin Deshmukh (Vivekanand Education Society's Institute of Management Studies & Research, Maharashtra)
- Shreedhara K.S. (University BDT College of Engineering, Karnataka)
- Pran Hari Talukdar (Kaziranga University, Assam)
- Neelendra Badal (Kamla Nehru Institute of Technology Sultanpur, Uttar Pradesh)
- Ashish Kumar (ITS Engineering College, Greater Noida, Uttar Pradesh)
- Ajay Shankar (DIT University, Uttarakhand)
- Srinivasulu Reddy (NIT Trichy, Tamil Nadu)
- Haider Banka (IIT-ISM Dhanbad, Jharkhand)
- Rajesh Prasad (NBN Sinhgad School of Engineering, Maharashtra)
- Korra Sathya Babu (NIT Rourkela, Odisha)
- J. Suresh (SSN College of Engineering, Tamil Nadu)
- S. Karthikeyan (VIT AP University, Andhra Pradesh)
- A.C. Kaladevi (Sona College of Technology, Tamil Nadu)

Contents

CHAPTER 1

Introduction to Big Data Analytics

LEARNING OBJECTIVES

- Appreciate emerging trends in information extraction and analytics
- Understand data science and its overlapping disciplines
- Recognize the characteristics of big data
- Explore big data architecture
- Be aware of the composition of data science team

This chapter provides an introduction to big data. It presents emerging trends in information extraction, need for big data technologies, emergence of data science as a discipline, and characteristics of big data. The chapter also introduces its readers to the basic architecture of big data systems along with the frameworks and tools used in the industry. The chapter concludes by discussing the roles played by various members of a data science team and big data use cases.

1.1 OVERVIEW

The proliferation of the Internet has resulted in data explosion. The International Data Corporation (IDC) predicts a 40 times increase of digital data by 2020 when compared to 2012. Organization, management, analysis, dissemination, and knowledge discovery from such voluminous amount of data is a challenging task. Large-scale computing infrastructure along with reliable frameworks is necessary to handle such data-intensive computing tasks, while mathematical models for data analytics is essential to extract knowledge for such 'new-age' applications. State-of-the-art data processing necessitates the use of data management frameworks for acquiring, storing, cleaning, aggregation, representation, and dissemination of data and use of analytical engines for modelling, analysing, and interpreting data.

Online activities such as weblogs, text, videos, images, and sensors create large amounts of data that require modern and sophisticated systems for storage. This trend has initiated the birth of the big data era. In fact, data is the raw material of the current industrial revolution 4.0. Data is priceless as it forms the bottom-line for effective decision-making. True to the saying 'Data is the king', the future belongs to analytical solutions that can process data in real time. Data is a precious resource analogous to crude oil. Both have to be processed and refined to derive value. While refined oil has to be transported through pipelines to consumers, Internet and cloud computing technologies have made it easier for storage and usage of data. While crude oil reservoirs are diminishing, data resources are exploding.

Globally it is estimated that there are 98,000 tweets, 695,000 status updates, 11 million instant messages, 698,445 Google searches, 168 million emails sent, 217 new mobile users, and 1,820 TB data created every minute. Such an IT trend is possible through SMAC (Social, Mobile, Analytics, and Cloud). Data management has changed. In the past, data was maintained in premises and was primarily used for transaction processing using servers and consumed by desktop users. Examples of such applications include banking, inventory management, and employee information management systems. In the present, data is maintained in the cloud (outside the premises of the organization) and the applications require interactive management of real-time data. The output of such a data management system is consumed by a variety of devices including mobiles. Examples of such applications include Internet of Things (IoT), stock market prediction, personalized medicine, and precision agriculture. Figure 1.1 shows the characteristics of data handled by web generation applications including social media, sensors, and web logs. Such data is not only voluminous, but also heterogeneous and event-based.

In the field of healthcare, medical images, clinical reports, medical devices, and sensors create a huge influx of real-time data that can be used for effective diagnosis. Credit card companies identify fraudulent transactions based on the knowledge of billions of past transactions. Social networking sites like Facebook analyse data to generate valuable information for advertisement and identification of relationships among users. Genome sequencing fosters personalized medicine and treatment by predicting possible diseases using the patient's genetic makeup. Smart city initiatives analyse a lot of sensor data to predict the outcomes and control devices. Oil drilling platforms use data from 20k to 40k sensors to derive valuable information. We can conclude that the explosion of web applications and IoT has resulted in a new generation of data. Media, social networks, smart phones, Internet, sensors, archives, and log data are some common sources of voluminous, heterogeneous, real-time data that characterize the nature of current generation data—*big data*.

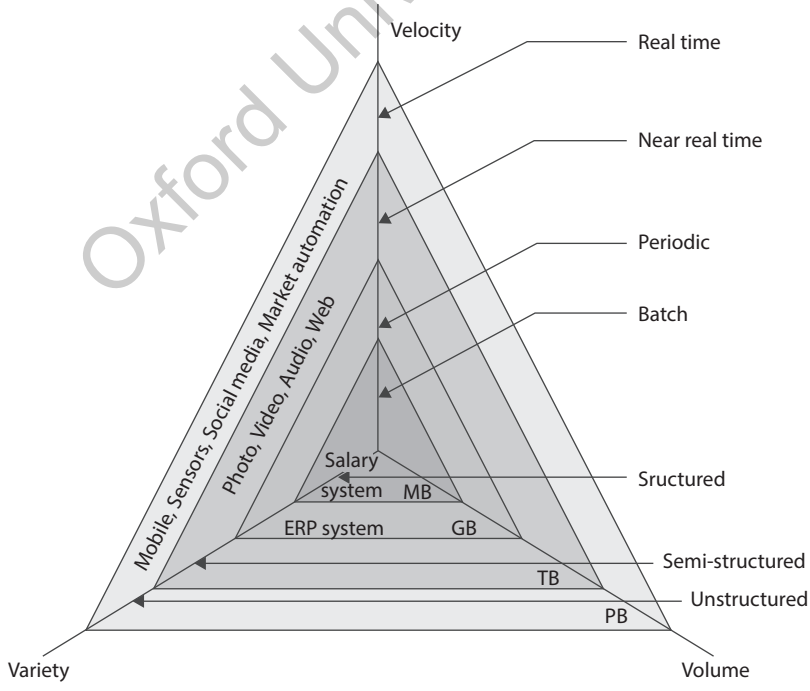

Fig. 1.1 Volume vs Variety vs Velocity of current generation of data

Information cannot be extracted from this data using conventional platforms and techniques. This has resulted in the birth of big data technologies.

Let us study the importance of big data technology in the current context. To study the progress of state-of-the-art technologies and techniques, Gartner uses the hype cycle. Gartner is a global research and advisory firm that advises on tools in domains like IT. The hype cycle provides a conceptual representation of the maturity of emerging technologies. The phases of a hype cycle include:

1. **Technology trigger** phase that initiates the cycle based on proof-of-concept and media interest.
2. **Peak of inflated expectations** result due to early success/failure stories of the new technology.
3. **Trough of disillusionment** shows a decrease of interest in the technology when experiments/implementations fail.
4. **Slope of enlightenment** results due to a better understanding of the business benefits due to the success stories of the emerging technology.
5. **Plateau of productivity** occurs when the technology moves into production stage with more adopters for the technology.

Currently, deep learning is at the peak of the hype cycle. Machine learning and data science are moving towards the trough, whereas big data and advanced analytics show more success. Predictive analytics and data mining are used extensively by industries in the production stage. Until 2010, the hype cycle included extreme transaction processing in the initial technology trigger phase. Big data was introduced as an emerging technology in 2011. In 2012, it progressed in the technology trigger phase and reached its peak by 2013. In 2014, it climbed down the trough of disillusion. However, big data was subsequently removed by Gartner from the hype cycle, as it has now become a part of all the emerging technologies such as IoT, content analytics, social analytics, and virtual reality. This means that big data technologies are already put in practice in various fields such as social media, healthcare, and agriculture. In 2015, machine learning was introduced in the place of big data in the disillusion phase and continued to be so until 2017. Since 2018, deep learning using multiple layers for learning based on the neural networks appearing in the hype cycle.

Until 2010, the term *Big Data* was virtually unknown, but by mid-2011 it became a state-of-the-art knowledge extraction technology. Big data cannot be converted into an asset unless it is analysed and insights are mined from it. This is where *big data analytics* comes into the picture. The process of mining relevant and useful information from the plethora of data being generated to make smart business decisions is called *big data analytics*. Analytics is a process of discovery, interpretation, and communication of meaningful data patterns for decision making. Data-driven analytics is highly effective in domains dealing with a large amount of recorded information such as business, healthcare, and government. Analytics uses the power of statistics, computer programming, and operation research to draw insight from data and communicate the results using visualization techniques. Analytics also supports organizations to consume the generated business data and to describe/predict insights that can improve the business. The following are the three sources of big data:

1. **Social data:** It comes from social media channel's insights on consumer behaviour.
2. **Machine data:** It consists of real-time data generated from sensors and web logs. It tracks user behaviour online.
3. **Transaction data:** It is generated by large retailers and B2B companies on a frequent basis.

INTERVIEW QUESTIONS

1. Why is the hype cycle important?
2. Why is analytics important for big data?
3. Enumerate the growth of big data as a technology.
4. What is analytics?

1.2 DATA SCIENCE

Analytics has been used for a long time. Analysing the geographic distribution of the population, the Swedish government in 1749 predicted their military requirements. By analysing nursing and hygiene, Nightingale (1850) predicted mortality rates. Doll in 1950 analysed the effect of smoking to predict lung cancer incidence. A study was done on the number of smokers and how many of them were affected by cancer. Simple statistical techniques were used and the analysis of data reveals correlations and patterns that can help us understand the following:

1. What has happened?
2. Why did it happen?
3. What is likely to happen in future?
4. What actions can be taken based on the analysis?
5. How to cause something to happen?

Thus the process of analysis can be descriptive, diagnostic, predictive, prescriptive, or cognitive in nature leading to knowledge discovery. Data analysis primarily deals with analysing past data and understanding it. This leads to knowledge discovery. *Data analytics* deals with using this knowledge to make smart business decisions in the future. Analytics is the heart of data science.

Peter Naur used the term *data science* as a substitute for computer science in 1960. Chikio Hayashi re-introduced the term in 1996 in the International Federation of Classification Societies (IFCS). *Data science* is the science of making sense out of data. It is related to cleansing, preparation, and analysis of structured/unstructured data to extract information from data. Data science uses a combination of techniques to increase the business value of data by obtaining insights from it. Data science combines mathematics, statistics, artificial intelligence, programming, problem-solving, domain knowledge, data cleansing, and data visualization techniques to extract knowledge. It also uses machine learning to make automated, smart, and informed decisions about data.

The following analytics disciplines are related to data science (see Fig. 1.2):

1. **Statistics** is basically a measure for an attribute(s) of a sample. It is extensively used in industrial, scientific, and social applications. This field of mathematics deals with the study of the collection, analysis, interpretation, presentation, and organization of data. Although statistics is a mathematical science, it has become domain-specific based on its application to a particular domain. Statisticians in the industry now function as data scientists. Industrial statistical science is used by non-statisticians like engineers to optimize on their engineering projects. Actuarial science is a subset of statistics used for the insurance domain. It extensively uses survival models. In fact, actuarial science is an active sub-domain of data science. Operations research (OR) and statistics are siblings. OR is a field of decision science that can be used to optimize business projects. Econometrics used in stock market analysis is a sub-field of statistical sciences dealing with forecasting using time-series approaches. It is essential that data analysts should possess deep knowledge in descriptive statistics and probability theory. The work of a data analyst lies in inference, which is the process of deriving conclusions that are solely based on what the researcher already knows; for example, running through a number of data sets to look for meaningful correlations between each other.
2. **Data mining** deals with designing algorithms to extract insights from data. It includes pattern recognition, feature selection, clustering, supervised classification, and some statistical techniques. Thus data mining is a subset of data science. Data mining is an application of computer science (rather than mathematical science) for a particular domain.
3. **Artificial intelligence (AI)** is a sub-domain under computer science, concerned with solving tasks using abstract intelligence. Such tasks are easy for humans, but hard for computers. Machine learning is a subset of AI that focuses on a narrow range of activities. It aims at applying AI to real-world problems.

4. **Machine learning** aims at applying AI techniques for mining data. It is an AI technique that is broadly used in data mining. It deals with designing algorithms (like data mining), but the emphasis is on designing automated systems by prototyping algorithms for production mode. Such systems automatically update/refine themselves by discovering new knowledge. To summarize, machine learning figures out the 'correct' action to be taken for a given AI problem based on the information of the surrounding environment. At its core, it deals with automated clustering and classification, rule-based systems, and scoring techniques. Machine learning uses a training dataset to build a model that can predict values. For example, future sales of a product can be predicted using current and past sales.

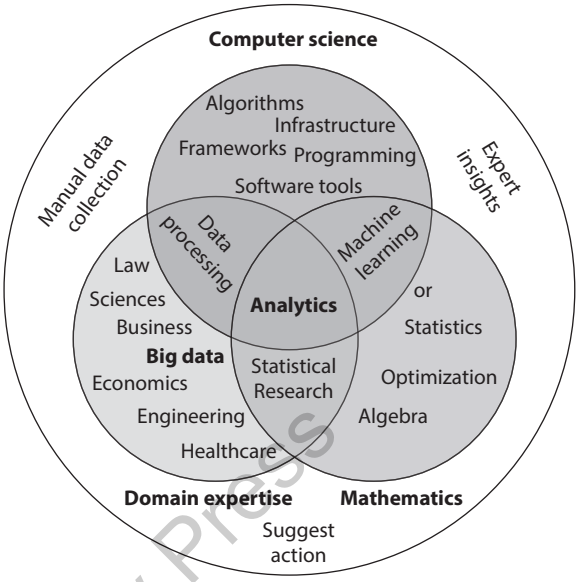

Fig. 1.2 Disciplines related to data science

5. Both machine learning and statistics converge on learning from data. While statistics is estimation-based, machine learning deals with automated learning. Machine learning is 'glorified statistics'. Compared to machine learning, data science need not obtain data from machines. It can be also collected manually through surveys. In addition, learning may not be involved in data science. Data science covers the whole spectrum of data processing, not just the algorithmic or statistical aspects. It includes architecture, data integration, machine learning, data visualization, business intelligence, automated data-driven decisions, and so on.
6. **Predictive modelling** draws its roots from statistics. These models predict future outcomes based on past data using statistical and machine learning techniques. For example, predicting the future sales of a product based on the past demands, is done by predictive modelling.
7. **Data analysis** primarily aims at gaining some insight on a dataset. This can be done by an analyst with/without the aid of tools. Analysis of data deals with assimilating patterns in the past data and understanding them. A stock market trader may use his experience to buy/sell a share without using an analytics tool.
8. **Data analytics** aims at using tools and techniques to discover knowledge from hidden patterns and to take effective actions for prediction. Business analytics uses business intelligence for data analysis. Data analysis deals with knowledge discovery. Data analytics uses this knowledge to make smart business decisions to enhance the business.

   There are 5 categories of analytics:

   (a) When analysis of data leads to describing patterns, it is called descriptive analytics
   (b) When knowledge discovery helps to understand the reason behind the occurrence of patterns, it is known as diagnostic analytics
   (c) When the knowledge discovered is used to predict future trends, then it is called predictive analytics
   (d) When the knowledge discovered can be used to suggest actions to be taken in future, it is known as prescriptive analytics
   (e) When analysis of the knowledge discovered identifies methodologies to suggest future actions, this is known as cognitive analytics

9. **Big data** is high-volume, high-velocity, and/or high-variety of data assets that can be used for effective decision making. Such data cannot be handled by conventional systems. Big data analytics can be called data science, but data science cannot be called big data analytics as it includes other technologies/techniques also. Both disciplines mine useful information from data, but, big data analytics involves mining useful information from raw data, whereas data science covers the whole spectrum of data processing, not just the algorithmic or statistical aspects. It can also include analytics done by experts manually.

   The term 'big data' is related to business intelligence (BI) and data mining. All the three aim to deliver business value through analysis of data. Big data differs from data mining and BI in the following means:
   (a) **Approaches used for data organization:** Traditional data warehouses organize data in the repository by aggregating it from multiple databases/systems. Such systems are poor in organizing and querying real-time operational/streaming data. Some examples of operational data include emails, chat transcripts, sensor data, click stream data, location data, surveillance data, etc. Big data technologies facilitate effective organization of real-time data from heterogeneous sources.
   (b) Techniques used in big data analytics harness new sources of data to extract intrinsic valuable information.
   (c) Data mining functions use sample data sets, whereas big data analytics uses real data.
   (d) Data mining aims at applying computer science to a particular domain. Big data analytics applies both mathematical techniques and computer science to an application.
10. **Data engineering** is a sub-field of software architecture basically dealing with hardware and frameworks to maintain data that is consumed by data scientists. Data engineers work in suggesting and implementing architectures to store, organize, and process data for different kinds of applications. They also work in transforming old architectures to newer ones to cope up with changing data trends.
11. **Business intelligence** deals with collecting useful data, analysing it, and visualizing it by creating data reports to extract valuable business insights.
12. **Deep learning** is a kind of machine learning using a category of mathematical models composed of simple blocks (function composition) of a certain type. To enhance prediction, these blocks can be adjusted as required.

Data science overlaps with the following:

1. **Computer science**—in developing data architectures, distributed architectures, optimizing data flows, programming, and using various structures to represent data.
2. **Statistics**—to analyse samples and design experiments.
3. **Machine learning and data mining** are extensively used by data science.
4. **Operations research techniques** aid data science in decision making.
5. **Business intelligence:** Data science techniques are applied extensively in business applications.

At the core of data science is analytics that deals with analysing data to discover knowledge. Analytics uses this knowledge to make decisions and prescribe actions in future.

INTERVIEW QUESTIONS

5. What are the types of analytics?
6. What are the disciplines that overlap with data science?

Contd

7. What is the work of data engineers?
8. Is it necessary for data science to incorporate machine learning?
9. What is data analytics?
10. What is the role of a data analyst?
11. What is data science?
12. Why is data science required?

1.3 BIG DATA CHARACTERISTICS

Volume, variety, and velocity of web generation of data form the key data management issues as discussed in the overview section. The American sociologist Charles Tilly in 1980 quoted the increased use of computer technology and statistics by the historians as 'none of the big questions has actually yielded to the bludgeoning of the big-data people'. NASA scientists (1997) had first documented their visualization issue using computer graphics as a big data problem. This is quoted as 'big data problem because it provides an interesting challenge for computer systems: data sets are generally quite large, taxing the capacities of main memory, local/remote disk'. *Big data is a high-volume, high-velocity and/or high-variety information asset requiring cost-effective and innovative forms of information processing for better insight, decision making, and process automation* (Gartner, 2012). Thus big data technologies use new tools to find relevant data and analyse its implications. There is no consensus on a unified definition of big data. Tom Davenport, co-founder of the International Institute for Analytics, predicted a shorter life for this technology due to this issue.

Some of the other definitions of big data include the following:

1. Oxford English Dictionary defines big data as 'data of a very large size, typically to the extent that its manipulation and management present significant logistical challenges'.
2. Wikipedia (2019) defines big data as 'a field that treats ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software'.
3. McKinsley in 2011 defines big data as 'datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyse.'
4. Roger Magoulas of O'reilly in 2005 defined big data as 'a massive volume of both structured and unstructured data that is difficult to store, analyse, process, share, visualize and manage with traditional database and software techniques.'
5. International Data Corporation (IDC) defines big data technologies as 'a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling high velocity capture, discovery, and/or analysis.'
6. Microsoft defined big data as 'the term increasingly used to describe the process of applying serious computing power—the latest in machine learning and artificial intelligence—to seriously massive and often highly complex sets of information.'
7. Big data is a term that describes hi-tech, high speed, high-volume, complex, and multivariate data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information.' (TechAmerica Foundation's Federal Big Data Commission, 2012).
8. Gartner in 2001 introduced 'the three V's—Volume, Velocity, and Variety' characterizing big data.

From the various definitions it can be summarized that big data stands out in dealing with large volume of heterogeneous and complex data that cannot be handled by traditional warehousing solutions. It deals

with effective organization and efficient extraction of valued information from the voluminous, heterogeneous, real-time data aggregated from multiple and autonomous sources, using statistical and machine learning techniques.

The characteristics of big data include the following:

Volume Volume is synonymous to the word 'big' in big data. It refers to the scale at which the data is growing. It is estimated that 2.5 exabytes of data is generated daily. IDC reports that data volume would reach 40 zettabytes by 2020 due to explosion of the Internet, scientific, and business applications.

Variety Big data encompasses structured data such as customer information, semi-structured data like XML, and unstructured data from documents, videos, images, sensors, smart devices, audios, social networks. The unstructured data grows 15 times faster than structured data. Only 10% of the data available globally is structured. Facebook generates 30+ petabytes of unstructured data in the form of web logs, pictures, and messages.

Velocity Big data systems deal with static and dynamic streaming data from business applications such as trading, telecom, wearable sensors in healthcare, banking, and the Internet. Around 90% of the data available was generated between 2015 and 2019. Facebook stores petabytes of streaming data of its 1.8 billion active monthly users to process their queries in real-time. Usage of smartphones for shopping and other activities has necessitated the service providers to deal with streaming data sources that demand real-time analytics.

Value Data received from sources in big data environment has a low value relative to its volume. However, high value can be inferred from such data using suitable analytical techniques. The objective of a data scientist is to extract additional value from data sources that can be static or dynamic, large or small, structured or unstructured, historic, transient, or externally available. This valuable information can provide business insights for better decision making. Added value can elevate data to be treated as an asset that can be traded. Big data systems can also extract value from fuzzy data using semantics. Infrastructure requirements of big data systems to store and analyse data is high leading to heavy investments. Hence inferring business value from big data includes estimating financial and infrastructure requirements, analysing streaming data, integration with legacy data and discovering new business opportunities. This helps to estimate the viability of the solution. Thus financial and technical value is also to be considered.

Veracity Introduced by IBM, this term represents dependability, reliability, and certainty of data. It deals with data quality. Accuracy and certainty of information is lost due to noise, latency, approximation, and ambiguity of voluminous amount of data analysed in a big data application. Veracity also implies the trustworthiness of the data and its insights. If the data is of a higher value, then it becomes more trustworthy. For example, customer sentiments, though valuable, are subjective and hence uncertain. Big data analytics should be capable of handling imprecise and uncertain data. Veracity can be maintained if datasets are cleaned and ranked to avoid making decisions on imprecise and uncertain data. Business leaders usually do not trust the data given for decision making. This is because poor quality data leads to wrong decisions and incurs heavy loss to the organization and the economy.

Validity This implies correctness and accuracy of data used for extracting useful information. It is a measure of data authenticity. Valid data leads to right decisions. Validity deals with data quality, governance, and master data management. To understand the impact of a cyclone in a locality, data from weather prediction systems can be used along with tweets from the locality. This improves the validity of the information.

Variability This term was introduced by Statistical Analysis System (SAS) to represent a dynamic evolving behaviour of a data source. Data velocity is not consistent and is characterized by periodic peaks and troughs. As big data sources are heterogeneous, technologies should enable connection, cleansing, and transforming such data to make it suitable for analytics. For example, in a news recommendation system, timeliness of the data is important. Election news varies based on the reporting time. Outdated news should be removed to obtain better value.

Visualization This facilitates better understanding of the value of the information drawn from data sources. The major objective of visualization is to communicate information clearly and effectively to its intended users. Charts and graphs are the common tools used to visualize large amounts of complex data. Interactive tools also help to dig deep into data to extract the required insight. Visualization should be targeted for its intended users.

Viscosity This measures data velocity relative to the time scale of events under consideration. It gives the time lag between the occurrence and usage of an event. There is resistance in integrating heterogeneous data sources that produce data at different speeds and the processing required to turn data into actionable insights. Viscosity is a measure of such resistance. Improved streaming services, flexible and adaptable integration bus and suitable event processing can handle this issue. Only highly viscous data is stored; less viscous data calls for quick action.

Virality It is a measure of the rate at which information disperses in a people-to-people network. It also measures the spread and sharing of data/messages in the nodes of a network for analysis. Indirectly, it measures the activities of the sender/receiver and the spread of data.

Volatility It is a measure of the rate of loss/stability of data. It refers to the time period for which the data is valid and should be stored. Beyond this, the data is no longer relevant. For example, information related to election news, cricket matches, and tennis tournaments are highly volatile.

Venue This deals with the source of data and the approaches to obtain it. For example, tweets obtained from Twitter may be used to recommend products to the customers.

Vocabulary It deals with data models and data structures used to organize, store, and analyse data. Different modelling approaches can be used for different problem domains. These can be validated using different techniques.

Versatility Big data is flexible enough to be used differently under different contexts. Big data is evolving to satisfy the needs of many organizations, researchers, and governments. It is used in various domains in different ways voluntarily. Hence it is versatile.

Different organizations add V's based on their perspective of big data.
Big data can be categorized based on the following parameters:

(a) **Data dimensions,** which are related to the characteristics of data handled by big data systems as discussed by the V's in the previous section.
(b) **Infrastructure** used including cloud, grid, and cluster environments.
(c) **Frameworks and tools** used to handle key-value, column-oriented, document, and graph data stores such as Hadoop, Spark, Google, Oracle, Graph, MongoDB, and so on.
(d) **File systems** used in these frameworks such as local, distributed, or high performance file systems.
(e) **Programming models** used such as MapReduce, graph query, thread/task level parallel models.
(f) **Analytics techniques** used including descriptive, diagnostic, predictive, prescriptive, or cognitive analytics on a variety of data including video, text, Web, audio, and sensor data.

INTERVIEW QUESTIONS

13. Define big data.
14. Why is viscosity relevant for big data?
15. What are the three basic V's of big data?
16. What type of infrastructure is suitable for big data?
17. Define veracity.
18. What do you mean when you say that big data should 'add to value'?
19. Where does big data come from?
20. How can businesses benefit from big data?

1.4 ARCHITECTURE OF BIG DATA SYSTEMS

1.4.1 Traditional Data Systems

Traditional data systems like RDBMS is three-layered, consisting of the physical, logical, and view layers.

Physical layer Being the lowest level of abstraction, this layer uses low-level complex data structures for storing data. This includes B+ trees, R-trees, and so on.

Logical layer This layer is extensively used by database administrators for database design, performance tuning, backup, and recovery. This layer maintains the metadata describing the type of the data and the relationships among the data.

View layer This layer offers tools for querying and processing data. It hides the details and complex organization of data.

While the physical layer is used for storing data, the logical layer is used to obtain an abstraction to query and process data using the view layer.

1.4.2 Core Layers of Big Data Systems

Big data systems follow a layered architecture with the following four basic layers (Fig. 1.3). Other layers offer common services to these layers.

Data storage layer As current applications have to work with large amount of heterogeneous real-time data, it becomes a necessity to handle the heterogeneity using different data stores. Such an approach to identify an effective data store for a particular data is called *polyglot persistence*. It harnesses the power of multiple databases at the same time. To store a large amount of unstructured data, Hadoop Distributed File System (HDFS) can be used. Simple Storage System (S3) can be used for object-based storage. The functionality of this layer is handled by the following two sub layers:

1. **Physical layer** that organizes data over distributed storage in a cluster of nodes connected by high speed networks. This layer is designed to handle large volume of heterogeneous real-time data.
2. **Data layer** is a logical layer that maintains data blocks and the global namespace to access data. This layer maintains tools to organize, access, and retrieve heterogeneous data including data maintained as key/value pairs, documents, or in relational, column-oriented, and semi-structured formats. This layer hosts a large variety of tools that can be used by data scientists, data engineers, and other users to organize and access heterogeneous data. It also performs data backup for large volumes of rapidly changing data.

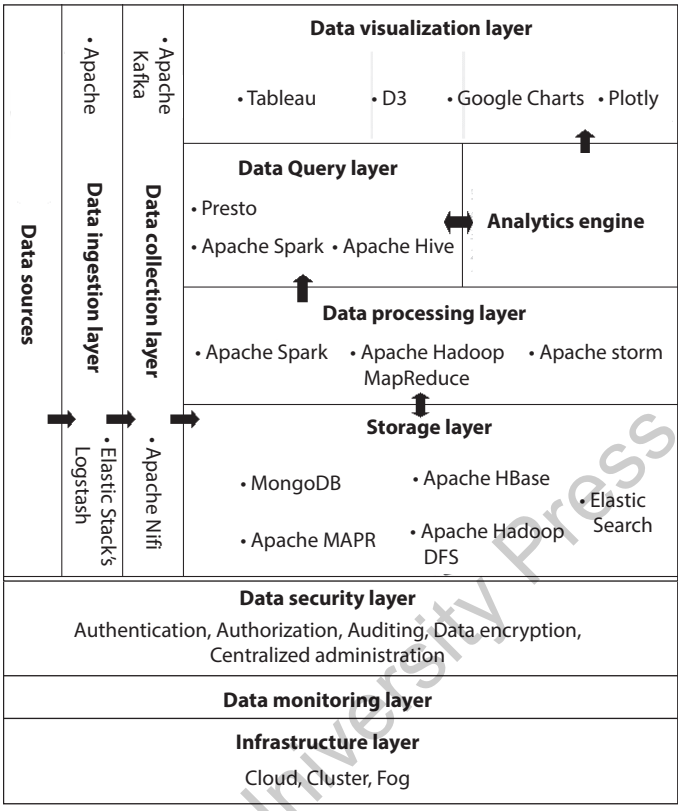

Fig. 1.3 Architecture of big data systems

Data processing layer Data collected in the storage layer is processed in this layer in batch or real-time modes.

1. Batch processing is used for offline analytics. Hadoop is a batch processing system that uses MapReduce programming technique.
2. Real-time processing is used for online analytics. Apache Storm processes streaming data in real time to make decisions based on the criticality of the event. Spark is a time-efficient, in-memory data processing engine that can execute streaming, machine learning, or SQL workloads iteratively. Flink can take care of out-of-order data also.

The data processing layer offers software tools such as MapReduce, Spark, Neo4j as well as tools for statistical modelling and machine learning. This layer also handles and maintains data replication and deduplication mechanisms suitable for a particular computation technology.

Data query layer This layer aims at obtaining data value or valuable insights from the processing layer. The following are some frameworks used for analytical querying:

1. **Hive:** This framework induces structure into the unstructured data stored in Hadoop, so that it can be queried using an SQL-like language called Hive-QL. Hive is used by data analysts to query, summarize, explore, and analyse unstructured data to obtain actionable business insight.

2. **Presto:** Used by Facebook, it is an open-source distributed SQL query engine capable of executing interactive analytic queries on heterogeneous data sources of different sizes.

Analytics engine extends the functionality of the data processing layer with domain-specific tools for decision making. Tools in this layer perform descriptive, diagnostic, predictive, prescriptive, or cognitive analytics.

Data visualization layer This layer presents the value of data to the users in an understandable format. Dashboards, graphs, and tables are some tools used for visualization. Google charts is a benchmark for visualizing huge data sets with ease. Tableau is a free visualization tool for big data. D3 is not a visualization tool, but is a programming tool as the user should be knowledgeable on JavaScript to visualize the collected data effectively.

1.4.3 Service Layers

The following layers offer common services to the core layers.

Data ingestion layer True to the saying, 'A good beginning makes a good end', the effectiveness of this layer determines the value of information extracted. Data coming from multiple sources is prioritized, validated, categorized, and routed to the destination for effective storage and access. Data may be ingested in batches at periodic time intervals or in real-time as it arrives. As real-time data is involved, ingestion also involves detecting the changed data. Ingestion of data from heterogeneous data sources is a challenging task. Further, evolution of data sources and consumer applications impacts the ingestion process. Some of the commonly used tools in the ingestion layer include the following:

1. **Flume** is a distributed, reliable, and highly available framework based on streaming flows to efficiently collect, aggregate, and move large volumes of real-time data. It can aggregate large volume of streaming data into Hadoop for storage and analysis. Flume can handle transient spikes and scales horizontally.
2. **Elastic Logstash** aggregates data from multiple sources to simultaneous processes and routes it to Elastic Search Storage.
3. **Sqoop** supports bulk data transfer between Hadoop and structured stores such as Oracle and MySQL.

Data collector layer The objective of this layer is to transport data from the ingestion layer to the rest of data pipeline. A messaging system is used for decoupled communication between senders and receivers. Kafka is a message-oriented middleware used for data collection. Kafka collaborates with Storm, HBase and Spark for real-time analysis of streaming data.

Data security layer This layer provides authentication, authorization, audit, data encryption, and centralized administration for big data systems. In Hadoop stack, Knox can be deployed centrally to provide restricted access to Hadoop through a firewall. Kerberos can be used to provide a strong authentication. Ranger can be used for authorization, auditing, and centralized administration. HDFS encryption can be used for data secrecy.

Data monitoring layer It includes tools for monitoring the performance at infrastructure, framework, analytics engine, data store, and application levels.

Infrastructure layer This layer provides the hardware to host various big data frameworks in cloud infrastructure that is highly scalable and preferable. This architecture can also be layered in the fog environment to perform edge analytics.

1.4.4 Differences between Traditional and Big Data Systems

Traditional data systems differ from big data systems in their infrastructure, framework, and processing capabilities in the following ways:

1. Traditional databases were designed to handle transactions for operational and historical data using data warehousing tools, query languages, and online transaction/analytical processing tools. Big data systems support consistency, availability, and partitioning of large data sources for effective decision making using suitable analytical tools that support intelligence.
2. Traditional database frameworks are designed to handle structured data like relational data. The file system can be centralized or distributed. Big data frameworks are designed to handle data in large scale. They should be capable of handling voluminous amount of heterogenous real-time data from multiple sources.
3. The infrastructure required for handling traditional databases is small/medium scale with centralized control. Big data systems require massively distributed, scalable large scale infrastructure on commodity hardware.

In traditional systems, data warehouse is segregated from the operational database that uses online transaction processing (OLTP). Traditional databases are designed to provide business intelligence. However, they are incapable of handling large volumes of rapidly changing data in business or scientific applications. Processing in data warehouses include

1. online analytical processing operations such as slice-and-dice, drill down, drill up, and pivoting;
2. data mining techniques such as pre-processing, modelling, classification, and prediction for knowledge discovery.

Table 1.1 provides a comparison of traditional and big data systems.

Table 1.1 Traditional vs big data systems

S. no.	Property	Traditional database systems	Big data systems
1.	Volume	Data is segregated as operational and historical data and handled separately. If the volume of historical data is large, filtering is used. Extract transform load (ETL) operations are used to extract information	Capable of handling large volume of operational and historical data simultaneously from various sources
2.	Velocity	Transaction orientation limits data velocity	Real-time data is obtained from various sources like the Web, sensors, devices
3.	Variety/ Heterogeneous data formats	Semi-structured/structured data like XML and relational data	Structured data like relational, semi-structured data like XML, unstructured data like text, video streaming
4.	Languages	Query languages like SQL	NoSQL query/programming languages like MapReduce, Neo4j, Hive, document querying
5.	Platforms	OLTP, Relational database management systems (RDBMS)	Decision support tools with machine learning, statistical modelling for text, video, image analytics, graph analytics, i-memory analytics, statistics/predictive analytics

Contd

Table 1.1 (Contd)

S. no.	Property	Traditional database systems	Big data systems
6.	Data handling	Data distribution is usually centrally controlled and maintained in a structured data format	Data is distributed over multiple storage/computer nodes in multiple data formats
7.	Infrastructure	Centralized architecture with less scalability	Scale out infrastructure for efficient storage and processing
8.	Workloads	Data is usually static; operational and analytical workloads are handled separately	Handles both batch and stream processing efficiently. Operational big data workloads are easier to manage and implement using NoSQL systems like MongoDB. Complex analytical workloads can be analysed using MapReduce
9.	Backup	Implemented using already established mechanisms using replication	Owing to large volume of data, differential backup mechanisms are preferred
10.	Data recovery	Using replication	Replication depends on criticality of data
11.	Theorem	CAP theorem with ACID properties	CAP theorem with BASE properties
12.	Stakeholders	Administrators, developers, end-users	Data scientists, data analysts, data engineers, end-users

INTERVIEW QUESTIONS

21. What is the functionality of physical, logical, and view layers in traditional systems?
22. What is the theorem on which big data frameworks are based upon?
23. What are the commonly used tools to manage operational and analytical workloads in big data systems?
24. What are the problems in traditional databases?
25. What is Kafka?
26. List some tools for aggregation.

1.5 ROLES IN DATA SCIENCE TEAM

The members of a data science team play various roles to provide solutions for state-of-the-art information extraction problems (see Fig. 1.4). These roles are listed here:

Data scientist A data scientist designs and conducts experiments by accumulating and cleansing raw data, analysing it using suitable statistical models, inferring insights, and visualizing and communicating the results clearly to the intended audience. Data scientists are knowledgeable in applied statistics, algorithms, and machine learning. They possess good data analysis and communication skills. They are capable of modifying existing approaches or designing new techniques for data analysis. Building a model to predict the number of fraudulent customers in the vehicle insurance domain is the job of a data scientist. A data

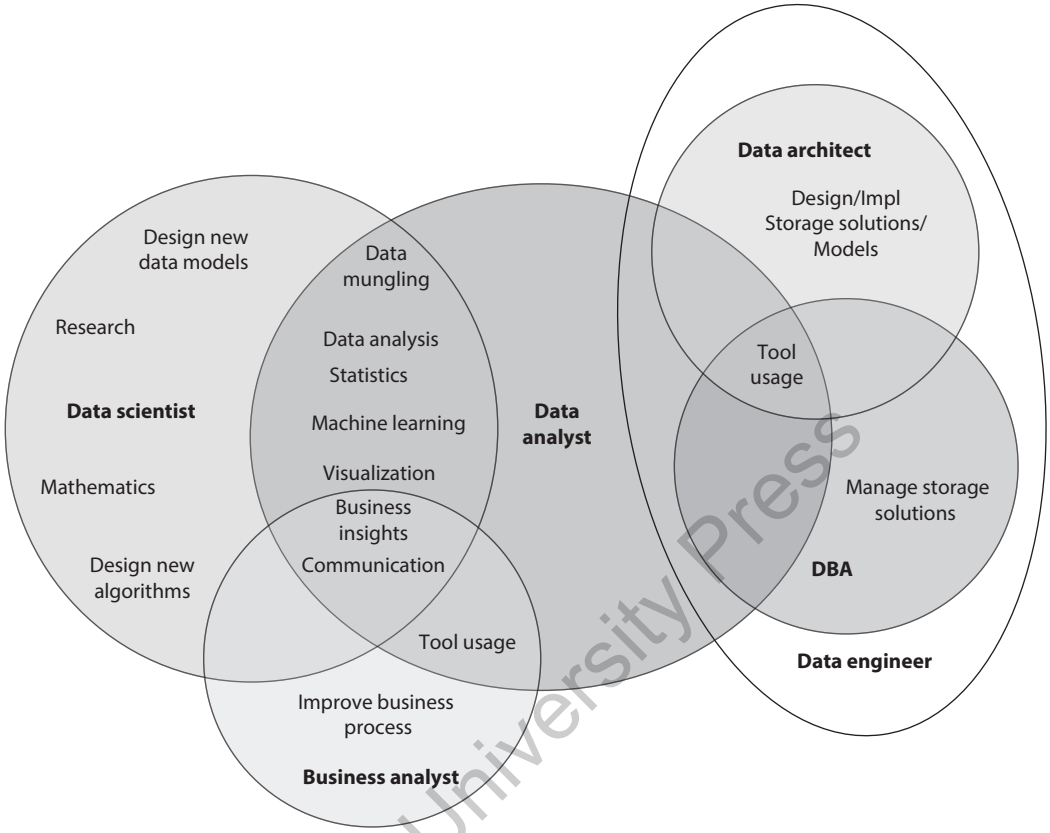

Fig. 1.4 Roles in data science team

scientist performs the role of a statistician who models data sets and a data manager who communicates the business insights with business leaders. They have sound knowledge on the theory and implementation techniques behind data science and the business implications of data. It is not necessary that a data scientist should be highly proficient in handling data analytics tools.

Statistician A statistician collects, analyses, and interprets qualitative and quantitative data using statistical theories and methods to obtain knowledge that can be used to transform businesses. Modern statisticians are also proficient in using tools.

Data analyst A data analyst collects, processes, and analyses data just like a data scientist. Thus a data analyst is knowledgeable in data munging (transformation of raw data into a format fit for processing), programming, analysing using statistical and machine learning techniques, using tools, and visualizing data. A good analyst obtains valuable business insights from the process of analysis and effectively communicates the results to business leaders. He/She is also proficient in using tools to solve problems. A data analyst differs from a data scientist in not possessing mathematical or research background to invent new algorithms. Data analysts typically require the following:

Programming skills Knowledge of programming languages such as R and Python are extremely important for any data analyst.

Statistical skills and mathematics Descriptive and inferential statistics and experimental designs are also a must for data analysts.

Machine learning skills These skills are essential to design and develop systems capable to update and refine themselves by discovering new knowledge.

Data wrangling skills This refers to the ability to map raw data and convert it into another format that allows for a more convenient consumption of the data.

Communication and data visualization skills These skills communicate the results to the intended users in an understandable format.

Business analyst This person is a data analyst interested in improving the business from the insights obtained using data analysis. Business analysts link the knowledge mined with actionable business insights and communicate the same to the business leaders.

Data architect A data architect creates a blue print for the data management system, so that data from disparate data sources can be effectively captured, integrated, organized, centralized, protected, and maintained. The data architect has an extensive knowledge on data modelling, database architecture and architectural patterns, data warehousing architecture, and database management tools. An architect plays a major role in the big data era by creating trustworthy systems capable of handling huge volume of heterogeneous real-time data. A data architect also suggests suitable approaches for data acquisition, recovery, and maintenance of data stores.

Database administrator (DBA) He/She manages and maintains the database so that suitable database access is offered to relevant users. A DBA ensures data availability by implementing suitable backup and recovery mechanisms.

Data engineer A data engineer uses software engineering skills to develop, construct, test, and maintain scalable frameworks for data storage and processing. A data engineer obtains the model from the data scientist and implements it using suitable languages or tools. Data architects deal with designing solutions, patterns, and frameworks to manage data models. A DBA manages data storage solutions. Data engineers possess the skills of data architects and database administrators. Data engineers implement and manage data models and data warehousing solutions. They are proficient in using extract transform load (ETL) tools.

Data science manager A data science manager plays the role of a project leader capable of building and managing the data science team by setting goals and priorities. He/She possesses both analytical and managerial skills such as leadership and communication skills.

INTERVIEW QUESTIONS

27. What is the role of a data scientist?
28. What are the skills required by a data scientist?
29. What is the role of a data engineer?
30. Differentiate the roles of data scientist, data engineer, data architect, and DBA.
31. What is the role of a data architect in a data science team?
32. What are the skills required by a data analyst?

1.6 BIG DATA USE CASES

Common use cases of big data include the following:

Sports Domain

Big data systems are extensively used in the sports domain to provide real-time analytics.

1. The movement of the players during their practice sessions can be tracked and analysed by the coach to improve on their performance. This approach was used by Atlanta Falcon.
2. Toronto Raptors collect large volume of data on a player's movements and styles. This data is analysed to provide personalized coaching and to improve the team performance.
3. Riot Games analyse 4TB of operational logs and 500GB of structured data to predict performance of a game.
4. Nike uses big data analytics for eco-friendly product design. For example, identifying an environment-friendly dyeing technique without using water.
5. Internet games are usually hosted using big data platforms on a cloud infrastructure.

Sentiment Analysis

The major challenge faced by any business is maintenance of customer satisfaction. Sentimental analysis can be used to predict changing customer interests, identify potential customers, forecast the demand and price of products. Such an analysis can be done using data from social media like Facebook and product reviews given by customers. The companies can then respond immediately to improve customer satisfaction. Such analytics can help the companies to connect with their customers effectively to outperform their competitors.

1. Delta Airlines collects tweets of their flyers to track their in-flight experience, flight schedules, hospitality, and so on. Sentimental analysis of the flyer tweets can help the support team to enhance their customer satisfaction by providing suitable guidance to their flyers.
2. Analysis of the sentiments of customers gathered from tweets can be used by retailers like Macy to identify their preferences based on demographics, pricing, seasons, and geography. This helps to make sensible and profitable advertisements.
3. Salesforce uses Radian6 to analyse social media conversations on the credibility of a company and its products by aggregating positive, negative and neutral sentiments.

Behavioural Analytics

Organizations harness the power of big data to understand customer behaviour and add value to their business. Amazon's product recommendation system recommends apt product to its customers based on their previous purchase history and purchase made by related users.

1. Cash back offers are given by a number of banks like Bank of America based on purchase histories of its credit/debit card customers.
2. 'Retaining an old customer is more profitable than obtaining a new customer.' Banks and retail stores can study the interests of its customer and obtain insights that can suggest the right action/offer/advertisement for the right customer.
3. Companies like Target leverage big data analytics on the purchase of certain products to predict life changes of their customers and make suitable promotions. For example, if pregnancy is predicted, Target advertises on baby-related products.

4. A large-scale retailer, Nordstorm monitors their customer shopping patterns related to their items of interest and shopping time to advertise suitably to its customers and provide a personalized shopping experience.
5. McDonald's uses the power of big data analytics to study ordering patterns, size of orders, and waiting time to optimize operations at its outlets and enhance its customer's experience.
6. Kohl's retail chain tracks the customer's browsing history to provide offers on items of interest to the customer. Such offers increase the probability of purchase of the item by the customer.

Customer Segmentation

This is nothing but grouping similar users based on their purchases and recommending suitable items for them based on their personal or group interests.

1. Personalized advertising leverages big data analytics to combine heterogeneous data sets including local and demographic data of the customers to gain insights on their interests, preferences, beliefs, and environment. Tim Warner uses such a strategy for targeted advertisements.
2. Amazon identifies a list of similar customers by analysing the browsing and purchase patterns of its customers and related items. Thus, Amazon recommends items to its customer not only based on his/her personal interest, but also based on the interest of similar groups of users and related items (like iPhone and power bank).
3. Pandora provides music recommendations based on the static profile, user interest, related songs, demographics, and location of its users. Netflix uses collaborative filtering algorithms to recommend movies to its users.
4. Customer behaviour patterns and transactions can be used for personalizing experiences using analytics and cognitive computing.
5. Analysing a patient's historical records, healthcare organizations improve on the effectiveness of their service. The popularity of wearable devices to monitor fitness activity, sleep patterns, and calorie intake help physicians and health insurance companies to predict health outcomes and behaviours. This helps in personalizing healthcare.

Prediction

One of the major analytics technique used by the analytics engine of big data systems is prediction of the outcomes based on historical information.

1. Purdue University tracks the performance of its students in various courses to predict their pitfalls and alert them in advance based on their performance in similar courses. Such alerts help to improve the success rates of the students.
2. Volkswagen improves the revenue of its service centres by analysing personal information of its customers along with vehicle data and information from the technical support team.
3. Earth observation systems use remote-sensing technologies to obtain necessary information required for weather forecasting, predict natural disasters and climatic changes.
4. Breast cancer data sets and patient details over a decade is used by Ayasdi to predict the relationships between leukaemia and breast cancer that can aid in novel cancer therapies.
5. Spread of diseases such as H1N1 and dengue can be predicted by healthcare organizations using location based information, search patterns on the Internet and social media.
6. The stock exchange data holds information on share transactions made by a large number of customers. This information can be analysed to provide stock recommendations to customers and to predict the stock prices.

7. Banks can predict loan eligibility by analysing the personal details of the applicant in real-time and target suitable customers.
8. Transportation department can use real-time traffic information to model and predict traffic patterns to handle traffic signals, public transport systems, traffic redirection, and so on. Smart city applications use sensor data to control various utilities. Power grid maintains electricity consumed by a particular node. This information can be used for better power distribution.

Fraud Detection

Big data systems help organizations to detect, prevent, and eliminate internal and external frauds. Unusual usage pattern of a debit or credit card can alert a bank of stolen card.

1. VISA uses big data analytics to detect frauds and implement mechanisms to overcome such vulnerabilities.
2. Intelligent Bureau of Canada (IBC) identifies fraudulent claims by analysing over 2 lakhs claims made in the past 6 years.
3. Analysing emails, phone calls, and transaction data frauds can be detected. This technique was used by JPMorgan Chase. Internal frauds can be identified by tracking employee communications.
4. Operational and historical credit card transactions made by billions of customers can be used to identify fraudulent transactions.
5. Black box of an aircraft registers heterogeneous information including audio (voice, sounds) and performance of various aircraft devices continuously. Analysing this data can draw insights when sabotage occurs.

1.6.1 Personalized Healthcare

Personalized healthcare (Fig. 1.5) can be a reality only if suitable platforms are available to store, process, analyse, and visualize large volume of heterogeneous real-time data. Since traditional technology is not equipped to this end, big data and semantic web related technologies come into the play.

From an architectural perspective, Hadoop can be integrated with existing data systems to complement their functionalities. Some of the functional areas offered by Enterprise Hadoop are Data Management, Data Access, Data Governance and Integration, Security and Operations.

Electronic Health Records (EHRs) Genomics, and Medical Imaging data are stored into the Hadoop Distributed File System (HDFS). Then Pig and Hive tools are used to clean and prepare data. Identifying cohorts (similar patients) is a herculean task from millions of EHRs. For example, to predict the survival rate of heart failure patients, data from sources such as lab results, medications, and patient demographics should be analysed and cohorts should be created by analysing ejection rate of heart failure patients. For this, clinical trials of heart failure patients should be analysed and EHR of patients should be studied. Pig queries are used to filter patients with heart failure in level 1 (descriptive), followed by patients with ejection rate > threshold in level 2 and clinical trials are analysed with the knowledge rules given by experts (diagnostic). Based on this, personalized treatments can be predicted for a patient (predictive). Then, suitable actions can be taken to offer this treatment (prescriptive).

The data processing layer extracts the big data driven phenotype. The analytics layer uses the following:

1. Descriptive analytics to evaluate various statistics and visualize them using charts.
2. Diagnostic analytics using survival analysis and regression to correlate survival rate of patients with heart failure.

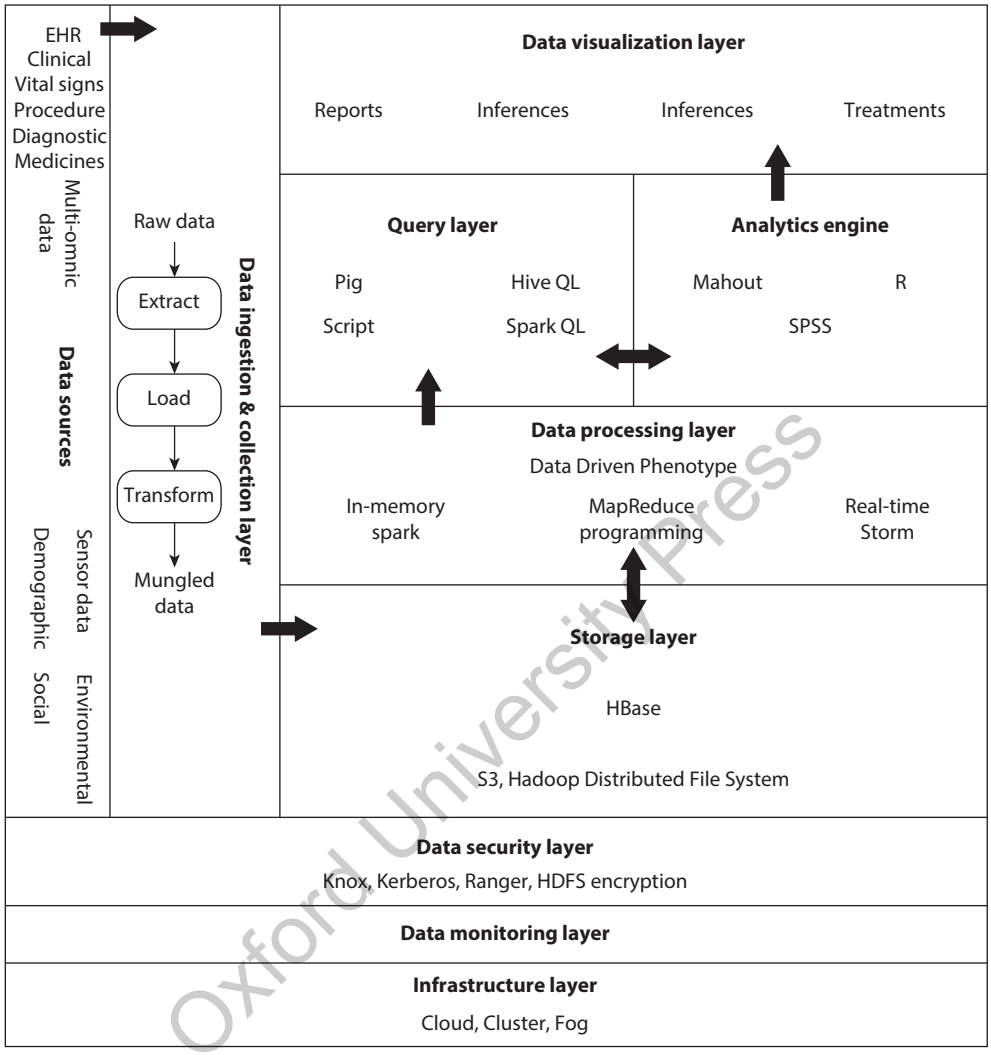

Fig. 1.5 Big data architecture for personalized healthcare

3. Predictive analytics using classification, clustering, and inferential analysis to predict survival rate for a new patient.
4. Prescriptive analytics for treatment plan and decision support.

INTERVIEW QUESTIONS

33. How can big data be used in the sports domain?
34. What are the common tools used in the data security layer?
35. List a use case of big data in fraud detection.

1.7 ADVANTAGES OF BIG DATA

Based on the characteristics of big data, the following are some of the advantages of big data systems:

1. Analysing big data can help in understanding the behaviour of customers and optimizing/personalizing the business processes to their needs.
2. Big data can help in scientific research. For example a particle accelerator like the Large Hadron Collider can produce more than 6MB of data/second. Identifying which data to ingest, store, and analyse requires a big data platform.
3. It can be used in medical field by
   (a) storing, maintaining, and distributing patient health records between multiple healthcare organizations. This paves the way for federated healthcare services for patients.
   (b) providing personalized medical services by decoding DNA strings along with clinical and behavioural data to provide a better cure. Big data platforms can store and analyse heterogeneous information to provide results to a query. For example, in healthcare, patients gene data, images (CT etc.), along with clinical data can be analysed to give suitable diagnostic results.
4. It can aid in ingesting large amounts of sensor data, maintaining them, and analysing them in real time.
   (a) Can monitor real-time data and identify natural calamities or man-made disasters using information from various types of sensors
   (b) Data from smart meters can be used to optimize energy grids
   (c) Help for safe and autonomous driving by using vision, traffic data, and sensors
5. Law enforcement by accessing vast amounts of stored information from multiple sources to provide useful analytics. For example, to identify criminal activities, travel data (flight records) can be used along with social media activities.
6. It can be used in educational applications including online course delivery and monitoring.
7. Along with suitable data integration strategies, big data can be de-duplicated and stored by removing all redundant data and thus help in reducing storage space.

1.8 CHALLENGES FACED BY BIG DATA SYSTEMS

Some of the challenges faced by big data systems include

1. Big data systems have to manage a large amount of heterogeneous real-time data. Traditional database systems cannot be used. Cluster processing frameworks are required for processing this data in parallel.
2. Big data systems depend on the availability of a large amount of historic data for analytics. Hence it needs to be analysed for longer durations to leverage the benefits.
3. Big data is unstructured and schema-less; hence it cannot be stored and managed in traditional database systems like relational systems.
4. Big data analysis may be misleading if sufficient amount of data is not available, quality of data is not good, or data is skewed. Data validation and pre-processing also cannot be standardized in big data systems.
5. Speedy updates can lead to skewness in data if redundancy is not removed.
6. Big data systems are costlier when compared to traditional systems. Platforms and tools used should be carefully selected to minimize cost and for proper functionality. Generally such tools require more training.
7. Big data analysis violates data privacy of customers as customer information related to their purchase patterns and transactions can be easily used to the advantage of business organizations.

8. The results of the analysis should be maintained with privacy as there is a chance that they may be misused against a particular person, group, or nation.
9. Big data has to be partially processed near its place of collection/storage to avoid the communication cost involved in the transfer of a large amount of data.

SUMMARY

Big data and data science play a major role in the organizations involved in information extraction. Big data technology is used by organizations to obtain useful insights from available data to improve on their businesses. This chapter starts by introducing the readers on the need for big data technology. Data science is an all-encompassing discipline that includes big data. Various terminologies under data science related to big data are explained in this chapter. This is followed by a definition of big data along with its characteristics. Traditional database architectures cannot be used to store and analyse large amounts of data and there is a need for big data architecture. This chapter then illustrates the components of big data systems. Unlike traditional systems, big data systems require additional expertise for storage and management. The members of the team who manage a big data project possess specific roles. The roles played by specific members of the team is illustrated. Globally big data technology has been adopted by a number of industries. Some such use cases are discussed. These use cases are also applicable to Indian organizations. Finally, the architecture of the healthcare system using big data technology is illustrated. The chapter ends with a note on the advantages and challenges faced by big data. It thus introduces its readers to the basic concepts in big data technology and helps them gain an understanding of the need and usage of big data technology by various organizations.

DEFINITIONS

- The hype cycle provides a conceptual representation of the maturity of emerging technologies.
- Data analysis primarily deals with analysing past data and understanding the data. This leads to knowledge discovery. Data analytics deals with using this knowledge to make smart business decisions in the future.
- Statistics is basically a measure for an attribute(s) of a sample.
- Data mining deals with designing algorithms to extract insights from data. It includes pattern recognition, feature selection, clustering, supervised classification, and some statistical techniques.
- When analysis of data leads to describing patterns, it is called descriptive analytics.
- When knowledge discovery helps to understand the reason behind the occurrence of patterns, it is called diagnostic analytics.
- When knowledge discovered is used to predict future trends, it is called predictive analytics.
- When knowledge discovered can be used to suggest actions to be taken in future, it is called prescriptive analytics.
- When analysis of knowledge discovered identifies methodologies to suggest future actions, it is called cognitive analytics.
- Data engineering is a sub-field of software architecture basically dealing with hardware and frameworks to maintain data that is consumed by data scientists. Data engineers work in suggesting and implementing architectures to store, organize, and process data for different kinds of applications.
- Business intelligence deals with collecting useful data, analysing it, and visualizing it by creating data reports to extract valuable business insights.

- Big data is a high-volume, high-velocity, and/or high-variety information asset requiring cost-effective, innovative forms of information processing for better insight, decision making, and process automation.
- Data science is the science of making sense out of data. It is related to cleansing, preparation, and analysis of structured/unstructured data to extract information from data.
- A data scientist designs and conducts experiments by accumulating and cleansing raw data, analysing it using suitable statistical models, inferring insights, visualizing, and communicating the results clearly to the intended audience.
- A statistician collects, analyses, and interprets qualitative and quantitative data using statistical theories and methods to obtain knowledge that can be used to transform businesses.
- A data analyst differs from a data scientist in not possessing mathematical or research background to invent new algorithms.
- A business analyst is a data analyst interested in improving the business from the insights obtained using data analysis.
- A data architect creates a blue print for the data management system, so that data from disparate data sources can be effectively captured, integrated, organized, centralized, protected, and maintained.
- A database administrator (DBA) manages and maintains the database so that suitable database access is offered to relevant users.
- A data engineer uses software engineering skills to develop, construct, test, and maintain scalable frameworks for data storage and processing.
- A data science manager plays the role of a project leader capable of building and managing the data science team by setting goals and priorities.

MULTIPLE CHOICE QUESTIONS

1. _____ is the heart of data science.
   a) Analytics b) Machine learning
   c) Statistics d) Mining
2. _____ is a subset of statistics used for the insurance domain.
   a) Actuarial science b) Econometrics
   c) Operations Research d) Prediction
3. While _____ is estimation based, _____ deals with automated learning.
   a) statistics, machine learning
   b) statistics, AI
   d) AI, statistics
   d) statistics, analytics
4. When knowledge discovered can be used to suggest actions to be taken in future, it is known as _____ analytics.
   a) predictive b) cognitive
   c) diagnostic d) prescriptive
5. _____ aims at applying computer science to a particular domain. _____ applies both mathematical techniques and computer science to an application.
   a) AI, Machine learning
   b) Data mining, Big data analytics
   c) Machine learning, BI
   d) BI, Big data analytics
6. _____ deals with trustworthiness and data quality.
   a) Validity b) Variability
   c) Veracity d) Visualization
7. _____ measures data velocity relative to time scale of events.
   a) Virality b) Variability
   c) Viscosity d) Value
8. _____ deals with data models, data structures used to organize, store, and analyse data.

a) Vocabulary
b) Versatility
c) Variability
d) Value

9. While _____ layer is used for storing data, the _____ layer is used to obtain an abstraction to query and process data.
   a) physical, logical
   b) logical, view
   c) physical, view
   d) ingestion, logical

10. The _____ layer aims at obtaining data value or valuable insights from the processing layer.
    a) physical
    b) collector
    c) ingestion
    d) query

11. _____ support consistency, availability, and partitioning of large data sources.
    a) Big data systems
    b) Three-tier system
    c) RDBMS
    d) Networked systems

12. _____ can be used to predict changing customer interests, identify potential customers, forecast the demand and price of products.
    a) Sentiment analysis
    b) Behavioural analysis
    c) Customer segmentation
    d) Customer analysis

13. _____ is used in Amazon's product recommendation system.
    a) Sentiment analysis
    b) Behavioural analysis
    c) Customer segmentation
    d) Customer analysis

14. _____ analytics using survival analysis and regression correlates survival rate of patients with heart failure.
    a) Predictive
    b) Cognitive
    c) Diagnostic
    d) Prescriptive

REVIEW QUESTIONS

1. What is a hype cycle? What is its significance?
2. Elaborate on the current technology trends in knowledge extraction.
3. What are the types of analytics?
4. Compare and contrast the following:
   (a) Data science and machine learning
   (b) Data analytics and data analysis
   (c) Big data and business intelligence
   (d) Machine learning and statistics
   (e) Data science and data engineering
   (f) Big data and data mining
   (g) Business analyst and data analyst
   (h) Data architect and data engineer
   (i) Traditional and big data systems
5. What is data science? Elaborate on its related disciplines and their correlations.
6. Define big data.
7. What are the characteristics of big data? Describe by giving suitable examples.
8. Describe in detail the various roles in a data science team.
9. Discuss the architecture of big data solution for the financial domain.
10. Provide suitable use cases for big data in the financial domain.

CRITICAL THINKING QUESTIONS

1. The city's transportation department is interested in studying the relationship between the temperature and the number of passengers that ride the main bus line, in order to better serve their customers. Suggest a suitable environment for the same. (Hint: Section 1.5)
2. Big data has reduced the cost of treatment in healthcare industry. Justify. (Hint: Section 1.3)

3. Can data ingested and used in applications like Uber be used for traffic monitoring? (Hint: Section 1.3)
4. Can big data platforms be used to identify availability of usable water around the world? How? (Hint: Section 1.3)
5. Identify the type of analytics in the following problems. (Hint: Section 1.2)
   (a) Consider a hotel chain, what type of analytics can be used to identify whether a hotel in this chain has met its target sales?
   (b) Let us assume that we know various parameters in a desert region like temperature pressure, etc. that are related to the occurrence of a storm (cause a storm). What type of analytics can be used to prevent a storm in the desert?
   (c) Identify the type of analytics that can predict a natural calamity and identify strategy to move people to safety.

ANSWERS TO MULTIPLE CHOICE QUESTIONS

1. (a) **2.** (a) **3.** (a) **4.** (d) **5.** (b) **6.** (c) **7.** (c) **8.** (a) **9.** (a) **10.** (d)
11. (a) **12.** (a) **13.** (b) **14.** (c)