Big Data, Cloud and Analytics

Block

2

CLOUD COMPUTING AND BIG DATA TECHNOLOGIES

Editorial Team

Prof. R. Prasad
IFHE (Deemed-to-be-University), Hyderabad

Dr. Sanjay Fuloria
IFHE (Deemed-to-be-University), Hyderabad

Dr. Sindhuja
IFHE (Deemed-to-be-University), Hyderabad

Dr. Nasina Jigeesh
IFHE (Deemed-to-be-University), Hyderabad

Content Development Team

Prof. Anirudh Prakhya
IFHE (Deemed-to-be-University), Hyderabad

Dr. Y. V. Subrahmanyam
IFHE(Deemed-to-be-University), Hyderabad

Prof. R. Muthukumar
IFHE (Deemed-to-be-University), Hyderabad

Prof. V. Srinivasa Murthy
IFHE (Deemed-to-be-University), Hyderabad

Dr. K. Veena
IFHE (Deemed-to-be-University), Hyderabad

Prof. KV Murali Mohan
IFHE (Deemed-to-be-University), Hyderabad

Prof. M. Venkata Dharma Kumar
IFHE (Deemed-to-be-University), Hyderabad

Proofreading, Language Editing and Layout Team

Prof. Jayashree Murthy
IFHE (Deemed-to-be-University), Hyderabad

Mr. K. Venkateswarlu
IFHE (Deemed-to-be-University), Hyderabad

Ms. C. Sridevi
IFHE (Deemed-to-be-University), Hyderabad

Ref. No. BDCA-SLM-IFHE – 112022B2

For any clarification regarding this book, the students may please write to The ICFAI Foundation for Higher Education (IFHE), Hyderabad specifying the unit and page number.

While every possible care has been taken in type-setting and printing this book, The ICFAI Foundation for Higher Education (IFHE), Hyderabad welcomes suggestions from students for improvement in future editions.

Our E-mail id: cwfeedback@icfaiuniversity.in

Centre for Distance and Online Education (CDOE)
The ICFAI Foundation for Higher Education
(Deemed-to-be-University Under Section 3 of UGC Act, 1956)
Donthanapally, Shankarapalli Road, Hyderabad- 501203

BLOCK 2: CLOUD COMPUTING AND BIG DATA TECHNOLOGIES

The learner needs to understand the handling of structured and unstructured data using analytics and move to decision making using big data and allied information management. We got introduced to various concepts of big data like Hadoop as well as the generation aspects of big data and its applications in healthcare and marketing. It's time now, to know about big data technologies. This involves understanding various methods to use big data for decision making in business. The focus will then shift to how to store, handle, analyze and use unstructured data and finally to big data in information management at organizations.

Unit 5 – The learner gets to know the technology behind Hadoop (the open source technology) and the people working behind it. He/she also learns about the relation between cloud and big data and how to put them to use in business applications. *Big Data and Cloud Technologies* deals with the big data concept, Hadoop, the existence of a parallel world. Old versus new approaches of data discovery, following the way people work, open source technology for big data and the relation between cloud and big data.

Unit 6 – In addition to the above technologies, the learner needs to understand business intelligence and predictive analysis, mobile business intelligence and crowdsourcing concepts for business decision making. *Big Data Technologies and Terminologies* deals with predictive analytics used as a service, business intelligence, mobile BI in the main perspective, crowdsourcing, analytics, inter and trans-firewall analytics, R&D approach, adoption of new technology, big data technology terms and data size. By this time the learner will have a fair understanding of handling various big data activities at business.

Unit 7 – The learner needs to get a grasp of challenges in big data storage, dashboards using big data, and cloud computing in big data management for business decision making. *Cloud Computing and Big Data Management for Decision Making* covers overview, challenges in analyzing and managing, big data storage considerations, from operational dashboards to evidence-based decision making, cloud computing in big data management.

Unit 8 – Big data generates varieties of unstructured data and the learner has to get the skills to use this unstructured data for analyzing and utilizing various analytics to get the best benefit out of unstructured data. *Handling Unstructured Data* spans over various formats of unstructured data- text, audio, images, office software, web pages and video; issues in storing unstructured data, usage patterns, different ways of analyzing and managing unstructured data, comprehensive file archiving solution, opportunities and problems with unstructured data. By this time, complete confidence will be gained by the learner about handling big data.

Unit 9 – It is time now to get hold of platforms handling big data and analytics, limitations on computations and evolving technologies related to working with big data for the management of information within the organization. *Information Management* covers the foundation of big data, computing platforms that can handle big data & analytics, computation and storage of big data, computational limitations and emerging technologies related to big data.

Unit 5

Big Data and Cloud Technologies

Structure

> *"If someone asks me what cloud computing is, I try not to get bogged down with definitions. I tell them that, simply put, cloud computing is a better way to run your business."*
>
> \- Marc Benioff, Chair, Co-CEO, and Co-Founder, Salesforce

5.1 Introduction

Definitions apart, cloud computing provides the resources to analyse ever growing Big Data to run business more efficiently with customer satisfaction at the core.

In the previous unit, we shared advantages and disadvantages of using big data in health care. The data collected from business transactions require tools for analysis, validation, visualization, and management. Big data computing is a tool used for multi-dimensional information mining and business analytics. We expect a 400-fold increase in digital data from 2012 to 2020. Big data technologies lead to large-scale storage, faster computing at a lower price and also the preservation of large volumes of data by providing high availability of storage. The present unit discusses big data technology, Hadoop frame work, open source technologies and benefits of cloud technology.

5.2 Objectives

At the end of the unit, you will be able to -

- Describe the four dimensions of big data and how these dimensions are supported by technology
- Discuss about Hadoop framework, Hadoop architecture and hadoop distributed file system

- Differentiate old and new approaches of data processing
- Define open-source technologies for big data analytics
- Discuss how the cloud provides services for big data management

5.3 Big Data Dimensions

Technology is changing every day, leading to changes in how data is produced, processed and analyzed. New technologies are helping in evolving new data sources to capture more and more data. Technology helps in processing this data quickly and efficiently.

Big data technologies have been making an impact on new technologies for handling big data. Big data is characterized as possessing four dimensions called the 4Vs, they are-

- Volume
- Velocity
- Variety
- Veracity

The fifth V (Value/Valor) is also used of late, which denotes the quality of the data.

Volume

Volume means the amount of data, i.e. the volume of data that is growing. The task of big data is to convert such voluminous data into valuable information. Hadoop is an open-source software that helps in the fast storage and retrieval of big data. Here, the data is stored in commodity servers and is run in clusters.

According to the IDC report, the volume of data will reach 40 Zeta bytes by 2020. The rate of increase at present is 400 times. The volume of data is growing rapidly due to several applications like business, social, web and scientific explorations. Examples include Twitter data feeds, Facebook news feeds, clickstreams on a web page and mobile applications.

Velocity

Velocity means the rate at which data is increasing. The highest velocity data streams directly into memory. The velocity depends on the growing speed of business applications such as trading, transaction of telecom and banking. Another reason for the speed is growing number of internet connections. Mobile applications too have a large user population and high network traffic.

Variety

It refers to different data types such as structured, like relational databases, semi-structure like XML and unstructured, like audio, video and text.

Veracity

Veracity means uncertainty or inaccuracy of data. In many cases, the data will be inaccurate and needs a lot of statistical and analytical processes for data cleansing. After this process, the data is used for decision making.

5.3.1 Hadoop's Parallel World

Apache Hadoop is an open-source framework that allows distributed processing of large databases using simple programming written in Java. It provides distributed storage across clusters of computers.

Hadoop was developed by Doug Cutting and Mike Cafarella in 2005. Doug named it after his son's toy elephant and now it has become the trademark for Apache Software Foundation.

Functionalities of Hadoop:

- Handles data warehousing including search, log processing and workloads.
- Stores all kinds of data and performs a variety of analyses and transformations on the data.
- Stores terabytes and even petabytes of data.
- Handles system failure automatically without losing data.
- Runs on clusters of servers where each server has a local CPU.

Hadoop Architecture

Hadoop framework has four modules as follows.

- Hadoop Common
- Hadoop YARN
- Hadoop Distributed File System (HDFS)
- Hadoop MapReduce

Hadoop Common

Hadoop Common is a Java library that provides necessary java files and scripts which are required to start Hadoop.

Hadoop YARN

Hadoop YARN is a framework for managing cluster resource and job scheduling.

Hadoop Distributed File System (HDFS)

HDFS is a critical component of Hadoop. Hadoop cluster uses this as a storage system. When the cluster receives the data, then HDFS breaks it into pieces and distributes it among different servers in the cluster. The copy of each piece of data will be stored on more than one server. HDFS provides a list of commands in the file system as a shell that is available to interact with the file system.

Hadoop MapReduce

Hadoop MapReduce is an agent that distributes the works to all servers participating in the cluster and collects the results. It handles a large amount of data in parallel on large clusters in a fault-tolerant manner.

There are two different tasks performed by Hadoop MapReduce.

- **The Map Task:** This task takes input data and breaks individual elements into tuples (key/value pairs).
- **The Reduce Task:** This task takes the output of a map task as an input and combines the data tuples into a smaller set of tuples. This task is always performed after the map task.

Advantages of Hadoop

- It allows users to quickly access distributed systems.
- It is efficient in distributing the data to different servers automatically.
- It is fault-tolerant and easily available because the Hadoop library detects and handles the failures by itself.
- Any number of Servers can be added and removed from the cluster.
- It supports all platforms because it is java based.

5.3.2 Old vs. New Approaches

New approaches and open-source inventions make it easy to store, manage and analyze the data and improve the ability to handle unstructured data. The following table illustrates the difference between old and new approaches.

Table 5.1 compares traditional data warehouse and big data approaches.

Table 5.1: Comparison of Traditional Data and Big Data Approaches

S. No.	TRADITIONAL DATA	BIG DATA
01.	Traditional data is generated at the enterprise level.	Big data is generated in the outside and enterprise levels.
02.	Its volume ranges from Gigabytes to Terabytes.	Its volume ranges from Petabytes to Zettabytes or Exabytes.
03.	Traditional database system deals with structured data.	Big data system deals with structured, semi-structured and unstructured data.
04.	Traditional data is generated on an hour or day basis.	Big data is generated more frequently mainly per second.
05.	The traditional data source is centralized and is managed in a centralized form.	The big data source is distributed and it is managed in distributed form.

Contd…

06.	Data integration is very easy.	Data integration is very difficult.
07.	Normal system configuration is capable to process traditional data.	A high system configuration is required to process big data.
08.	The size of the data is very small.	The size is more than the traditional data size.
09.	Traditional database tools are required to perform any data base operation.	Special kinds of database tools is required to perform any data base operation.
10.	Normal functions can manipulate data.	Special kinds of functions can manipulate data.
11.	Its data model is strict schema-based and it is static.	Its data model is flat schema-based and it is dynamic.
12..	Traditional data is stable and inter related.	Big data is not stable and not inter related.
13.	Traditional data is in a manageable volume.	Big data is in huge volume which is unmanageable with the normal tools.
14.	It is easy to manage and manipulate data.	It is difficult to manage and manipulate data.
15.	Its data sources include ERP transaction data, CRM transaction data, financial data, organizational data, web transaction data, etc.	Its data sources include social media, device data, sensor data, video, images, audio, etc.

Source: https://www.geeksforgeeks.org/difference-between-traditional-data-and-big-data/

Example: How Hadoop Based Database Structure Helps Netflix with Insights to Improve its Recommendation System

Netflix was the most liked entertainment brand in America with millions of customers hooked to its streaming services. The company generated huge data in Realtime from user actions. This data was stored in data structures based on Hadoop and other technologies. The data was then analysed to find insights into individual customer choices like when did someone pause, did he restart etc. Based on this, the company improved its recommendations and the customers like it.

Source: https://data-flair.training/blogs/big-data-case-studies/, 2022, Accessed on 10/08/2022.

Check Your Progress - 1

1. Which of the following is not one of the four dimensions of big data.

   a. Volume

   b. Velocity

   c. Veracity

   d. Variety

   e. Value

2. Identify the data format which is supported by big data but is not supported by traditional data warehousing.

   a. Unstructured

   b. Semi-structured

   c. Structured

   d. XML

   e. Relational

3. The open-source framework Apache Hadoop is written in which programming language?

   a. C++

   b. Java

   c. Net

   d. PHP

   e. Python

4. Apache Hadoop stores how much of data?

   a. Petabytes

   b. Exabytes

   c. Zettabytes

   d. Yottabytes

   e. Megabytes

5. Which of the following breaks received data into pieces and distributes among different servers in the cluster?

   a. Hadoop common

   b. Hadoop YARN

   c. Hadoop Distributed File System (HDFS)

   d. Hadoop Map

   e. Hadoop MapReduce

Activity 5.1

Consider you are the marketing head of a car manufacturer. You want to make a highly consumer-centric car. You are aware that knowing consumer preferences through various sources like social media and product reviews would help you to understand the taste and preferences of the customers. You can make use of big data to predict consumer perception of a specific car. Also, you know that data analytics usage would help you predict trends, market patterns, and future demands so that the best car for the target market can be designed. Analyze the approach, old versus new, that would be used by you for the purpose.

5.4 Data Discovery

Today we have all become accustomed to ‘Search’, that is “Googling”. This is an example of data discovery. Another software. QlikTech offers a quick search method wherein the users get results just by typing a relevant word in any order.

5.4.1 Data Discovery: Work the Way People’s Minds Work

In an interview, Anthony Deighton, QlikTech’s CTO said, “Business Intelligence needs to work the way people’s minds work. Users need to navigate and interact with data any way they want to – asking and answering questions on their own and in big groups or teams.” Deighton is responsible for guiding the product strategy and leads all aspects of the company’s R&D efforts for its product suite.

Data discovery is ever-evolving. The more you offer to the customer, the more becomes their demand. QlikTech is a reputed BI vendor with a considerable presence in Denmark. Still, it preferred to go for a tie-up with Discus Communication, a local communication firm, to meet the growing demand of Danish customers.

Activity 5.2

You are a Human Resource Manager in a software multinational. You are assigned the task of making sure there is enough supply of human resources for an upcoming software project. The observed problem with other HR managers is that they often fail to unravel hidden talents within organizational pockets and persistently feel the need to hire external talent which comes with its own set of challenges.

You want to discuss with the IT head to check how big data with Natural Language Processing can help understand the HR queries and analyze the performance review feedback, reviews of projects and come up with the employee talent profile data. You want to explore how this would help in the near-real-time to act as a tool for workforce planning. Explore how big data can help in data discovery about the availability of the right human resource for your projects.

5.4.2 Open-Source Technology for Big Data Analytics

These days big data is very widely used. Open-source technology is the core of big data initiatives. There are nine open-source big data technologies. These technologies are available under Apache License 2.0.

1. Apache Hadoop
2. R
3. Cascading
4. Scribe
5. Elastic Search
6. Apache HBase
7. Apache Cassandra
8. MongoDB
9. Apache CouchDB

1. Apache Hadoop

Apache Hadoop is the most popular technology for storing structured, semi-structured and unstructured data that comprise big data. It was developed by ‘Doug Cutting’ to support his work on an open-source search engine called ‘Nutch’. He named it after his son’s toy elephant ‘apache’.

2. R

R is an open-source programming language designed for statistical computing. It was developed by Ross Ihaka and Robert Gentleman at the University of Auckland in 1993. It has become a famous tool for statistical analysis for large data sets.

3. **Cascading**

   Cascading is an open-source software that is used for ad targeting, bioinformatics, predictive analysis, and web content mining. It was developed by Chris Wensel. It allows users to create and execute data processing workflows on Hadoop clusters.

4. **Scribe**

   Scribe is a web server developed by Facebook in 2008. It aggregates data from a large number of servers in real-time. Now Facebook is handling tens of billions of messages a day.

5. **Elastic Search**

   Elastic search is a distributed open-source search server. It was developed by Shay Banon. It supports real-time search without a special configuration.

6. **Apache HBase**

   Apache HBase is an open-source non-relational distributed database. It is used to run Hadoop Distributed File System (HDFS). It provides quick access to large quantities of sparse data. Facebook adopted HBase in 2010 to serve the messaging platform.

7. **Apache Cassandra**

   It is another open-source distributed database management system (NoSQL data store). It was developed by Facebook to manage Inbox search features in 2010. It is also used as a back-end database for few companies like Netflix.

8. **MongoDB**

   MongoDB is another popular NoSQL data store. It stores structured data with dynamic schemas. It was developed by the founders of DoubleClick and adapted by large enterprises like The New York Times, MTV networks, Disney interactive media group, etc. It is available under the GNU Affero General Public License and the language drivers are available under an Apache License.

9. **Apache CouchDB**

   Apache CouchDB is also another open-source NoSQL database. It was developed by IBM as a storage system for large database objects in 2005. It uses JavaScript as query language and JSON to store data and HTTP for an API. It is used by BBC for its dynamic content platforms.

5.4.3 The Cloud and Big Data

Cloud service provides the most optimized big data management system. Oracle big data cloud service is designed to securely run big data workloads and technologies with simple operations. Oracle’s big data SQL cloud service makes organizations analyze data across Hadoop and NoSQL.

Oracle’s big data cloud service delivers big data cloud security with the highest data security levels currently available for Hadoop. Network encryption prevents network sniffing from capturing protected data. Cloud provides additional security for all Hadoop services.

Why Big Data in the Cloud?

1. Managed: Cloud manages the cluster for you like single-click patching and upgrades.
2. Dedicated: It delivers consistently and is high in performance.
3. Connected: It integrates with other Oracle services using big data connectors.
4. Integrated: It stores data in the storage cloud service and processes with Hadoop clusters.
5. Comprehensive: Oracle’s big data cloud service enables you to easily deploy a complete big data management system using Oracle SQL with unified query.
6. On-demand: It provides self-service Hadoop clusters which are easy and rapid.
7. Elastic: A small cluster can be grown into a large cluster to handle petabytes of data.
8. Secured: Securing data is critical to big data solutions. Oracle big data cloud service provides strong authentication, authorization and auditing of data in Hadoop with just a single click.

Example: How Honda Used Qlik to Streamline its Business Processes and Create more Efficient Operations

Honda (the automobile major) had some 400 plus companies with a staff strength of 2 lakh. The group recognized the need for digitization. It wanted to add value at every stage by every employee by using processed data available in an easily understood and usable form. It chose Qlik for data discovery and visualization. The exercise resulted in more efficient operations, next generation services. Also, the data culture was spread company wide. Two key success indicators: it took just one day to prepare plan documents instead of earlier one month. The second one was: 50 reports were condensed to one.

Source: https://www.qlik.com/us/-/media/files/resource-library/global-us/direct/case-studies/cs-honda-undergoes-digital-transformation-en.pdf?rev=cca1b9b642934f1fb8e388d06a22c472, 2021, Accessed on 10/08/22 .

Check Your Progress - 2

6. Doug Cutting developed Apache Hadoop and named it after which of his son’s toy?

   a. Car

   b. Elephant

   c. Plane
   d. Train
   e. Lion

7. Scribe is a web server developed by which company in 2008?
   a. Facebook
   b. Twitter
   c. LinkedIn
   d. WhatsApp
   e. Viber

8. Which of the following is an open-source non-relational distributed database?
   a. Apache Hadoop
   b. Apache HBase
   c. Apache Cassandra
   d. Apache CouchDB
   e. MongoDB

9. Which property of the Cloud makes a small cluster grow to large cluster to handle petabytes of data?
   a. Integrated
   b. Dedicated
   c. Comprehensive
   d. Elastic
   e. On-demand

10. Which of the following is a programming language designed for statistical computing for large data sets?
   a. Apache Hadoop
   b. R
   c. Cascading
   d. Scribe
   e. Elastic Search

5.5 Summary

- Big data is the next generation of data warehousing which facilitates cost efficiency for enterprises.
- The world is becoming more transparent and the amount of data is increasing rapidly every year in which most of the new data is unstructured. It is very difficult to analyze unstructured data because it represents 95% of new data.

- Apache Hadoop is an open-source platform that supports the processing of unstructured data.
- In this unit, we discussed the architecture of the Hadoop and the working of Hadoop distributed file system. We also discussed how the cloud supports big data.

5.6 Glossary

API (Application Program Interface): API is a tool for building software applications with a set of routines and protocols. It specifies how software components should interact when using Graphical User Interface (GUI) components.

DoubleClick: DoubleClick is an internet advertisement technology. It was developed by O'Connor in 1995.

Hadoop: A Hadoop is an open-source software framework written in java that supports the processing of large data sets in a distributed computing environment.

HTTP (Hypertext Transfer Protocol): HTTP is the foundation of data communication for the World Wide Web (WWW). It is an application protocol for distributed information systems.

JSON (JavaScript Object Notation): It is a lightweight data-interchange format that makes humans write and read easily. It also makes machines easy to parse and generate.

NoSQL Database: A NoSQL database is a mechanism for storage and retrieval of data which means other than tabular relations used in relational databases.

5.7 Self-Assessment Questions

1. What are the four dimensions of big data technology?
2. Explain the four modules of Hadoop architecture.
3. What are the advantages of the Apache Hadoop framework?
4. Differentiate between the old and new approaches of data processing.
5. 'Why big data in the cloud?' Justify.

5.8 Suggested Readings/Reference Material

1. Maleh, Yassine. Shojafar, Mohammad. Alazab, Mamoun. Baddi, Youssef. Machine Intelligence and Big Data Analytics for Cybersecurity Applications (Studies in Computational Intelligence, 919) 1st ed. 2021 Edition.
2. Ahmed, Syed Thouheed. Basha, Syed Muzamil. Arumugam, Sanjeev Ram. Patil, Kiran Kumari. Big Data Analytics and Cloud Computing: A Beginner's Guide, 2021.

3. Saleem, Tausifa Jan. Chishti, Mohammad Ahsan. Big Data Analytics for Internet of Things 1st Edition.

4. Jones, Herbert. Data Science: The Ultimate Guide to Data Analytics, Data Mining, Data Warehousing, Data Visualization, Regression Analysis, Database Querying, Big Data for Business and Machine Learning for Beginners Hardcover – 10 January 2020.

5. Maheshwari, Anil. Data Analytics Made Accessible: 2023 edition Kindle Edition.

6. Mayer-Schönberger, Viktor. Cukier, Kenneth. Big Data: A Revolution That Will Transform How We Live, Work, and Think Paperback – October 26, 2021.

5.9 Answers to Check Your Progress

1. (e) Value

Big data is characterized as possessing four dimensions called the 4Vs. They are Volume, Velocity, Variety, Veracity. Another dimension V (Value/Valor) is also used to characterize the quality of the data.

2. (a) Unstructured

Big data supports unstructured data format which is not supported by the traditional data warehouse. It supports different data types such as structured-like relational databases, semi- structured-like XML and unstructured-like audio, video and text.

3. (b) Java

Apache Hadoop is an open source framework that allows distributed processing of large databases using simple programming written in Java. It provides distributed storage across clusters of computers.

4. (a) Petabytes

Hadoop handles data warehousing including search, log processing and workloads. It can store all kinds of data and performs variety of analysis and transformation on the data and it stores terabytes and even petabytes of data.

5. (c) Hadoop Distributed File System (HDFS)

Hadoop Distributed File System (HDFS) is a critical component of Hadoop. Hadoop cluster uses this as a storage system. When cluster receives the data, then HDFS breaks it into pieces and distributes it among different servers in the cluster.

6. (b) Elephant

Apache Hadoop is an open source framework that allows distributed processing of large databases. It was developed by Doug Cutting. Doug named it after his son's toy elephant and now it has become the trademark for Apache Software Foundation.

7. (a) Facebook

Scribe is a web server developed by Facebook in 2008. It aggregates data from large number of servers in real time. Now Facebook is handling tens of billions of messages a day.

8. (b) Apache HBase

Apache HBase is an open-source non-relational distributed database. It is used to run Hadoop Distributed File System (HDFS). It provides quick access to large quantities of sparse data. Facebook adopted HBase in 2010 to serve the message platform.

9. (d) Elastic

Oracle big data cloud service is designed to securely run big data workloads and technologies with simple operations. The elastic property of the cloud makes small cluster grow to large clusters to handle petabytes of data.

10. (b) R

R is an open-source programming language designed for statistical computing. It was developed by Ross Ihaka and Robert Gentleman at the University of Auckland in 1993. It has become a famous tool for statistical analysis for large data sets.

Unit 6

Big Data Technologies and Terminologies

Structure

"The goal is to turn data into information and information into insight."

- Carly Fiorina, Former CEO, Hewlett Packard

6.1 Introduction

It is not just enough for organizations to invest in collecting huge data and analyse to get useful information. Organizations need to act on insights provided by the information to add value to its stakeholders.

In the previous unit, we shared about the Elephant Room, Hadoop's Parallel World, and data discovery, open source technology for big data analytics, and the cloud and big data.

Technology is changing the way the data is produced, processed, analyzed and consumed. Besides, technology is evolving with new data sources. Therefore, more data gets captured every day and it is a big task to process the unstructured data into structured data. Technology is helping in processing the data quickly and efficiently and plays a pivotal role in making informed decisions. In this unit, we will be discussing the different technologies supporting big data analytics and the different terminology we are using in big data analysis.

6.2 Objectives

At the end of the unit, you will be able to:

- Discuss how predictive analytics moves into the limelight
- Describe how the companies use the software as a service BI (SaaS BI)
- Analyze the role of mobile business intelligence in big data analytics
- Discuss crowdsourcing analytics and its uses
- Differentiate inter and trans-firewall analytics

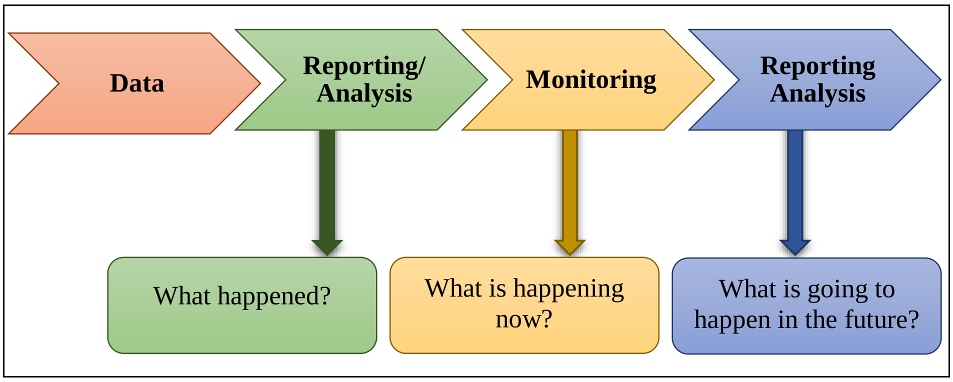
6.3 Predictive Analytics Moves into the Limelight

Predictive analysis means making predictions about future events. Predictive analysis uses techniques like statistics, artificial intelligence and data mining to analyze the present data and generate predictions about the future. It also identifies risks and opportunities for the future.

The predictive analysis makes organizations more proactive and forward-looking. It provides decision options to get benefits from the predictions.

Figure 6.1 presents the value chain of predictive analysis.

Figure 6.1: Predictive Analysis Value Chain



Source: ICFAI Research Center

Predictive Analysis Process is as follows:

- **Define Project**

  Predictive analysis defines the business objectives and project outcomes and identifies data sets for that project.

- **Data Collection**

  Predictive analysis collects data from multiple sources for analysis and also for customer interaction.

- **Data Analysis**

  Predictive analysis analyzes the data using the process of inspecting, transforming and modeling and discovers useful information.

- **Statistics**

  Predictive analysis uses statistical methods to validate the data assumptions and hypotheses.

- **Modeling**

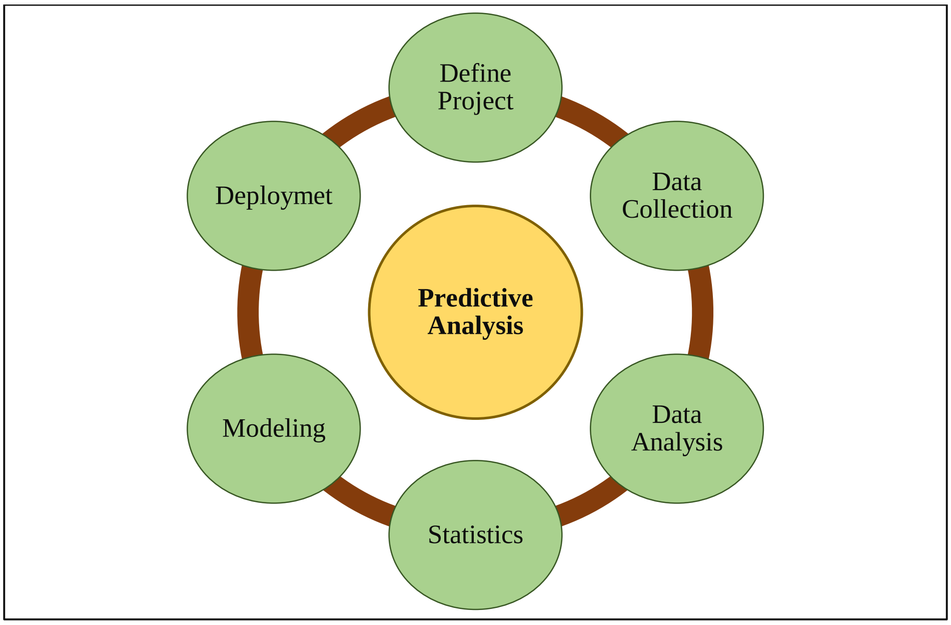
  Predictive analysis creates models for the future so that the management can choose the best model for optimizing the assumptions.

- **Deployment**

  Predictive analysis deploys the analytical results into the decision-making process to get better results.

Figure 6.2 presents predictive analysis process.

Figure 6.2: Predictive Analysis Process



Source: ICFAI Research Center

Example: Tata Motors uses Predictive Analytics To Reduce Cost of Acquisition of Customer (CAC) by 75%

Tata Motors was the second largest car manufacturer in India. The company wanted to reduce the Cost of Acquisition of customer who visited their show room. They took the help of Predictive Analytics (based on machine learning model) for this.

Contd….

The new deployment resulted in 75 percent reduction in CAC within 3 months. The model was also able to suggest locations where new dealerships were to be opened. The time taken to start a new dealership also came down from around 5 to 6 months to few weeks.

Source: https://infiniteanalytics.com/2022/06/20/how-tata-motors-used-ai-to-reduce-cac-by-75/, Accessed on 11/08/2022

Activity 6.1

Predictive Analytics Moves into the Limelight

Mr. Iyer, an IT consultant was hired by an online grocer to address delivery delays and also to organize its business data. The portal plans to improve its service with the help of predictive analysis; for optimizing route and truck space based on order history. Suggest the sequence of steps Mr. Iyer needs to follow to organize information and deployment of a predictive analysis system at the portal.

Answer:

6.4 Software as a Service BI

What is Software-as-a-Service Business Intelligence (SaaS BI)?

SaaS BI is a combination of Business Intelligence (BI) and Software-as-a-Service (SaaS).

SaaS BI is a cloud-hosted application that delivers BI tools, like dashboards to enterprises like SaaS, offering lower costs, less complex and being easier to deploy.

SaaS BI Cost - Single User

SaaS BI offers a significant application use of adding users to the application, over on-premise applications. It is much easier and also cost-efficient when you decide on the option of adding users to a SaaS. It can be 1 or 100 or 1000 users.

What does adding a single user to a SaaS dashboard mean?

It is as simple as paying for an additional "seat". At the SaaS level, it is supplying the user with all necessary rights for accessing the appropriate dashboard, using a web browser.

But one should remember that for this facility, the BI application requires installing the software, syncing the application and also linking all other computers in the network. It may be a disadvantage.

SaaS BI and Accessibility

Another advantage SaaS BI provides is accessibility. SaaS applications can be accessed at any location, with any device; and all you need is an internet connection. As only a web browser is required, it is easier to deploy the application and view the information. It also makes the application easier to deploy because it only requires a web browser.

SaaS Dashboards - BI in the Cloud

A SaaS dashboard is a representation of the expression of BI in the cloud. The dashboards summarize various key facts as needed and designed for the business, for visual/quick consumption leading to a good information system.

A SaaS dashboard helps enterprises to reach a large audience quickly and effectively. Such deployment encourages users to view business-critical information.

Example: How South African Bank Nedbank Used Microsoft's SAAS BI Tool to to Identify Projects which Are Likely to Fail

Nedbank Group (South Africa based) was the 4th largest bank in the country offering various financial products. The operations of the bank were customer centric. With that as its core philosophy , the bank had been deploying technology. The company wanted people in all departments to use data for their needs without depending on analysts and scientists. The company implemented SAAS, BI tool from Microsoft Power BI . The tool enabled people at all levels to generate dashboards for themselves. One of the dashboards dealt with Project Risk Indicators. It gave insights to predict the risk of projects failing. The user could drill down to find attributes contributing to the overall predictive project risk

Source: https://customers.microsoft.com/en-us/story/848386-nedbank-banking-capital-markets-power-bi, 2020 (Accessed on October 9, 2020)

6.5 Mobile Business Intelligence is going Mainstream

Mobile devices are ensuring optimal performance and the best user experience.

Aberdeen Group, the US-based international market research firm, from their surveys found that smartphones, tablets, iPad and Android units, are not the only preferred device for mobile and general access.

Mobile device access to BI apps is in great demand from high level executives. Their preference is to explore opportunities and solve problems. Many executives prefer a tablet device to desktop and laptop PCs, even in the office

There are two approaches to mobile BI.

- Running an Application (App)
- Web browser-based access

The advantages are, apps allow users for off-line access when they are not connected to the internet. Whereas web browser based access needs an internet connection for mobile business intelligence.

The future of mobile BI is very bright because it offers accurate, reliable, and secured information to users compared to desktops. This flexible platform allows for better decisions and maximum productivity.

iPhone was introduced in 2007 and iPad, in 2010. On the first day, 3 lakh iPhones were sold in the market. The figure rose to forty lakhs in 2011. The number of smartphones sold in 2011 was 385 million. This shows the acceptance of Mobile BI.

Example: How Ascentis used Mobile BI to Share Dashboards with Staff Working Away from Office

Ascentis was a workforce management company. It had many Key Performance Indicators (KPIs) to be monitored at each department. Also, some of the departmental employees worked away from office. The company used Mobile BI dashboards from Klipfolio. The software enabled to publish live URLs of the dashboards to remote working staff. For the staff working from office, the dashboards were displayed on TV walls in each department with relevant metrics data.

Source : https://www.klipfolio.com/blog/mitigate-performance-issues-with-klipfolio February 15, 2022 Accessed on 11/08/2022

6.6 Crowdsourcing Analytics

Crowdsourcing is a modern business term that means the process of obtaining needed services, ideas, or content from a large group of people from an online community.

The internet has provided a venue for crowdsourcing and impersonalization. This led to artistic projects because individuals are less conscious of the scrutiny and identity of their work.

In an online environment, people pay attention to the specific needs of a project. Communication with other individuals takes a back seat.

For example:

Crowd Voting: It means gathering opinions and judgments from large groups on a certain topic. The following are some crowd voting sites.

- http://outsideinmarketing.wordpress.com/2010/05/08/four-types-of-crowdsourcing/
- http://uxpost.com/post/7425534086/real-time-interactive-crowd-voting

Crowd Funding: It is the process of funding the projects via the internet by multitudes of people. The following are some crowd funding sites.

- My Dream Store - A place to create and sell custom t-shirts without any upfront costs, without any risks or hassles.
- Evangel - Crowdfunding Platform
- Fundlined - Fund the Change You Want To See!
- Wishberry - GO FUND YOURSELF!
- http://www.thehotstart.com/
- Fund a Peer - Homepage
- Indiegogo: Crowdfunding for what matters to you - start now!
- Crowd Funding Platform India-Start51
- Homepage | Ketto.org
- Pik A Venture - Empowering Entrepreneurship and Creativity!!
- http://fundmydream.in/

Example: How Crowdflower Lower Costs and Reduced Data Errors

Crowdflower realised that getting its big data cleaned up without errors for analysis was a time consuming and costly process with no guarantee of error free data just employing inhouse staff. It adopted crowdsourcing option and employed 50 Lakh people who accessed and cleaned up data through internet. The data errors had reduced, and the costs came down a lot.

Source: https://www.simplilearn.com/big-data-vs-crowdsourcing-ventures-revolutionizing-business-processes-rar367-article (Oct 11, 2021) Accessed on 11/08/22

Check Your Progress - 1

1. Which component of predictive analysis validates the data assumptions and hypothesis?
   a. Define project
   b. Data collection
   c. Data analysis
   d. Statistics
   e. Deployment
2. Which of the following means the expression of Business Intelligence in the SaaS dashboard?
   a. Network
   b. Web browser
   c. Internet

   d. Data mart
   e. Cloud

3. The demands for BI apps on mobile devices are primarily coming from which of the following employees?
   a. Managers
   b. Low-level executives
   c. Middle-level executives
   d. High-level executives
   e. Customers

4. Which of the following provides good venue for crowdsourcing?
   a. Cloud
   b. Database
   c. Internet
   d. Mobile device
   e. Personal computer

5. Which of the following is the process of gathering opinions and judgments from large groups?
   a. Crowdfunding
   b. Crowd purchasing
   c. Crowd search
   d. Crowd voting
   e. Crowd wisdom

6.7 Inter and Trans-Firewall Analytics

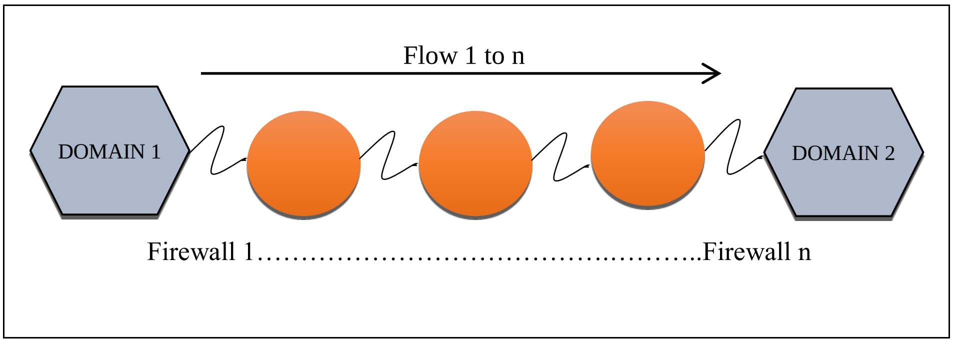
The data in the world is doubling every year as the business is transforming rapidly. Analytics and data-driven decision making are needed to develop effectively.

Today's organizations have been generating huge amounts of transactional data for the last 20 years. Transactional data has been growing fast and social media content is adding volumes to this phenomenon. The volumes of information the business groups collect about their customers and partners through social media like LinkedIn, Twitter and Face book have changed the concept of transactional data completely.

Inter-firewall: Inter-firewall means multiple firewalls in the same Enterprise network. It provides security to the local domain and applications. It also protects from internally generated traffic.

The inter-firewall may exist if any two firewalls on a network take different filtering processes on the same traffic. Figure 6.3 presents inter firewall analytics.

Figure 6.3: Inter-Firewall Analytics



Source: ICFAI Research Center

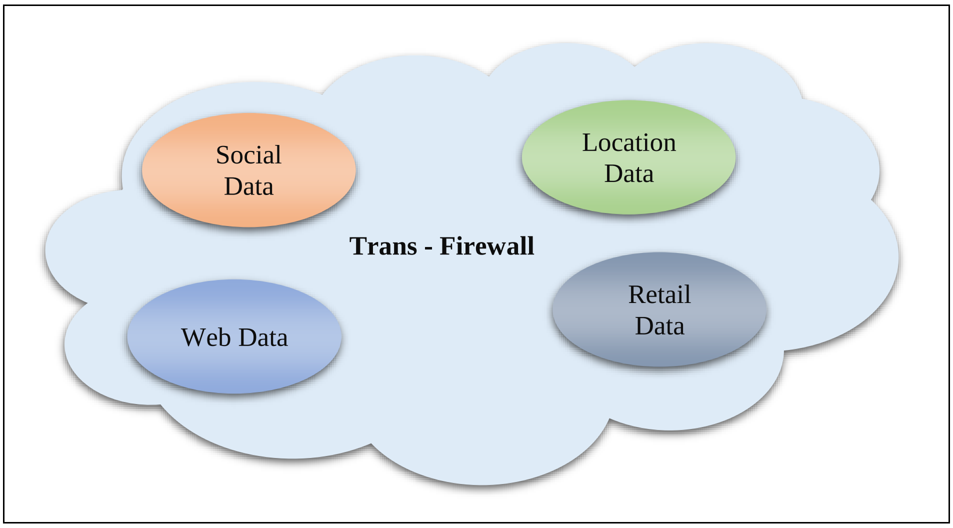
Trans-Firewall

The enterprises started collaborating on various insights emanating across the value chain. Now supply has connected multiple companies and concepts like CPFR, VMI, etc. enabled enterprises to create considerable value to the end-consumer.

Let’s take an example from the health care industry. Useful insights about affluent consumers can be generated by collaborating on big data and related mining insights from various sources like pharmacies delivering the drugs, the health insurance provider, and the drug manufacturer.

More progressive companies are leveraging the large volumes of data like social data, location data, etc. crossing their firewall boundaries. Figure 6.4 shown firewall analytics.

Figure 6.4: Trans – Firewall analytics



Source: ICFAI Research Center

This trend is popular and is referred to as - a move from inter-organization to available trans-firewall analytics. Companies were already doing intra-firewall analytics with data available within the organization's boundaries. In the future, they will be collaborating with other companies, which facilitate trans-firewall analytics and inter-firewall analytics.

Example: How Hygeia Hospital Participated in Project Musketeers and Shared Medical Data while Maintaining the Privacy and Security

Hospitals generated large volumes of medical data (imaging) coming from CT scans. But individual hospitals were not able to collaborate and share data for experts to analyse this data and benefit. The reason for this being security, privacy issues. Hygeia Hospitals was part of project Musketeers which had hospitals and technology analytics companies working to create a platform which enables hospitals to share data (trans firewall) while preserving the privacy of data (source) and security.

Source: https://h2020musketeer.medium.com/federated-learning-to-support-diagnostic-imaging-for-next-generation-ai-powered-healthcare-20bdf8ed9412 (October 26, 2021) Accessed on 12/08/2022

6.8 R&D Approach Helps Adopt New Technology

Every company's executives make innovation a top priority for their firm. This we can observe in their annual report. For example, a market executive analyzes market and customer data to make changes in the new product to attract the customer.

Technology brand is a vital part of the skill set for organizations. After spending years implementing data management and analytic stack, they find it difficult to make changes. At the same time, the company should take care to minimize the risk while adopting new technologies.

The CIO of R&D, VP of Technology Strategy & Innovation at Visa Inc., Joe Cunningham said that the R&D organizations have "small 'r' and big 'D'. He means that the organization should have fewer goals for research but more for development. It means the team should concentrate more on application of technology.

Adding Big Data Technology

Cunningham took different approaches to add new big data technology. There are two approaches.

i. **Practical Approach:** Begin with the problem on hand and then find the appropriate solution.

ii. **Opportunistic Approach:** Look at available technology for taking benefits and then find a suitable application at home for it.

These two approaches perform the following activities:

- **Play:** Installing the technology in the lab to be familiar with the new technology.
- **Initial Business Review:** Validating and ranking the technology by talking with the business owners.
- **Architecture Review:** Assess the validity of the architecture and ensure its standards.
- **Pilot Use Cases:** Finding a suitable-use case to test the technology.
- **Transfer from R&D to Production:** Moving the technology from research to production.

6.9 Big Data Technology Terms

Table 6.1 introduces the terms of big data technology.

Table 6.1: Terms in Big Data Technology

Algorithm	Any mathematical method with a process, placed in the developed software, to perform data analysis.
Analytics	Using mathematical algorithms and allied software coupled with statistics to derive meaning from data.
API	It is a set of programming standards and instructions and facilitates building web-based, software-based applications.
Business Intelligence (BI)	Business intelligence comprises identification, extraction, and analysis of data.
Cassandra	A columnar database used in big data applications, being managed by The Apache Software Foundation. It is also an open-source database.
Clojure	Clojure is a dynamic programming language based on LISP. It is well suited for parallel data processing. It uses the Java Virtual Machine (JVM).
Cloud	Cloud is hosted remotely. It refers to any Internet-based application or service which can be used and paid for
Clustering Analysis	It is a data analysis process. It helps in identifying similarities and differences among data sets. Such identification helps the process to take the benefit of bringing similar data sets together.

Contd….

Crowd Sourcing	The act of bringing people together for submitting a task or problem to complete or find solution working together.
Dashboard	A graphical platform presenting data of high-level to benefit managers with a quick report on status or performance, indicating static or real-time actual happened data. It is accessible on a mobile device or desktop.
Data Marketplace	It can be defined as a place to sell and buy data online.
Data mart	A Data mart is a minimized version of a data warehouse. It is designed to facilitate use by an individual and specific department, unit or set of functional users in the organization.
Data Migration	The process of moving data. It could be (1) between various computer systems, or (2) between different storage types or formats.
Data Mining	The process of using predictive techniques and statistical analysis techniques to extract inherent patterns or knowledge from voluminous data sets.
Metadata	Metadata indicates data about data. It is a data that describes another data. Meta in technology circles means - "an underlying definition or description". Examples: author, date modified, and file size. The concept is applicable for unstructured data: Video, images, web pages, spreadsheets, etc.
MongoDB	An open-source NoSQL database managed by 10gen.
NoSQL	NoSQL handles large data volumes that do not necessarily follow a fixed schema. It is designed to suit f non-relational models and very large databases.
OLAP - Online Analytical Processing	Three operations: consolidation, drill-down, and slice and dice, are used in the process of analyzing multi-dimensional data. a. Drill-down - users can drill deep to see the underlying details b. Consolidation - total summing up of available data c. Slice and dice – Choice to select subsets - view them from various perspectives

Contd….

Online Transactional Processing - OLTP	Providing access to large amounts of transactional data, in a spirit to derive meaningful analysis.
Petabyte	One million gigabytes or 1,024 terabytes.
Pig	It is a data flow language and execution framework. It is mostly used for parallel processing computation.
Predictive Analytics	Provides the user the capability to analyze existing data using statistical functions, individual or more datasets; aiming to mostly predict future events or trends.
Unstructured Data	Data that probably has no identifiable structure. Examples: video, images, the text of email messages.
Semi-structured Data	Data that is not structured or following any formal data model. It has the capability of providing other ways of describing the data and related hierarchies.
Structured Data	Data that is well organized by, or has a similarity to, some existing determined structure.
Structured Query Language (SQL)	It is a programming language. The objective is to facilitate queries to manage data and retrieve data from relational database systems.
Sentiment Analysis	It is a set of statistical functions which can be applied or operated on the comments people make on the web. It could be through social networks, indicating their feelings about a product or company.

Source: http://data-informed.com/glossary-of-big-data-terms/

Activity 6.2

Rahul works as a Business Analyst at IntelliSoft Ltd., which specializes in offering big data solutions. He joins a life insurance firm and was asked to study and find out the insurer’s premium defaulters and predict good and bad customers to generate app-based alters to insurance buyers and to display the real-time updates to the firm’s managers. Suggest a couple of big data technology-based tools for Rahul to meet the firm’s needs.

Answer:

6.10 Data Size

Find here the various data sizes in computer technology.

- The smallest unit of data is a bit.
- A single bit can have a value of either 0 or 1.
- The fundamental unit of measurement for data is a byte or eight bits.
- A byte can store 28 or 256 different values, which is sufficient to represent standard ASCII characters, such as letters, numbers, and symbols.
- A file contains thousands of bytes, therefore, a file is measured in kilobytes.
- A large file with images, videos and audio contain millions of bytes, therefore it is measured in megabytes.
- A storage device can store thousands of files, therefore it is measured in gigabytes or terabytes.
- The multiple storage devices are measured in large units of measurement like petabytes, exabytes, zettabytes and yottabytes.

Figure 6.5 presents the international nomenclature for data sizes.

Figure 6.5: The List of all the Standard Units of Measurement from the Smallest to the Largest

COMMON DATA STORAGE MEASUREMENTS	
Unit	**Value**
bit	1 bit
byte	8 bits
kilobyte	1,024 bytes
megabyte	1,024 kilobytes
gigabyte	1,024 megabytes
terabyte	1,024 gigabytes
petabyte	1,024 terabytes

Unit	Value	Size
bit (b)	0 or 1	1/8 of a byte
byte (B)	8 bits	1 byte
kilobyte (KB)	1000^1 bytes	1,000 bytes
megabyte (MB)	1000^2 bytes	1,000,000 bytes
gigabyte (GB)	1000^3 bytes	1,000,000,000 bytes
terabyte (TB)	1000^4 bytes	1,000,000,000,000 bytes
petabyte (PB)	1000^5 bytes	1,000,000,000,000,000 bytes
exabyte (EB)	1000^6 bytes	1,000,000,000,000,000,000 bytes
zettabyte (ZB)	1000^7 bytes	1,000,000,000,000,000,000,000 bytes
yottabyte (YB)	1000^8 bytes	1,000,000,000,000,000,000,000,000 bytes

Adapted from ICFAI Research Center

Check Your Progress - 2

6. Which of the following security is provided by the Inter-firewall?
   a. Local domain
   b. Social data
   c. Location data
   d. Web data
   e. Retail data

7. Which of the following activity validate and rank the new technology talking to the business owners?
   a. Play
   b. Initial business review
   c. Architectural review
   d. Pilot use cases
   e. Moving from R&D to production

8. Which of the following is the approximate size of a large file with images, videos and audio?
   a. Kilobytes
   b. Megabytes
   c. Petabytes
   d. Exabytes
   e. Zettabytes

9. 1 Gigabyte = how many bytes?
   a. 1000^1
   b. 1000^2
   c. 1000^3
   d. 1000^4
   e. 1000^5
10. Which of the following is a place where people can buy and sell data online?
    a. Data mart
    b. Database
    c. Data marketplace
    d. Data migration
    e. Data mining

6.11 Summary

- In the process of big data analytics, enterprises will move from business intelligence to predictive analytics. This generates predictions about the future, identifying risks and opportunities for the future.
- Software as a service BI is cloud-hosted and it delivers BI tools like dashboards to enterprises. The software has low cost, less complexity and can be deployed easily.
- Mobile devices are ensuring optimal performance and the best user experience and are widely used in BI applications.
- Crowdsourcing is a modern business process of obtaining needed services, ideas or content from a large group of people from an online community.

6.12 Glossary

Algorithm: Any mathematical method with a process, placed in the developed software, to perform data analysis.

API: It is a facilitating access or building web-based software-based applications and is a set of programming standards and instructions.

Crowd Sourcing: The act of bringing people together for submitting a task or problem to complete or find a solution working together.

Data Migration: The process of moving data. It could be between various computer systems or between different storage types or formats.

Data Mining: The process of using predictive techniques and statistical analysis techniques to extract inherent patterns or knowledge from voluminous data sets.

Predictive Analytics: Provides the user the capability to analyze existing data using statistical functions, individual or more datasets. It is used to predict future events or trends.

Sentiment Analysis: It is a set of statistical functions which can be applied or operated on the comments people make on the web. It could be through social networks, indicating their feelings about a product or company.

6.13 Self-Assessment Test

1. Explain how predictive analytics moves into the limelight.
2. What is the role of mobile business intelligence in big data analytics?
3. What is Crowdsourcing? What are the different techniques we use?
4. What is the difference between inter firewall and trans-firewall analytics?
5. What are the different activities we follow in R&D approach?

6.14 Suggested Readings/Reference Material

1. Maleh, Yassine. Shojafar, Mohammad. Alazab, Mamoun. Baddi, Youssef. Machine Intelligence and Big Data Analytics for Cybersecurity Applications (Studies in Computational Intelligence, 919) 1st ed. 2021 Edition.
2. Ahmed, Syed Thouheed. Basha, Syed Muzamil. Arumugam, Sanjeev Ram. Patil, Kiran Kumari. Big Data Analytics and Cloud Computing: A Beginner's Guide, 2021.
3. Saleem, Tausifa Jan. Chishti, Mohammad Ahsan. Big Data Analytics for Internet of Things 1st Edition.
4. Jones, Herbert. Data Science: The Ultimate Guide to Data Analytics, Data Mining, Data Warehousing, Data Visualization, Regression Analysis, Database Querying, Big Data for Business and Machine Learning for Beginners Hardcover – 10 January 2020.
5. Maheshwari, Anil. Data Analytics Made Accessible: 2023 edition Kindle Edition.
6. Mayer-Schönberger, Viktor. Cukier, Kenneth. Big Data: A Revolution That Will Transform How We Live, Work, and Think Paperback – October 26, 2021.

6.15 Answers to Check Your Progress

1. (d) Statistics

Predictive analysis uses statistical methods to validate the data assumptions and hypothesis.

2. (e) Cloud

SaaS dashboard means the expression of BI in the cloud. The dashboards summarize the key facts about the business for easy consumption.

3. (d) **High-level executives**

The demands for mobile device access to BI apps are primarily coming from high-level executives. They want the ability to explore and solve problems or pursue opportunities. Many executives are leaving their desktop and laptop PCs in favor of a tablet device even when they're in the office.

4. (c) **Internet**

The Internet provides a good venue for crowdsourcing since individuals tend to be more open because they are not being physically judged.

5. (d) **Crowd voting**

Crowd voting means gathering opinions and judgments from large groups on a certain topic.

6. (a) **Local domain**

Inter-firewall means multiple firewalls in the same Enterprise network. It provides security to the local domain and applications. It also protects from internally generated traffic.

7. (b) **Initial business review**

The activity ‘Initial Business Review’ performs validating and ranking the technology by talking with the business owners.

8. (b) **Megabytes**

A large file with images, videos, and audio contains millions of bytes, therefore it is measured in megabytes.

9. (c) **1000^3**

1 Gigabyte = 1000^3 bytes = 1000000000 bytes.

10. (c) **Data marketplace**

A place where people can buy and sell data online is called a data marketplace.

Unit 7

Cloud Computing and Big Data Management for Decision Making

Structure

"No great marketing decisions have ever been made on qualitative data."

– By John Sculley, CEO of Apple Inc.

7.1 Introduction

Good business decisions can be taken only based on quantitative data. Big Data analytics enable businesses to take better decisions.

In the previous unit, we discussed predictive analytics, software as a service BI, the role of mobile business intelligence in big data analytics, crowdsourcing and inter and trans firewall analytics.

Big data has become a critical challenge of the present time. It is created by the accumulation of data sets over years as well as the generation of voluminous data because of e-commerce, retail growth etc. It is normally defined in terms of volume, velocity and variety. Therefore, handling big data needs special technology like Hadoop, SAP HANA, etc.

In business, the related data is generated from customers, competitors and the companies themselves. It is either transactional or decision-making-based. The data can be both quantitative and qualitative. Technological evolution has made "data capturing and its analysis" easy and convenient. Any volume of data can be captured and warehoused. It is up to the decision-makers to decide whether to use data or to lose it. The data available is structured, semi-structured and unstructured and is available in many forms such as numeric text, tables, graphs, etc.

In this unit, complete working methods with big data, moving from operational dashboards to evidence-based decision making and the role of cloud computing in big data management are discussed.

7.2 Objectives

After going through this unit, you will be able to:

- Explain the meaning of big data and its elements
- Describe the challenges of big data implementation
- Explain analyzing and managing big data
- Discuss evidence-based decision making using big data
- Define the role of cloud computing in big data management

7.3 Big Data Features

Business houses and science-based institutions have been using data for a very long time. They have been using very small to huge databases. But the present-day technologies like cloud storage and cloud computing facilitate the storage of huge volumes of data. The data dealt is in petabytes, exabytes, zetta and yottabytes. The data in petabytes and exabytes is usually referred to as big data.

Doug Laney of Gartner group defined big data as big data is high volume, high velocity and/or high variety information that requires new forms of processing to enable enhanced decision making, insight discovery and process optimization".

Velocity in the above definition means the speed at which the data is handled.

Internet of Things (IoT) and big data are closely intertwined. Although they are not the same, it is very difficult to talk about one without the other.[1]The internet of things (or industrial internet) operates at a machine-scale, by dealing with machine-to-machine generated data. This machine-generated data create discrete observations (e.g., temperature, vibration, pressure, humidity) at very high signal rates (1,000s of messages/sec). Added to this, the complexity that the sensor data

[1] https://jaxenter.com/relationship-between-iot-big-data-138220.html

values rarely change (e.g., temperature operates within an acceptably small range). However, when the values do change the ramifications, the changes will likely be important.

Consequently, to support real-time edge analytics, we need to provide detailed data that can flag observations of concern. We also need to ensure that it will not overwhelm the ability to get meaningful data back to the core (Data Lake), which is used for more broad-based, strategic analysis.

Brisk Insights into Handling Big Data

Prominent among the technologies that handle big data is Hadoop.

Hadoop is a distributed file system and MapReduce processing technology that stores and processes data by dividing workloads across several thousands of servers.

Precisely, it is a unique “divide and conquer” approach.

Another approach is SAP – HANA (High-Performance Analytic Appliance) which is a very powerful server with terabytes of memory and which is used to compress the data as one unit. It uses a brute-force method. The dataset is compressed in memory and analytics is performed on it.

Further, to bring meaning out of big data, one needs to analyze this voluminous data through the latest technologies and tools.

Challenges of Big Data

Usually, the saying is that big data is for computers and small data is for people. To understand the challenges of big data, one has to get into the characteristics of big data.

The definition given by John Naisbitt makes us aware of the challenges of big data.

“We have, for the first time, our economy based on a key resource (information) that is not only renewable but self-generating. Running out of it is not a problem, but drowning in it is.”

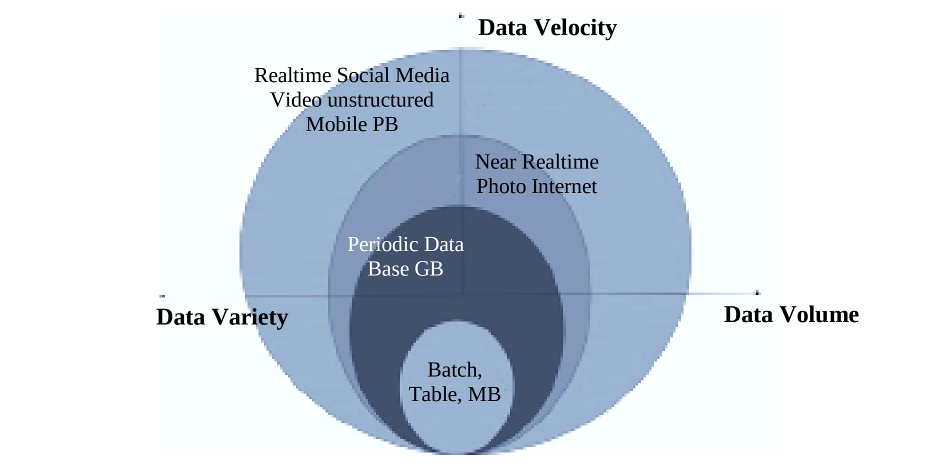
Characteristics of Big Data

Let us spend some time understanding the definition of big data given by Gartner business consulting.

Part of the definition is as follows (see Figure 7.1):

“Big data is high volume, high velocity and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization”.

Figure 7.1: Big Data Characteristics



Source: ICFAI Research Center

Example: How Tata Motors Partnered with Tata Elexis to Create a IOT Based Platform

Tata Motors had introduced "connected vehicle Platform" to introduce next generation customer experiences. This platform was based on IOT (Internet Of Things) devices gathering sensor data and analysing using Big Data analytics tools to create the enhanced experience to the customers. Unlike earlier, the customer experience was not just at the time of purchase. The vehicle upgradation could also be done through this platform without reference to year and model.

Source: https://www.tata.com/newsroom/business/driving-connected-experiences-tata-motors (December 2020), Accessed on 12/08/2022

7.4 The Challenges of Big Data Management

IEEE Report of 2012 by Rick Merritt[2]points out that data has grown 10-fold between 2010 and 2015 and multi-fold subsequently. This will raise issues of technological concerns — both software and hardware — to handle data for storage, representation and analysis.

Let us discuss the challenges of big data management. Data offers huge insights and advantages for decision makers. But terabytes and petabytes of data will be pouring in day-in-day-out. Conventional methods of data analysis have proved to be of no use in bringing out the information from these data mountains. IT teams are burdened with ever-growing requests for data and ad-hoc reports. Data visualization is gaining popularity since it gives a quick view and understanding of the data.

[2] https://www.eetimes.com/document.asp?doc_id=1262205

Organizations that are handling big data are relying on statistics, choosing an appropriate sample of the population and then the sample is analyzed. The results were attributed to the population based on the behavior of the selected sample, which is not a tested (conventional) method for drawing inferences about the population.

Companies are using SaaS (software as a service) visual analytics to represent the data in a graphical and pictorial form which gives the decision-makers a better understanding of the voluminous data.

The following are the challenges of big data in brief:

- **Matching the Speed of Results/ Inferences**

  Analysis needs to be performed on big data with a very quick turnaround time which was not possible by conventional methods. Some companies are using hardware capabilities like huge memory and powerful parallel processing to crunch large volumes of data fairly and quickly. Another method used is to apply grid computing where several machines are used for computing. Both the methods are used to analyze and present data with a quick turnaround time almost in real time.

- **Clear and Comprehensive Understanding of the Data**

  To get meaning out of data, one needs to have excellent domain expertise and an understanding of multiple methods of analyzing available data. People who are handling the analysis of big data must have a fair understanding of the incoming data. They also need to be cognizant of the sources of the data, who the end-users are and what will they look forward to inferring from the data, etc. This is, as well, a requirement for any data analysis situation, specifically when the data is large.

- **Quality of Data**

  One is aware of the fact that garbage-in will result in garbage-out. Good quality data will only ensure good analysis and proper inferences. We need to use proper statistical tools, techniques, data governance and information management system. It is always better to have proper data quality checking mechanisms so that good quality data is always in.

- **Discrepancies in the Data**

  As part of data quality, one usually encounters missing values, outliers, etc. Missing values might occur in the collection or at the entry point. Since data is huge, a 1% missing value is also a big issue in data analysis. One may use the method of averages or interpolations. Outliers are the numbers that fall outside the normal range if the data is numeric. If the data is qualitative, the outliers need to be treated properly. Even a 1% deviation is quite a good volume and is critical for analysis and inferences. One may need to segregate outliers and analyze them separately.

- **Inferences about the Data**

  Representing and drawing inferences about the data is the key to decision making. The Use of visual representations with the latest software and the use of advanced statistical tools and techniques (multivariate data analysis techniques) help the decision-maker to understand and gather actionable intelligence.

  To conclude on the challenges of big data, the three Vs, apart from the veracity and other data-related issues, are best addressed by a combination of advanced software and improved hardware.

Example: Yamaha Partners with Oracle and Loqate to Ensure Higher "Quality of Data" for Analytics to Improve Customer Experience

Yamaha had serious problems with customer data in terms of duplicates, incorrect addresses, or multiple addresses etc. This was hampering data analysis and hence customer experience. The company worked with Oracle and Loqate to address this big data challenge. After the new project was in place, the email verification software verified real time verification of emails so as to get clean data. False registrations and wrong deliveries were reduced and Physical address verification was also accomplished.

Source: https://www.loqate.com/en-us/customers/yamaha-motor-oracle/, 2022, Accessed on 12/08/2022

7.5 Analyzing and Managing Big Data

Let us start this section with some definitions given by experts on data, statistics, and knowledge.

"Data is just like crude. It's valuable, but if unrefined it cannot be used".

Michael Palmer (2006)

(Source: http://ana.blogs.com/maestros/2006/11/data_is_the_new.html)

The above definition, an expanded version of Humby's quote (*"Data is the new oil"*), emphasizes the need for analyzing data.

"The need for three Rs in reading, writing and arithmetic is well understood. These do not take us far unless we acquire the fourth R, reasoning under uncertainty for taking decisions in real life".

Prof. C.R. Rao (2014)

(Source:www.crraoaimscs.org/Stat_Day_2014_TJR)

This definition given by Prof. C.R. Rao reiterates the need for data analysis, particularly large volumes of data.

Walmart handles more than 1 million customer transactions per hour[3]. Facebook deals every day with 2.7 billion likes, 300 million photos uploaded and

[3] http://www.grabstats.com/stats/2036

500 terabytes of data[4]. Decoding the human genome is taking a week now when compared to the 10 years taken earlier.

Big data is often managed by hybrid cloud technology because of the need to analyze huge volumes of data. It has three characteristics —volume, variety, and velocity —as discussed earlier.

The following Table 7.1 gives an idea about terabytes and petabytes.

Table 7.1: Big Data Numbers

Name	Byte	Gigabyte	Terabytes	Petabytes
Value	10^0 Bytes	10^9 Bytes	10^{12} Bytes	10^{15} Bytes

Big data comes in different shapes and sizes:

- **Structured data:** Most of the analysts used to deal with structured data which can be included in the database. For example, the number of items sold and the number of customers who visited the outlet.
- **Semi-structured data:** It has some structure but not in the form of tables as in a database; it includes EDI (Electronic Data Interchange) formats and XML (Extensible Markup Language) formats, etc.
- **Unstructured data:** It includes text, images, audio, documents, e-mails, tweets, blogs, etc. Un- structured data amounts to 80% of the data in use in the current time.

While analyzing big data, a few terms are heard very commonly. Text analytics, sentiment analytics, voice analytics, moment analytics and machine to machine analytics.

7.5.1 Analysis of Big Data

Big data is analyzed for a better understanding of the customer and also to identify target customers. Retailers can predict what products will have maximum sales. Optimizing the business can be done by synthesizing the data of customers from social media and make decisions on whether to maintain an inventory of certain goods.

One of the problems in dealing with large amounts of data is that it is not always easy to identify patterns. Thus we need data visualization software, which helps in converting streams of text and numbers into graphical representations that reveal a greater picture of the data.

Association rule mining is a methodology that is used to discover unknown relationships hidden in big data. Rules refer to a set of identified frequent item sets that represent the uncovered relationships in the dataset. The underlying idea is to identify rules that will predict the occurrence of one or more items based on

[4] https://gizmodo.com/5937143/what-facebook-deals-with-everyday-27-billion-likes-300-million-photos-uploaded-and-500-terabytes-of-data

the occurrences of other items in the dataset. It is an unsupervised machine learning method. This means that no direct guiding output data is given to find the patterns.

7.5.2 Management of Big Data

In many commercial environments, large quantities of data are accumulated in databases from day-to-day operations. This lays the foundation for mining association rules. In retail, for example, "customer purchase data" is collected daily at the checkout counters of city stores or from online shopping stores. The accumulated data items are often market basket transactions. Managers of stores are interested in analyzing the collected data to learn the purchasing behavior of customers. This enables a large variety of business-related applications based on the identified rules in the data.

Sequential pattern mining is a data mining task specialized for analyzing sequential data, to discover sequential patterns. More precisely, it consists of discovering interesting sub-sequences in a set of sequences. Here the interestingness of a sub-sequence can be measured in terms of various criteria such as its occurrence frequency, length and profit. Sequential pattern mining has numerous real-life applications since data is naturally encoded as sequences of symbols in many fields such as bio-informatics, e-learning, market basket analysis, texts and webpage click-stream analysis.

Big data use

When probed with powerful visualization tools, big data can produce a wide variety of insights allowing to test relationships between multiple variables, detect anomalies and compute the statistics from basics to advanced level for prediction.

So far we have discussed big data and issues of big data analytics. We have also briefly discussed the data analysis tools and techniques, from the basic to the advanced level.

Big data analytics makes use of advanced tools like data mining, statistics, predictive analytics and natural language processing. They are used against large volumes of data. It allows analysts, researchers and business users to make better and faster decisions.

7.5.3 Insights into Tools that are used for Analysis of Big Data

Hadoop is an open source, Java-based programing framework that supports the processing and storage of extremely large data sets in a distributed computing environment. It is part of the Apache project sponsored by the Apache Software Foundation. Hadoop – native data processing and analysis options- include the following:

- Apache Hive (provides data accessing and data warehousing similar to Structured Query Language -SQL),

- Apache Mahout (supports machine learning on top of Hadoop– for finding patterns in data),
- Apache MapReduce(for searching, filtering and sorting–ways to generate useful nuggets from big data), and
- Apache Pig (a language for writing MapReduce jobs).

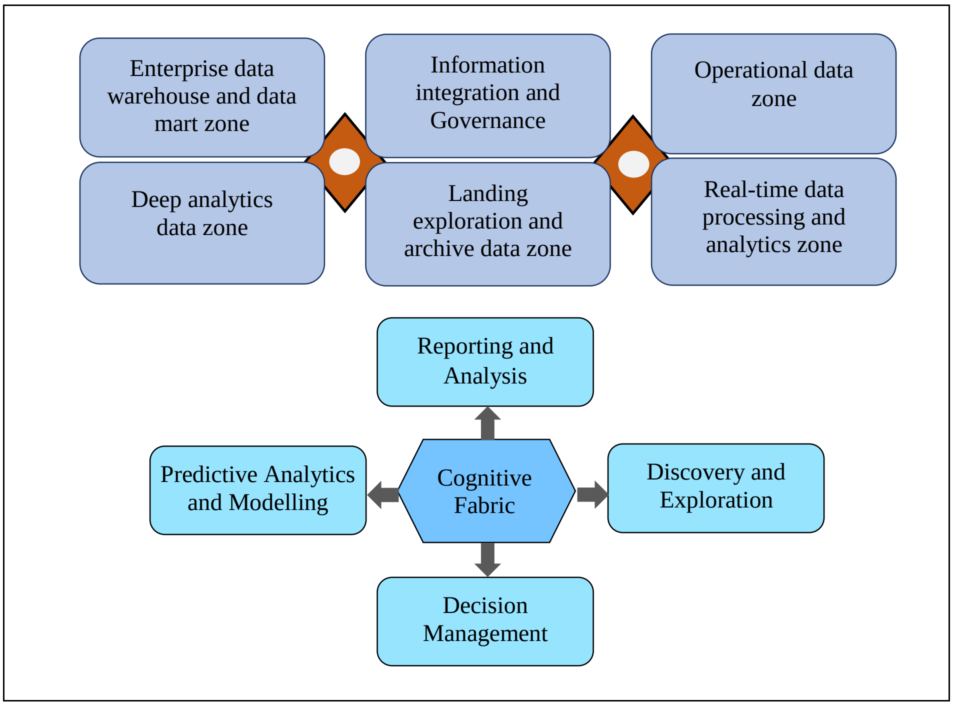
Alternative SQL access/analysis options: Hive is slow by relational database standards and it does not support all SQL-analysis capabilities. The following will give an advantage of working on big data and SQL operations, Teradata, Query Grid, Oracle, big data, SQL, IBM, and SQL.

Analytics and BI (business intelligence) options designed to run on Hadoop: These tools are a blend of SQL and BI. This has the facility of querying with big data and data-oriented advanced analytics capabilities. Examples include Apache Spark, Apache Strom, SAS Visual Analytics and Data Meer. Many of these analytic engines run on Hadoop 2.0's YARN resource management system.

The value of big data analysis is often used to find correlations among disparate data sets or insights hidden in semi-structured or highly variable data sources. The tools are used to summarize, compare, represent and forecast customer behavior or customer-related queries.

Figure 7.2 details IBM's next-generation architecture for big data analytics and usage of big data solutions.

Figure 7.2: IBM's Next Generation Architecture for Big Data Analytics


Enterprise data warehouse and data mart zone
Information integration and Governance
Operational data zone
Deep analytics data zone
Landing exploration and archive data zone
Real-time data processing and analytics zone
Reporting and Analysis
Predictive Analytics and Modelling
Cognitive Fabric
Discovery and Exploration
Decision Management

Source:ICFAI Research Center

Example: Wells Fargo Is Planning to Use Quantum Computing to Speed Up the Modelling and Analytics for Fraud Detection

Fraud detection methods currently deployed by Wells Fargo and other financial institutions was creaky and the result was weeks long delays in onboarding new customers. The analytics needed high speed computing power. The bank was working with MIT scientists to see how quantum computing can enhance the computing power and reduce the waiting times. This would have a marked effect on the bank's operational efficiency.

Source: https://www.cio.com/article/400745/wells-fargo-prepares-to-(-a-quantum-leap.html (June 17, 2022) Accessed on 13/08/2022

Activity 7.1

Big Data on Cloud Platform

A private hospital has a chain of orthopedic clinics across Mumbai. It manages its surgical equipment inventory using an existing IT solution that is well-connected over the data network. But to handle complex surgeries, it requires certain imported surgical equipment. It approaches IT solution providers for developing larger software to handle its imported equipment operations electronically and to analyze cost and transport requirements from a pool of suppliers.

Mr. Ram Sastry, who is the project manager, gave Cloud Platform's big data application a suitable alternative. If you were the technical lead, how would you support his choice to convince the client?

Answer:

Check Your Progress - 1

1. Which of the following is not a characteristic of big data?
   a. Volume
   b. Velocity
   c. Variety
   d. Totally Structured Data
   e. Veracity

2. Which of the following is an example of unstructured data?
   a. Database
   b. Word File
   c. Doctor’s Prescription
   d. File Management System
   e. Note Pad Text Document

3. Which of the following is the right set of challenges for big data?
   a. Matching the Speed of Results/Inferences, Clear and Comprehensive Understanding of the Data, Quality of Data, Discrepancies in the Data, Size about the Data
   b. Matching the Speed of Results/Inferences, Clear and Comprehensive Understanding of the Data, Quality of Data, Volume of the Data, Inferences about the Data
   c. Matching the Speed of Results/Inferences, Clear and Comprehensive Understanding of the Data, Quality of Data, Discrepancies in the Data
   d. Matching the Speed of Results/Inferences, Clear and Comprehensive Understanding of the Data, Quality of Data, Discrepancies in the Data, Inferences about the Data
   e. Matching the Speed of Results/Inferences, Clear and Comprehensive Understanding of the Data, Discrepancies in the Data, Inferences about the Data

4. Which of the following visualization term refers to large volumes of data?
   a. Archiving
   b. Pooling
   c. Organizing
   d. Viewing graphically
   e. Retrieving

5. In mathematical terms, how much is the size of petabyte?
   a. 10^{12}
   b. 10^{13}
   c. 10^{15}
   d. 10^{24}
   e. 10^{18}

7.6 Big Data Storage Considerations

Every organization generates some amount of data in some form or other and many others in the organization use the data as per their requirements. Proper data storage is a huge challenge. The following will take you through the data storage evolution.

7.6.1 Data Storage Evolution

The evolution in data storage has been marked from magnetic tapes of the 1920s to floppies, CDs, DVDs, pen drives, hard disks to holographic data storages. Present-day organizations are using cloud-based storage in which businesses are assured of security for their data, disaster recovery and faster access due to high bandwidth.

The issues concerning data storage of big data have been initiated in the earlier section. Moving further, the currently available internet speed and infrastructure raise storage and transport issues, as it takes a long time for data transfer. This is addressed by transferring only the required data and running the code. Only the required details need to be retrieved.

7.6.2 Data Storage Issues

Data storage and management issues like access, utilization, updating, governance and reference are the major stumbling blocks. The sources of the data are varied in terms of size and format. The method of collection, unlike the collection of data using manual methods, involves some rigorous protocols. To ensure accuracy and validity, digital data collection has become much more flexible and methodologies are being developed. Other issues in big data storage are privacy and security. The most important thing is, it includes conceptual, technical as well as legal significance. The personal information of an individual is combined with external data sets. The person may not prefer to share this secretive information with others.

7.6.3 Data Accessing and Sharing

Another important issue is data accessing and sharing. In the present day’s open society, anybody including the government and regulatory agencies can access information from any corner. Thus, the agencies which manage and maintain data should be very careful in giving access to the databases. Besides, one needs to be cognizant of data ownership issues. From the point of human resource issues, big data is an upcoming technology and people need to be trained in this area to handle the storage and manage it.

Technical Challenges: Following are the technical challenges that accompany big data:

i) *Fault tolerance:* With the latest technologies like cloud computing, there could be a failure of servers. The failure must be within acceptable levels and it should not lead to starting from scratch.

ii) *Scalability:* The scalability issue of big data led to cloud computing. Cloud computing aggregates disparate workloads with varying performance goals into very large clusters, and the respective storage requires the latest hardware technology like solid-state drives.

iii) *Quality of Data:* Quality of the data is another issue. Big data is used for predictive analytics and other advanced analysis, hence it requires good quality data. Quality checking is a big challenge and needs a robust methodology.

Example: PostFinance (Swiss Bank) Deploys Cloud Based Cloudian Hyperstore Object Storage for Scalability and Cost Reduction

PostFinance offers provided banking services to 30 lac customers. The bank aimed to provide seamless customer experience across the online and offline channels. The bank was using Hadoop for storage for big data analytics, but it was not able to scale up as per growing data. The bank went with Cloudian Hyperstore Object storage whereby the bank could index, store, and analyse more big data.

Source: https://cloudian.com/resource/case-studies/postfinance/, 2022, Accessed on 13/08/2022

7.7 Moving from Operational Dashboards to Evidence-Based Decision Making

The present era is of big data. People believe that big data = big opportunity and big money.

"Concepts without percepts are empty; percepts without concepts are blind".

-German Philosopher
Immanuel Kant (~1780)

This quote gives us a hint that neither intuition/theory (concepts) nor data (percepts) can exist on their own. They need to be combined scientifically for getting the meaning out of the data given.

Operational dashboards are informative templates that give quick information on the subject to the concerned decision-makers. Dashboards have become front-end and the first line of access for business intelligence. The best way to get insights into an organization and its performance is through dashboards.

For example, looking at sales operations, looking into the performance of salespeople, individual product categories, understanding customer behavior, and demographics.

These can be monitored together or separately as part of broader initiatives. The value this brings to the business is significant. Once companies gain regular insights into these performances, they dig deeper into the data. For instance, sales of products/services can identify (1) what products or services are most successful, (2) how to drive sales based on the leads, (3) how demographics is related to sales, etc. An integrated approach to building solutions for the decision-making using the data from multiple sources will help the organizations better.

Areas where organization's benefit from the dashboards include:

- Time-saving.
- Money saving.
- Insight into customer behavior.
- Aligning strategy with tactics.
- Ensuring a widespread, goal-driven approach and
- Performance culture.

Let us now examine the operational and analytical dashboards.

Operational dashboards manage internal and intra-daily business processes, frequently changing current performance metrics or key performance indicators. The areas span from sales, marketing, helpdesk and supply-chain. Overall operational dashboards are best suited to departments that require low latency data feeds and a continual view into happenings within the business unit. In short, operational dashboards are meant to help an organization understand if its performance is "on or off" the target and how much it is on/off the target in real-time.

Analytical dashboards focus more on gaining insights from a volume of data collected over time –last month or last quarter – and use this to understand what happened. This helps in deciding, why and what changes should be made for the future. For instance, organizations may want to compare trends over time or products performance across regions. Companies want to measure the success of marketing campaigns by combining sales data with product placements and marketing campaign strategy. Analytical dashboards use sophisticated advanced statistical models with what-if analysis. These dashboards are regularly used by business analysts and experts, who are responsible for preparing the reports for general consumption. In short, analytical dashboards are meant to help an organization visualize the real-time key metrics and performance indicators. Dashboards give insights into a situation.

7.7.1 Evidence-based Decision-making

Evidence-based decision-making will facilitate the decision-maker to make decisions based on data/information.

Business success depends on gaining new insights faster than the competition and turning these insights into good decision-making, which in turn depends on good data. Many organizations have huge volumes of data. But the dilemma is which data/information is to be used. This is where evidence-based management comes in. It appears simple and straight that organizations have problems in collecting reliable and relevant information. However, the technique of finding relevant data among the huge amounts of datasets is available through analytical tools.

Hence, the data can be used to turn it into information and knowledge for the benefit of the organization. Top organizations use this approach to ensure that they collect the most relevant information to support key decisions. The big data management techniques come into the picture when the right data is picked.

Relevant analytical tools can be operated on the big data for evidence-based decision-making.

The following steps are in use for evidence-based decision making:

- Define the objective - research problem - formulate the hypothesis.
- Collect the data from data warehouses of big data/cloud spaces.
- Analyze the data using big data analysis techniques.
- Validate the hypothesis and present the findings.
- Take decisions based on this information.

Big data management and analysis will give us information based on conclusions that can be used for validating the hypothesis/the business statement proposed.

The data analysis for decision making from big data requires real-time, stream computing capabilities. These enable one to combine multiple information sources to facilitate decision and initiate action. The stream computing helps business move from predefined decision paths to advanced analytical techniques making use of software like R. Most of the organizations see a decrease in storage and other infrastructure costs after implementing a real-time analytics platform.

One needs to emphasize that big data power does not erase the need for human insight. The companies that are moving from operational dashboards to evidence-based decision-making need to take caution. There is either too much evidence present or the evidence does not apply. A combination of operational dash board inputs, evidence-based inputs and decision makers' expertise leads to a better decision.

Example: AIG (Global Insurance Major) Goes for Analytical Dashboards to Move from "Gut Based" to Evidence Based Decision Making

AIG was a major global insurance company for both life and general Insurance products. The company had been taking up data analytics projects for improve its operational efficiency. But it was finding it a challenge to present the output of analytics models to be fully understandable to users at all levels so that the people can take evidence-based decision making moving away from gut based. For this, it took the help of a consultant who created world class analytical dashboards. The consultant reviewed the existing dashboards, interviewed people at various levels and understood the requirements. Then they went about re-architecting the dashboards.

Source: https://www.door3.com/our-work/aig-business-intelligence-dashboard/, 2022, Accessed on 14/08/2022

7.8 Role of Cloud Computing in Big Data Management

Till a decade back, few functionaries of the business used to generate data and the entire organization used the data. Now, many business entities are generating data. Data is growing exponentially in different formats, unstructured, structured and semi-structured.

What is Cloud Computing?

The basic concept of cloud computing is pay, use, store and be secure

Distributed computing on the internet was introduced way back in 1950 by IBM as RJE (Remote Job Entry process). In 2006, Amazon provided the first public cloud, Amazon Web Service. Usually, cloud computing has three components — client computers, distributed servers and data centers. Clients are the people who make use of the cloud facility; they usually use mobiles, thick and thin client systems. Data center is the collection of servers, where the application is placed and accessed via the internet. Distributed servers are in geographically different locations but perform as if they are next to each other. The central server administers the data traffic and attends to the client demands, etc. It functions through special software called middleware.

SaaS: Software as a Service (SaaS; pronounced /sæs/) is a software licensing and delivery model in which software is licensed on a subscription basis and is centrally hosted. It is sometimes referred to as "on-demand software", and was formerly referred to as "software plus services" by Microsoft.

PaaS: Platform as a Service (PaaS) or application Platform as a Service (aPaaS) is a category of cloud computing services that provides a platform allowing customers to develop, run and manage applications without the complexity of building and maintaining the infrastructure typically associated with developing and launching an app.

IaaS: Infrastructure as a Service (IaaS) is a form of cloud computing that provides virtualized computing resources over the Internet. IaaS is one of the three main categories of cloud computing services, alongside Software as a Service (SaaS) and Platform as a Service (PaaS).

Types of cloud deployment models in use are:

- Public Cloud
- Private Cloud
- Community Cloud and
- Hybrid Cloud

Cloud services are popular since they can be customized. One need not buy a software; it reduces the complexity of networks, enhances affordability, and its services are reliable, efficient and scalable.

The following Table 7.2 gives a comparison of cluster, cloud and grid computing.

Table 7.2: Distributed vs. Grid vs. Cloud Systems

	Clusters	*Grids*	*Clouds*
SLA	Limited	Yes	Yes
Allocation	Centralized	Decentralized	Both
Resource Handling	Centralized	Distributed	Both
Loose coupling	No	Both	Yes
Protocols/API	MPI, Parallel Virtual	MPI, MPICH-G, GIS, GRAM	TCP/IP, SOAP, REST, AJAX
Reliability	No	Half	Full
Security	Yes	Half	No
User friendliness	No	half	Yes
Virtualization	Half	Half	Yes
Interoperability	Yes	Yes	Half
Standardized	Yes	Yes	No
Business Model	No	No	Yes
Task Size	Single large	Single large	Small & medium
SOA	No	Yes	Yes
Multilatency	No	Yes	Yes
System Performance	Improves	Improves	Improves
Self service	No	Yes	Yes
Computation service	Computing	Max. Computing	On demand
Heterogeneity	No	Yes	Yes
Scalable	No	Half	Yes
Inexpensive	No	No	Yes
Data Locality Exploited	No	No	Yes
Application	HPC, HTC	HPC, HTC, Batch	SME, interactive apps
Switching cost	Low	Low	High
Value Added Services	No	Half	Yes

Source: Sadashiv, & Kumar. (n.d.). Cluster, Grid and Cloud Computing: A Detailed Comparison. Department of Computer Science Faculty of Sciences, University of Porto. Retrieved October 31, 2022, from https://www.dcc.fc.up.pt/~ines/aulas/1314/CG/Presentations/Goncalo/papers/06028683.pdf

Activity 7.2

Cloud for Archrivals

An apparel manufacturing unit in Mumbai has a number of outlets across the country. They want to develop a system to monitor CCTV footages of all their stores and store them on a cloud platform. The aim is that, if needed, they can view these footages at their headquarters to manage products, address missing issues and the related complaints with law enforcing agencies. Suggest the cloud deployment approach that can be used in order to archive the data on cloud and retrieve it when needed.

Answer:

7.8.1 Cloud Computing and Big Data Management

Adaptability of big data is increasing in the recent years due to competitive price, safety and facility to hold big data. Researchers are of the opinion that this data can be analyzed and used for better business environment and for a better society. For example, analysis of which person purchases which item on what day, analysis of the density of vehicular activity on coming week in different routes with maximum levels of accuracy etc., help many decisions in business and governance.

Big data on servers: Big data processing on clouds might involve hundreds of entities such as application servers accessing data and this will generate massive read/write requests. Hence, large numbers of servers are interconnected such that the failure of a few of the servers does not stop the entire service.

Data read write: One of the techniques is distributed data store, specifically, distributed Key Value Store (KVS).With this technique, a data structure composed of keys and values is distributed and is stored in a number of servers. Read and write take place on one server according to the request specified by a key to return its response. It offers the best performance because the load can be prevented from getting concentrated on one server even if many requests are pouring in. The chances of failures are less since data is copied and stored on many servers at a time and processed. Thus, failure of some servers does not lead to loss of data.

Complex Event Processing: Another methodology is parallelization of complex event processing (CEP). CEP refers to the technology that processes and analyzes in real-time, the complicated and massive event series those are constantly generated in real-world activities and operations. This model allows the queries to be dynamically distributed, system-wise, according to the event, stream, characteristics and load of the data. This in turn enables optimization of the processes.

Dynamic load balancing CEP has the following technical challenges:

Applying real-time, cost-effective CEP in the cloud has been an important goal in recent years. Distributed stream processing systems- DSPS have been widely adopted by major computing companies such as Facebook and Twitter for performing scalable event processing in streaming data. However, dynamically balancing the load of the DSPS' components can be particularly challenging due to:

- The high volume of data,
- The components' state management needs, and
- The low latency processing requirements.

Systems should be able to cope with these challenges and adapt to dynamic and unpredictable load changes in real-time.

One approach makes the following contributions:

(i) Formulate the load balancing problem in distributed CEP systems as an instance of the job-shop scheduling problem, and

(ii) Present a novel framework that dynamically balances a load of CEP engines in real-time and adapts to sudden changes in the volume of streaming data by exploiting two balancing policies. The detailed experimental evaluation using data from the Twitter social network indicates the benefits of this approach in the system's throughput.

In the present day’s business or real-life situations, innovations are the key differentiators. Big data and cloud computing will facilitate these innovations. The data is analyzed in multiple ways to bring out the innovations much beyond data mining.

Example: Walmart Rebuilds Applications before Moving to Cloud so as to take Full Advantage of the Cloud Infrastructure

Walmart used hybrid cloud for its ever-expanding IT infrastructure requirements. It had its own data centres in private cloud. But it also used Google Cloud for some of its applications where the company felt that use of public cloud was beneficial for some of the applications in view of scalability and other benefits. As a strategy, they were not moving legacy applications to the cloud. They were rebuilding the apps to take the benefit of special features available in the public cloud.

Source: https://www.ciodive.com/news/walmart-cloud-strategy-investors/573175/ (Feb 28, 2020), Accessed on 14/08/2022

Check Your Progress - 2

6. What does CEP stand for?

   a. Communication Engine Processing

   b. Complex Element Processing

   c. Complex Event Processing

   d. Computerized Event Processing

   e. Customized Event Processing

7. What is the time required for data migration in a grid-based cloud implementation?

   a. Hours

   b. Days

   c. Weeks

   d. Months

   e. Minutes

8. Which of the following is not a cloud deployment model in use?

   a. Corporate Cloud

   b. Public Cloud

   c. Private Cloud

   d. Community Cloud

   e. Personal Cloud

9. Which of the following refers to The Key Value Store?

   a. Multiple databases

   b. Multiple applications

   c. Multiple sources of data

   d. Multiple servers

   e. Cloud-enabled environment

10. Which of the following is not services and delivery model in Cloud?

    a. SaaS

    b. PaaS

    c. IaaS

    d. gPaaS

    e. MaS

7.9 Summary

- Big data is a collection of structured, semi-structured and unstructured data which has volume, variety and velocity as its characteristics.
- Big data is a set of tools and techniques that require innovative/latest methods of integration to find hidden patterns from large data sets that are complex, diverse and massive.
- Qualities of data, understanding its nature, the inconsistencies of big data are major concerns while consolidating and integrating it on a single platform for analysis.
- The ability of a number of BI (business intelligence) and analytical tools to manage large volumes of data is restricted. This include like Apache Spark, Apache Storm, SaaS Visual Analytics and Data Meer.
- A big data case study to highlight the benefits of implementing a prediction mechanism for forecasting product demand of bakery product manufacturer in Europe.

7.10 Glossary

Business Intelligence: It is a technology-enabled data analyzing and presentation method for displaying information to help business executives and managers in making decisions.

Dashboard: It is a user interface which organizes and presents information in an understandable format such as a graph. Usually, dashboards are helpful in drawing inferences from the graphics, useful in the decision-making process.

Hadoop: It is an open-source distributed data processing framework to handle big data used in the cloud environment. It is sponsored by the Apache Software Foundation and is used for executing applications on systems with a large number of nodes without interruption.

Predictive Analytics: Predictive analytics is a method for extracting information from existing data sets in order to identify patterns and predict future outcomes and trends.

Unstructured Data: Any information which does not have a pre-defined structure to organize itself is called unstructured data. The unstructured content is generally text-based documents like building designs and medical machine prescriptions.

7.11 Self-Assessment Test

1. Briefly discuss the different characteristics of big data with suitable examples.
2. Discuss how big data will be helpful in managing the retail sector.

3. Give a few examples of structured and unstructured data found at a hospital.
4. Explain the challenges of managing big data.
5. List a few BI and analytical tools used on a cloud platform.

7.12 Suggested Readings/Reference Material

1. Maleh, Yassine. Shojafar, Mohammad. Alazab, Mamoun. Baddi, Youssef. Machine Intelligence and Big Data Analytics for Cybersecurity Applications (Studies in Computational Intelligence, 919) 1st ed. 2021 Edition.
2. Ahmed, Syed Thouheed. Basha, Syed Muzamil. Arumugam, Sanjeev Ram. Patil, Kiran Kumari. Big Data Analytics and Cloud Computing: A Beginner's Guide, 2021.
3. Saleem, Tausifa Jan. Chishti, Mohammad Ahsan. Big Data Analytics for Internet of Things 1st Edition.
4. Jones, Herbert. Data Science: The Ultimate Guide to Data Analytics, Data Mining, Data Warehousing, Data Visualization, Regression Analysis, Database Querying, Big Data for Business and Machine Learning for Beginners Hardcover – 10 January 2020.
5. Maheshwari, Anil. Data Analytics Made Accessible: 2023 edition Kindle Edition.
6. Mayer-Schönberger, Viktor. Cukier, Kenneth. Big Data: A Revolution That Will Transform How We Live, Work, and Think Paperback – October 26, 2021.

7.13 Answers to Check Your Progress Questions

1. **(d) Totally Structured Data**

   Big data is distributed and is a mix of both structured and unstructured data.

2. **(c) Doctor’s Prescription**

   A medical prescription is a classic example of unstructured data as it is generally handwritten.

3. **(d)** Matching the Speed of results/inferences, clear and comprehensive understanding of the data, quality of data, discrepancies in the data, inferences about the data.

4. **(d) View Graphically**

   Visualization refers to the viewing of large volumes of data in graphics mode for better understanding.

5. (c) **10^{15} bytes**

Petabyte is a memory measuring metric which accommodates 10^{15}bytes of content.

6. (c) **Complex Event Processing**

Complex Event Processing (CEP) refers to technology that processes and analyzes in real-time, complicated and massive event series that are constantly generated in real-world activities and operations.

7. (a) **Corporate Cloud**

Types of cloud deployment models in use include: public cloud, private cloud, community cloud and hybrid cloud. Corporate cloud is not a cloud deployment model in use.

8. (e) **Personal Cloud**

All other options are existing cloud options.

9. (e) **Cloud-enabled environment**

In a cloud-enabled environment, one of the techniques is distributed data store, specifically, distributed key value store (KVS). With this technique, a data structure composed of keys and values is distributive and is stored in a number of servers, and "read and write" take place in one server according to the request specified by a key to return its response.

10. (e) **MaS**

All other options are service models in cloud.

Unit 8

Handling Unstructured Data

Structure

"The skill of data storytelling is removing the noise and focusing people's attention on the key insights."

- Brent Dykes, data strategy consultant

8.1 Introduction

Big data is not about collecting, storing, and analysing huge data of various forms. The output should be presented to the end users in an understandable form providing critical insights which can be acted on.

In the previous unit, we have briefly looked at decision-making using big data. However, handling big data itself is a rigorous task that needs a careful categorization of data to understand how it must be handled. In this unit, we will learn the handling procedure of unstructured data.

Big data is made of both machine and human-generated unstructured, semi-structured and structured data. Unstructured data include social media posts, images of data like fingerprints, email messages, weblogs on the internet and many other contents. It is growing at an unprecedented pace, contributing to huge volumes of content.

Small volumes of unstructured data can be handled easily. But as the unstructured data becomes large, its analysis becomes quite complicated because of the heterogeneity of the data. The solution for this is to adopt available IT solutions. Basically, unstructured data is divided into two broad categories, textual and non-textual. Textual data include all the documents, e-mails, text files and chat files. The non-textual data covers graphics, sound, movies, etc. Analytics are applied to the data for analysis purpose. Such data will support the decision-making process in areas like retail business, e-commerce, education, healthcare and marketing domain. In addition, the IoT (internet of things) is growing and generating much larger volumes of data. To manage, store and provide security to such huge volumes of data is a complex task to handle. This unit concentrates on issues relating to unstructured data management practices. It describes various forms of data generated on social media and mobile to be analyzed and stored using cloud technology. The cost of storage, backup and recovery operations incurred by organizations are also discussed. This unit also gives an insight into different file archiving solutions along with the problems that exist for analyzing textual and non-textual content.

8.2 Objectives

After going through this unit, you will be able to:

- Explain the different formats of unstructured data that are generated in organizations with examples
- Discuss the mechanisms used for storing unstructured data
- Explain different ways of analyzing and managing unstructured data
- Discuss file archiving solutions possible to handle unstructured data
- Explain the opportunities and problems with unstructured data

8.3 Different Formats of Unstructured Data

Data is the heart of any business activity and it is maintained by users in different formats based on the requirement. For example, it may range from a person's name you want to remember, an address entry in a diary or a diagnostic lab report. When this data is organized in the form of a report, it is known as information (i.e. data with a meaningful purpose).

There are two basic types of information, based on the structural standard, that are being followed. The first category is structured data, which is generally stored in a database based on a predefined format, known as metadata. The second category data is the unstructured and semi-structured one, which do not have a defined structure of its own.

Unstructured data: Unstructured data is everywhere. In fact, most individuals and organizations conduct their lives around unstructured data. Just as with structured data, unstructured data is either machine-generated or human-generated.

Here are some examples of machine-generated unstructured data:

- **Satellite images:** This includes weather data or the data that the government captures in its satellite surveillance imagery. Just browse Google Earth, and you get the picture.
- **Scientific data:** This includes seismic imagery, atmospheric data and data on high energy physics.
- **Photographs and videos:** This includes security, surveillance and traffic videos.
- **Radar or sonar data:** This includes vehicular, meteorological, oceanographic and seismic profiles.

Different formats of unstructured data include:

i) **Text Data:** Most of the available information in companies is in the form of text documents stored in text databases. These can be a large collection of documents originating from multiple sources like research papers, books, digital libraries, and e-mail messages. Nowadays most of the information in government, industry, business, and other areas is stored electronically, in the form of text databases. For example, Word Docs, PDFs, letters, and other written documents.

ii) **Audio Files:** Many companies provide IVR (interactive voice response) facilities to keep track of sales and service requests. All such calls made by customers are recorded by the call center staff to provide better service. For example, customer-care IVR recordings, voicemails, emergency phone calls are all audio recordings.

iii) **Videos:** Many online portals provide free storage space to upload videos which can be shared among friends and family members. For example, WhatsApp, video clips, personal video recordings and YouTube uploads.

iv) **Presentations:** Companies maintain quarterly performance-based presentations for projecting the quality assurance needs and for reporting to top management and stakeholders. Some companies store these presentations as a knowledge management resource for later use. Similarly, presentations are also made to train executives and promote new products and services. All these resources are important and they need to be stored for future use and reference.

v) **Images:** Pictures and illustrations are used by companies to showcase the working of their products. They are also used for marketing on social media and for sending promotional emails to prospective customers.

vi) **Messaging and Email:** It is the cheapest and most popular tool used by companies to promote, communicate, interact and serve customers during pre-sales, sales and post-sales activities.

vii) **Web pages:** Companies conduct web contests, surveys and a number of promotional activities on the internet, to engage customers and the public in their products/services. Other contents like newsletter, reminders, seasonal greetings, reward, and offers are also used to build the relationship with the users.

Unstructured data is generated by a variety of applications in different formats as discussed above. All these contents pile up to very large quantities of data, making it difficult for managers to manage its storage, archiving and recovery operations.

The following list shows a few examples of human-generated unstructured data:

- **Text internal to your company:** Think of all the text within documents, logs, survey results and e-mails. Enterprise information actually represents a large percent of the text information in the world today.
- **Social media data:** This data is generated from the social media platforms such as YouTube, Facebook, Twitter, LinkedIn and Flickr.
- **Mobile data:** This includes data such as text messages and location information.
- **Website content:** This comes from any site delivering unstructured content, like YouTube, Flickr or Instagram.

Semi-structured

Semi-structured data can contain both the forms of data. We can see semi-structured data as structured in form but it is actually not defined like a table definition in relational DBMS. Examples of Semi-structured Data are personal data stored in an XML file-HTML, RDF

Refer Figure 8.1 below for big data categories. Most of the data that is generated by organizations is unstructured in nature, amounting to nearly 80% of the total data volume. Examples of unstructured data also include specialized data such as spreadsheet files, word documents, graphics files, XML files-semi structured data and PDF files.

Figure 8.1: Big Data Categories with Examples

	Structured Data	**Unstructured Data**
Characteristics	• Pre-defined data models • Usually text only • Easy to search	• No pre-defined data model • May be text, images, sound, video or other formats • Difficult to search
Resides in	• Relational databases • Data warehouses	• Applications • NoSQL databases • Data warehouses • Data lakes
Generated by	Humans or machines	Humans or machines
Typical applications	• Airline reservation systems • Inventory control • CRM systems • ERP systems	• Word processing • Presentation software • Email clients • Toosl for viewing or editing media
Examples	• Dates • Phone numbers • Social security numbers • Credit card numbers • Customer names • Addresses • Product names and numbers • Transaction information	• Text files • Reports • Email messages • Audio files • Video files • Images • Surveillance imagery

Source: Taylor. (2021, May 21). Structured vs. Unstructured Data. Datamation. https://www.datamation.com/big-data/structured-vs-unstructured-data/

Example: How Swiss Re Used Analysis of Unstructured Data to Assess Flood Damage

Assessing damage due to floods and arriving at accurate estimates of insurance claims was a major challenge for insurance companies as the surveyors could not reach the actual locations. In view of this, the company could face excessive insurance claims. The company deployed a platform for analysis of satellite images (unstructured data) using analytics to accurately assess the quantum of claims to be made. This had led to faster claim settlement for the insured and savings for the company by avoiding excessive claims.

Source : https://www.swissre.com/risk-knowledge/mitigating- -risk/remote-sensing-technology-in-claims-assessment.html (23 Sep 2021), Accessed on 15/08/2022

8.4 Key Issues in Storing Unstructured Data

New technologies along with the best business processes are being adopted by organizations in different fields like education, manufacturing and healthcare due to growing competition. This has resulted in an increase in the amount of operational data at the workplace, such as spreadsheets, graphic designs, videos and audio files. All these data are unstructured in nature and consumes expensive storage infrastructure to store and process. There are software solutions which help in processing this unstructured data. Some of the important factors to be studied in this regard are as follows:

- **Cost:** Cost of data storage ranks third in the total project cost when compared to total IT infrastructure expenditure by the companies. This made companies to rework their storage cost strategy to minimize expenses by adopting low-cost alternatives. It is a challenging task for the organizations to handle this because of the increasing volume of data generated with time.
- **Impact on backup operations:** As the volume of data rises over time, requiring more and more storage capacities, it becomes difficult to handle backup operations by an organization. Failure to achieve total data protection can be risky while honoring service level agreements (SLAs) with clients. This causes loss of data, customer service delays and also the loss of goodwill. Unless huge investments are made by the companies to upgrade their backup system, it may lead to disruption of services.
- **Disaster recovery:** Unless the value of information is accessed and its importance is identified, it would be difficult for the organization to distinguish between very important data and less important data. This is an important factor as it allows deciding which data is critical and which is not. As a result, the companies can prioritize what data to be stored on primary disks. This helps in having a better data recovery strategy by avoiding non-mission critical data during recovery operations. It improves the speed and reduces storage requirements for data recovery.

- **Unstructured data complicates regulatory compliance and legal discovery:** Corporate and government regulations often require IT firms to store data for many years and to provide particular data on request. Unstructured data, in particular, is very difficult to categorize and locate for the purpose of archive/retrieval. When terabytes of unstructured data are involved, manual methods of searching can cripple an organization.

.Example: How Moderna Used Cloud Based Storage and other IT Infrastructure to Speed Up Vaccine Development

Moderna (Life sciences company) was well known for developing a vaccine for covid 19 in a record time. The development, storage and analysis of huge volume of unstructured data was needed. As a young company, the company's IT infrastructure strategy was based on Cloud (Amazon Cloud). Since the huge volumes of data could be stored and managed with cloud-based storage, the company could just concentrate on its core analysis activity. It could finish the sequence for its mRNA COVID-19 vaccine in two days using machine learning models created on AWS. The first clinical batch could be released twenty-five days later.

Source: https://aws.amazon.com/solutions/case-studies/innovators/moderna/, 2021, Accessed on 15/08/2022

8.5 Unstructured Data Usage Patterns

The main problem with unstructured data is that the users do not have an idea of how to prioritize data by deciding which data is important and which is not. As a result, the data accumulated may consist of duplicate files, files which are not accessed for years and obsolete data. Storing, protecting and archiving of such data is very costly and time-consuming operation.

Kryder's law refers to rapid increases in magnetic drive storage density over decades. The hard drives have transformed from storing a few thousand individual bits of information in the 1950s to the latest innovation of small, cheaper and high-volume drives storing gigabytes and terabytes of information. This phenomenon led to the development of different deployable technologies for both commercial and research needs. Volume of storage is the pivot of measurement as per Kryder's law. Generally, users do not delete unwanted files, due to which volume of data grows enormously. These result in capacities getting exhausted, irrespective of the type of storage being used such as direct, NAS (network attached storage) or SAN (storage area network). Companies have to forklift their storage capacities by adding additional servers to their existing ones, leading to more maintenance overheads.

Without having a well-defined strategy, this problem of unstructured data management will be there to stay forever, leading to additional cost for handling storage, backup and recovery operations. To address these issues, IT companies need solutions to handle the visibility of unstructured data with the ability to automatically sort itself. This may include data quality issues like automatically identifying the usage frequency of the data, duplicate data and unused data. Its file archiving technologies should be intelligent to automatically detect these issues so that a substantial reduction of storage management cost is possible. Some other techniques include:

1. Implementation of multi-tiered automated storage.
2. Making use of metadata for data categorization and applying storage policies regarding unstructured data.
3. Developing new optimized backup strategies.

Example: How Bioware utilized Dell's Automated Storage Management

Biodata was a Video Game developing company which had to work with huge volume of ever-growing unstructured data in the form of Video content. The company had been facing the problem of managing such huge data and share it among developers working across geographies and time zones. Also, availability of data without disruption was a major issue. Back up and restoration had to be robust. Scalability was an issue. Migration of data among different tiers was also a challenge. Once the frequency of usage for a piece of data had decreased, it needed to be automatically shifted to a lower tier. All this was achieved by going for Dell's storage automation solution.

Source: https://i.dell.com/sites/content/business/solutions/power/en/Documents/ps4q11.pdf (2021), Accessed on 15/08/2022

Activity 8.1

Healthcare Service Planning

A healthcare company provides services to underprivileged children and the destitutes who live in various state-run rehabilitation centers managed by various NGOs and self-help groups. As the network of NGOs is very large and widespread, the healthcare service provider wants to make use of a cloud-based solution to predict monthly usage of drugs and the number of lab visits needed by the people it serves. These monthly trends will help the company to effectively plan its quarterly fund requirements. Suggest a technology that suits and efficiently handles the above-given scenario.

Answer:

Check Your Progress - 1

1. Unstructured data is a combination of human and which of the following types of data?

   a. Machine

   b. Nuclear

   c. Research

   d. Social media

   e. Fingerprint

2. Which of the following is an example of structured data?

   a. Database

   b. PDF

   c. Biometric

   d. Voice

   e. Medical prescription

3. Which of the following is an example of human-generated unstructured data?

   a. Text data

   b. Audio files

   c. Videos

   d. Website content

   e. Photos

4. Which of the following means Data recovery?

   a. Deleting a record

   b. Appending a record

   c. Viewing a record

   d. Reconstructing data

   e. Querying data

5. What is Metadata?

   a. Processed data

   b. Client data

   c. Big data

   d. A software

   e. Description of structure, administrating the data

8.6 Different Ways of Analyzing and Managing Unstructured Data

Predictive analytics is one of the most popular methods used to predict the occurrence of an event, such as to know the trends and to find the possibilities of future occurrence of certain event out of the existing data universe. This is applicable in the field of digital commerce (e-Commerce), location awareness services, cloud computing, mobility and digital security.

But, most analytics-based technologies are tailor-made to suit structured data rather than for unstructured data. Hence there is a need to address the issue of handling unstructured data. Unstructured data includes social media tweets, blogs, posts and other forms. It also includes emails, images and open-ended surveys. There are proven IT Solutions for the processing of this type of unstructured data.

To analyze unstructured data, there are some basic steps or phases to be followed to prepare, organize and manage it.

These basic steps for preparing data for analysis are:

- **Identify Important Data Sources**

  The first step for a good analysis of unstructured data is to identify all sources of data which are essential and important. Use data that is absolutely relevant, avoid duplicate records and obsolete content for analysis.

- **Set Analytics Method and Goals**

  Without a purpose, any analysis is of no use. Hence having a predefined and focused goal is very much necessary. Work out as to what output is required, like whether the requirement is the quantity-based trends or a comparison between the products.

- **Technology Availability**

  Study technology options available on hand and assess their capabilities to meet the final requirements of analysis. Set up the information architecture along with factors for choosing data storage and retrieval. This should be based on scalability, volume and variety of data being analysed.

- **Real-Time Content**

  For online e-commerce companies, real-time data access is important to provide real-time quotes. It requires tracking of real-time events and provides an output based on the predictive analytic engine's output. The data considered for analysis should include content from social media engagement and machine-generated content. The technology platform should ensure that no data is lost from the real-time stream of multiple sources.

- **Save the Original Before Sending to Data Warehouses**

  With the advent of big data, storing information in its native format is very useful. It preserves metadata and other information that might assist analysis when required. So keep a copy of the same.

- **Prepare Data for Storage**

  While keeping the original file, clean up a copy from the noise in the content. For example, remove noise like white spaces and symbols while converting informal text to formal language.

- **Ontology Evaluation**

  Analysis allows building of relationships among sources and the extracted entities for designing a structured database as per specifications. Though time-consuming, it is worth doing for quality reasons.

8.6.1 Big Data: 9 Steps to Extract Insights from Unstructured Data

According to Gartner, about 80% of data held by an organization is unstructured data[5]. This is in addition to the voluminous diagnostic information logged by embedded and user devices. It is difficult to make meaning out of unstructured data. Organizations need to study both structured and unstructured data to assess meaningful business decisions, including determining customer sentiment, cooperating with e-discovery requirements and personalizing product for their customers. This must be done to ensure that the organization is on top of any network security threats and proper functioning of embedded devices.

However, by carefully studying the disparate sets of unstructured data, one can identify connections from unrelated data sources and find patterns.

There are nine steps to analyze unstructured data. They are:

1. **Make Sense of the Disparate Data Sources**

   Before one can begin, one needs to know what sources of data are important for the analysis. One information channel is log files from devices, but that source won't be of much help when searching for user trends. If the information being analyzed is only tangentially related to the topic at hand, it should be set aside. Instead, only use information sources that are absolutely relevant.

2. **Sign Off on the Method of Analytics and Find a Clear Way to Present the Results**

   The analysis is useless if it is not clear what the end result should be. One must understand what sort of answer is needed - is it a quantity, a trend or something else? In addition, one must provide a roadmap for what to do with

[5] https://www.datamation.com/applications/big-data-9-steps-to-extract-insight-from-unstructured-data.html

the results so that they can be used in a predictive analytics engine before undergoing segmentation and integration into the business's information store.

3. **Decide the Technology Stack for Data Ingestion and Storage**

   Even though the raw data can come from a wide variety of sources, the results of the analysis must be placed in a technology stack or cloud-connected information store. This enables easy utilization of the results. Factors that are important for choosing the data storage and data retrieval depend often on the scalability, volume, variety and velocity requirements. A potential technology stack should be well evaluated against the final requirements, after which the information architecture of the project is set.

   A few likely influential requirements are that the results of the analysis must be available in real-time, have high availability for access while still functioning in a real-time multi-tenant environment. Real-time access is crucial, as it has become important for e-commerce companies to provide real-time quotes. This requires tracking real-time activities, and providing offerings based on the results of a predictive analytics engine. Technologies that can provide this include Storm, Flume and Lambda. High availability is crucial for ingesting information from social media. The technology platform used must ensure that no loss of data occurs in a real-time stream. It is a good idea to use a messaging queue to hold incoming information as part of a data redundancy plan, such as Apache Kafka. The ability to function in real-time multi-tenancy environments is required if the results are required to avoid state changes and continue to be mutable data.

4. **Keep Information in a Data Lake until it has to be Stored in a Data Warehouse**

   Traditionally, an organization obtains or generates information, sanitizes it and stores it away. For example, if the information source is an HTML file, the text may have to be stripped and the rest discarded so that information is not lost during storage in a data warehouse.

   Previously, the extraneous pieces of information were first stripped away and the remaining data was used for arriving at conclusions. However anything useful that was discarded during the stripping process was lost forever. However, with the advent of big data, it has become a common practice to do the opposite, i.e., store first and sanitize later. With a data lake, information is stored in its native format until it is actually deemed useful and needed for a specific purpose.

5. **Prepare the Data for Storage**

   While keeping the original file, if one needs to make use of that data, it is best to clean up a copy. In a text file, there can be a lot of noise or short hand that can obscure valuable information. It is a good practice to cleanse noise like whitespaces and symbols, while converting informal text in strings to formal language. If it is possible to detect the spoken language, it should be categorized as such. Duplicate results should be removed, the dataset treated for missing values, and off-topic information removed from the dataset.

6. Retrieve Useful Information

Through the use of natural language processing and semantic analysis, one can make use of Parts-of-Speech tagging to extract common named entities, such as "person," "organization," "location" and their relationships. From this, one can create a term frequency matrix to understand the word pattern and flow in the text.

7. Ontology Evaluation

Through analysis, one can then create the relationships among the sources and the extracted entities so that a structured database can be designed to specifications. This can take time, but the insights provided can be worthy for an organization.

8. Statistical Modeling and Execution

Once the database has been created, the data must be classified and segmented. It can save time to make use of supervised and unsupervised machine learning, such as the K-means, Logistic Regression, Naïve Bayes, and Support Vector Machine algorithms. These tools can be used to find similarities in customer behavior, targeting for a campaign and overall document classification. The disposition of customers can be determined with sentiment analysis of reviews and feedback, which helps to understand future product recommendations, overall trends and guide introductions of new products and services.

The most relevant topics discussed by customers can be analyzed through temporal modeling techniques, which can extract the topics or events that customers are sharing via social media, feedback forms or any other platform.

9. Obtain Insight from the Analysis and Visualize It

From all the above steps, it all comes down to the end result. It is crucial that the answers to the analysis are provided in a tabular and graphical format, providing actionable insights for the end-user of the resultant information. To ensure that the information can be used and accessed by the intended parties, it should be rendered in a way that it can be reviewed through a handheld device or web-based tool, so that the recipient can take the recommended actions on a real-time or near-real-time basis.

The above steps are used for planning analysis of unstructured data and there are many business intelligence (BI) platforms that are already available from different companies like SAS, Tibco's Spotfire, SAP/NetWeaver and IBM/Cognos.

Example: Monzo (the bank) Uses "Social Listening" to Understand Customer Sentiments better and Act on them

Monzo was a new generation bank deploying IT for operational efficiency and customer satisfaction. It had been getting customer feedback from traditional sources but with the onset of social media presence, the company found a new way to engage with customers and prospects. The company realized, just being present on the social media was not adding much value. It realized the need to actively listen to the customers, analyse the data and decide what the customers were feeling about its campaigns, products, issues etc. This social listening had resulted in better decision making within the organization.

Source: https://www.brandwatch.com/case-studies/monzo/view/, 2020, Accessed on 15/08/2022

8.7 Comprehensive File Archiving Solution

Big data includes machine-generated data, IoT data like web search logs, satellite images and other data like audio, video and health records. Analytics are performed on such data to support the decision and support applications. On the other hand, human-generated unstructured data is also a constituent of big data, which is to be organized to suit data mining activities. Human-generated data includes contracts, tenders, CADs (computer-aided design) and content from mobile phones and social media. Summarization of these data is important for the organization. Organizing unstructured data is complex and time-consuming as it needs a strategy and integration method to a club with structured data.

For considering the implementation of a big data-based archiving system in an organization, the following aspects are important:

- Reducing primary storage consumption.
- Supporting data growth.
- Identifying non-active data.
- Multi-tier architecture for data storage.
- Ability to retain, categorize and in future mine the data.
- Compliance standards are followed to suit both structured and unstructured data.
- Density scaling.
- High throughput.
- Fast retrieval.
- Protection and disaster recovery.

Let us study the Strongbox technology to manage and handle unstructured data.

Strongbox: The usage of IT has been widespread across all walks of life. Data associated with these applications is also exponentially growing over time. Data from these IT-enabled applications need to be stored indefinitely. Its security and retrieval are also very important to the organization to access it when it is required.

Strongbox technology addresses these issues, within the reach of the organization's budget. Strongbox is a network-attached storage (NAS)-based technology to simplify data access. It provides cost-effective, linear tape file system (LTFS)-based storage for structured/unstructured data. Strongbox ensures complete data protection for all online and archive storage needs.

Advantages of using Strongbox are:

- Controls storage costs.
- Reduces TCO (Total Cost of Ownership) by more than 50%.
- Provides online and archive storage with an easy-to-use solution.
- Ensures long-term protection of files.
- Eliminates backup requirements of fixed content.
- Delivers performance to meet application workflow.
- Reduces the gap between data growth and budget.
- Provides an LTFS-based solution.

Example: Riveron (Business Advisory Company) Uses Strongbox Technology to Manage its Unstructured Data

Riveron was a business advisory company helping customers in the financial, technology and operations domain. The company used its domain expertise to deeply analyse the company's (which company its client wants to acquire or invest in) financial statements (unstructured data) to provide an uncluttered view of the financial performance, reliability of the data being used for decision-making related to investment or acquisition .The company used Strongbox technology to manage its data and drastically save time to give advice to its clients

Source: https://www.strongbox.ai/case-study-riveron, 2022, Accessed on 15/08/2022

8.8 Opportunities and Problems with Unstructured Data

It is estimated that about 80 to 85 percent of the user data is in an unstructured format. These are stored and processed on IT-enabled applications or technologies. Most of such data in smaller quantities can be analyzed by users in an optimized manner using their IQ and natural intelligence. But the same is not possible in the case of large amounts of unstructured data. The only alternative is to take the assistance of IT solutions available. Due to the existence of different kinds of unstructured data, the analysis is quite complicated. Basically, unstructured data is divided into two broad categories, namely, textual and non-textual data. Textual data includes all the documents, e-mails, text files and chat files. Similarly, non-textual data like graphics, sound, movies, etc., are also stored and processed.

There are a few sticky issues in this categorization. For example, a powerpoint presentation is a mix of both textual and non-textual data. So it is difficult to categorize this data.

i) Textual data

While dealing with unstructured textual data, there are many things that human beings can do that computers cannot. Let us see some instances:

a) **Context:** In the textual content, problems may arise because of context. For instance, the difference between the following statements can be marked — "John rides in a Mustang" and "John rides on a mustang". The first statement mustang means a car (Mustang is a popular car model by Ford); in the second case, it may be considered a horse (an American feral horse which is typically small and lightly built is called mustang). The human brain picks up all of these options almost instantaneously and understands implicitly. Computers cannot do the same unless exclusive commands are given on how to understand this statement.

b) **Style, structure and vocabulary:** Sometimes writing style, sentence structure and vocabulary used in emails and SMS messaging are different, though both come under the textual document. Here also the human is better placed in identifying these variations.

c) **Addressed to whom:** We use different styles while addressing friends, relatives and superiors in the office. The human brain captures these variations easily based on the situation, but for computers, it is a difficult task to handle.

The tools used to analyze textual data are search based on keyword/key phrases. Analysts used to prepare a large list of relevant keywords needed by computers to search large volumes of data. This method had its problems. If the search criteria are too narrow then there are chances of missing vital information, if it is too broad then the output may contain irrelevant data. Some advances are made by applying concepts like artificial intelligence, fuzzy approach and neural networks to address this type of problem. Neural networks are used to recognize simple tacit linkages between words and expressions.

ii) Non-textual data

Characteristics and schemes of handling non-textual data include:

a) **Common format:** In the case of non-textual data, the major problem for the computers is how to convert it into a common searchable form. To handle large amounts of this data by converting multiple forms of data into a common single format, this can be used for performing searches. For instance, Optical Character Recognition (OCR) systems are used to perform this conversion activity reliably.

b) **Audio:** Modern voice and language recognition systems are becoming more and more capable of converting voice recordings to text as technology advances. Many of the language recognition systems are based on neural networks and artificial intelligence principles.

c) **Graphics/Video:** Dealing with graphics or movie files is a very complicated problem as it requires a significant amount of processing power. Due to the presence of possible illusions, searching video content is possible by using a specialized predefined set of keywords. For example, identifying adult content and racist comments, etc.

Example: JP Morgan Asset Management Deploys an Artificial Intelligence-Based Tool to Help Managers Build a Theme-Based Climate Change Fund

JP Morgan Asset Management company used an AI based tool; ThemeBot to assist its managers to use text analytics and other tools to build theme-based funds for its customers and prospects. One such fund was related to Climate Change Fund. The tool used AI tool to process crores of news articles, research papers, regulatory filings, and profiles of companies to zero on companies that had a focus on reducing carbon footprint. The tool would work closely with human analysts who would short list around 100 companies from which the fund managers would choose the final composition.

Source: https://www.proactiveinvestors.co.uk/companies/news/953801/jp-morgan-launches-climate-change-fund-that-picks-stocks-using-ai-953801.html (30 June 2021), Accessed on 15/08/2022

Activity 8.2

Data Storage Solution

An automobile parts and service provider serves auto workshops by running a network of retail centers at various villages. It is a noble initiative, to reach and serve rural customers in their village to reduce their transport costs. Currently, the company does not have the proper hardware/software to securely store and manage its unstructured data, especially, the sales records of various parts at various centers. The company maintains many nodal centers at various district headquarters to consolidate and manage sales records procured from various rural retail centers in the surrounding villages in that district. They want to implement a cloud-based storage and data management technology connecting all nodal centers.

Suggest a cost-effective data storage technology suitable to address the situation. Also, list some of the major features of the technology suggested. They are planning to increase their operations if found beneficial.

Answer:

Check Your Progress - 2

6. Which of the following is an example of a non-textual data?
   a. Documents
   b. SMS messages
   c. databases
   d. Data warehouses
   e. Videos
7. Which of the following is not an advantage with the strongbox solution?
   a. Controls storage costs
   b. Reduces TCO by more than 50%
   c. Provides online and archive storage with an easy-to-use solution
   d. Ensures the long-term protection of files
   e. Focusing on a particular brand
8. Which of the following does NAS stand for?
   a. Network Attached Storage
   b. Network Access System
   c. Network Access Storage
   d. Network Attached System
   e. Network Accessing Software

9. Which of the following does SAN stand for?
    a. Social Area Network
    b. Storage Area Network
    c. Service Area Network
    d. Server Area Network
    e. Structured Area Network
10. What is the full form of LTFS?
    a. Linear Type Flow System
    b. Linear Tape File Service
    c. Linear Tape File System
    d. Linear Technology for Files and Services
    e. Large type File system

8.9 Summary

- Big data is a collection of structured and unstructured data, which includes unorganized data generated by humans and machines.
- Information which is not based on a well-defined structure is called unstructured data. Almost 80 percent of the existing data in organizations is unstructured.
- Managing, maintaining and analyzing unstructured data is very difficult due to the lack of uniformity in its content. Its volume is on the rise as more and more technology is being used by companies. Proper analysis is the key to effectively using big data for decision-making. Though analyzing unstructured data is complex and time-consuming, a few companies like SAP and IBM have developed many technology platforms to address these issues.
- As big data comprises data from various sources, operations to store, backup and recovery of data in a secure manner are very crucial for all business activities. Analyzing unstructured data has many unresolved problems which are better handled by humans when compared to computers. Some of the issues in this regard are the context, style of writing and the receiver. Artificial intelligence and neural networks are being used to overcome these gaps.

8.10 Glossary

Data Lake: Storage repository that holds a vast amount of raw data in its native format until it is needed.

Kryder's Law: Kryder's law represents an analysis of the hard drive's density and capability to store data over time. It correlates with Moore's law concept which says that the number of transistors placed on an integrated circuit should double every two years, resulting in predictable progress in microprocessor speed.

Linear Tape File System: It is a tape file system by IBM, which enables tape media to be read by the operating system when the magnetic tape is inserted into the tape drive.

Mutable Data: A mutable object can be changed after it is created. Custom classes are generally mutable.

NAS: Network-attached storage (NAS) is a file-based computer data storage server which facilitates data access to a heterogeneous group of clients connected over a data network.

Predictive Analytics: Predictive analytics is a method of extracting information from existing datasets in order to identify patterns and predict future outcomes and trends.

Real-time Multi-tenant Environment: A tenant is a group of users who share a common access with specific privileges to the software instance. Multi tenant refers to more than one and real-time addresses the immediate nature.

SAN: Storage area network (SAN) is a dedicated high-speed network that interconnects shared pools of storage devices with multiple servers.

State Changes: State refers collectively to the data stored in the object that determines the current properties of the object and a change refers to its alteration.

TCO: Total cost of ownership (TCO) is a cost estimate of the total cost incurred for purchasing the product along with the operational costs involved to maintain it. This helps buyers and owners in decision-making while purchasing a product or a system.

8.11 Self-Assessment Test

1. Briefly discuss the different forms of unstructured data known to you with suitable examples.
2. Discuss the relevance of big data in the banking industry.
3. Describe the key elements involved while preparing unstructured data for analysis.
4. Explain the following terms:
   a. Non-textual data.
   b. Backup and Recovery.
5. Does big data technology support decision-making in an organization? If yes, give a good example in the areas of marketing, education and insurance domains.

8.12 Suggested Readings/Reference Material

1. Maleh, Yassine. Shojafar, Mohammad. Alazab, Mamoun. Baddi, Youssef. Machine Intelligence and Big Data Analytics for Cybersecurity Applications (Studies in Computational Intelligence, 919) 1st ed. 2021 Edition.
2. Ahmed, Syed Thouheed. Basha, Syed Muzamil. Arumugam, Sanjeev Ram. Patil, Kiran Kumari. Big Data Analytics and Cloud Computing: A Beginner's Guide, 2021.
3. Saleem, Tausifa Jan. Chishti, Mohammad Ahsan. Big Data Analytics for Internet of Things 1st Edition.
4. Jones, Herbert. Data Science: The Ultimate Guide to Data Analytics, Data Mining, Data Warehousing, Data Visualization, Regression Analysis, Database Querying, Big Data for Business and Machine Learning for Beginners Hardcover – 10 January 2020.
5. Maheshwari, Anil. Data Analytics Made Accessible: 2023 edition Kindle Edition.
6. Mayer-Schönberger, Viktor. Cukier, Kenneth. Big Data: A Revolution That Will Transform How We Live, Work, and Think Paperback – October 26, 2021.

8.13 Answers to Check Your Progress Questions

1. **(a) Machine**

   Unstructured data is made of both machine and human-generated data. Such data can support the decision-making process in organizations in different domains such as education, healthcare and marketing.

2. **(a) Database**

   Database is an example of structured data as it is based on a pre-defined structure called metadata.

3. **(d) Website content**

   This comes from any site delivering unstructured content, like YouTube, Flickr, or Instagram.

4. **(d) Reconstructing data**

   Data recovery is a method to rebuild data which is generally done when there is a loss of data due to technical reason.

5. **(e) Description of the structure of data**

   Metadata is also called the catalog which describes the structure of the data.

6. **(e) Videos**

   Due to the presence of possible illusions, searching video content is possible by using a specialized predefined set of keywords.

7. **(e) Focusing on a particular brand**

   In strong box solution, all the other factors are advantageous except Focusing on a particular brand.

8. **(a) Network-Attached Storage**

   NAS is a file-based computer data storage server which facilitates data access to a heterogeneous group of clients connected over a data network.

9. **(b) Storage Area Network**

   Storage Area Network (SAN) is a dedicated high-speed network that interconnects shared pools of storage devices to multiple servers.

10. **(c) Linear Tape File System**

    LTFS stands for the linear tape file system. It is a tape file system, which enables tape media to be read by the operating system when the magnetic tape is inserted into the tape drive.

Unit 9

Information Management

Structure

"There's no such thing as knowledge management; there are only knowledgeable people. Information only becomes knowledge in the hands of someone who knows what to do with it."

- Peter Drucker

9.1 Introduction

This quote emphasises the pivotal role of people in the organization who will use the insights provided by the analysis to derive the knowledge which can lead to actions adding value to the customers.

In the previous unit we discussed about, handling unstructured data, different formats of unstructured data like text, web pages, video, audio, images and office software. We also covered the topics, key issues in storing unstructured data, usage patterns of unstructured data, different ways of analyzing and managing unstructured data, comprehensive file archiving solution and opportunities and problems with unstructured data. Information management is like combination of foundation and plumbing in a house. If the foundation of the house is not solid then we will get all kinds of structural problems which can cause the crashing of the house. Big data foundation has two major components, one is storing the data and the other is processing the data. To store big data, we use data warehouse and should adapt a proper file system. Today we are using Hadoop File System (HDFS) for big data storage. To process the big data, we use computing techniques and emerging technologies.

In this unit we will be discussing big data computing techniques and its limitations.

9.2 Objectives

After the completion of this, you will be able to -

- Define big data foundation and how HDFS has become so popular in big data storage
- Give examples of different computing platforms used for big data computation
- Define the parallel computing environments for big data computation
- Explain computational limitations of big data
- Relate to emerging technologies for big data

9.3 Foundation Of Big Data

There are two components in big data foundation. They are-

- Storing the data
- Processing the data

Storing the Data

Hadoop file system (HDFS) is the distributed data storage system used to store the big data. HDFS has become very popular because it stores the data automatically without any 'set up', whereas in traditional database we need to do 'set up' in order to store the data.

In traditional databases before dumping the data, first, we have to create a schema that is blue print of the database tables with rows and columns. But in HDFS we simply dump the data into the file without any blue print.

The traditional database has some limitations on the data storage whereas HDFS supports all types of data and size of data that a database can store.

Processing the Data

Processing the data means manipulation and calculations on big data. Today HDFS is using 'MapReduce' as a processing software for big data.

MapReduce is a parallel programming framework designed for distributed processing. It divides the workload into smaller workloads which are distributed. It uses Java programming for manipulations and calculations on big data. Further, it is fault tolerant because it automatically recovers and handles processing failures.

Exhibit 9.1 gives an example of MapReduce.

Exhibit 9.1: An Example of MapReduce

Here is a simple example. There are five files. Each file has two columns, key and a value in Hadoop terms. These are representing a city and the respective temperature recorded in the city during various days. It is a simple example to follow. A real application will not be as simple. It contains millions/billions of rows. The rows may not be neatly formatted. The key principles being discussed here remain the same, independent of how big or small the amount of data one needs to analyze.

20,Toronto	25, Whitby	22, New York	32, Rome
4, Toronto	33, Rome	18, New York	

Assume there are another four such files with similar data. The objective is to find the maximum temperature for each city across all available data files from this collected data. Same city may be represented multiple times in the files. We use the MapReduce framework, and break this down into five map tasks. Each mapper works on one of the five files. The mapper goes through the data in the files, and picks out the maximum temperature against each city. For the available data above, the results from mapper task is:

20, Toronto	25, Whitby	22, New York	23, Rome

Four other mapper tasks, working on four other files produced the following results:

18, Toronto	27, Whitby	32, New York	37, Rome
32, Toronto	20, Whitby	33, New York	38, Rome
18, Toronto	27, Whitby	32, New York	37, Rome
32, Toronto	20, Whitby	33, New York	38, Rome

All these five generated output streams are fed into the Reduce Tasks.

They get combined. and the task arrive with a single output value for each city as a final result:

32, Toronto	27, Whitby	33, New York	38, Rome

You can think of similar example of census conduct in Roman times and MapReduce data as above. This mapping of people to cities, in parallel, and then combining the results (reducing) is much more efficient than sending a single person and doing a serial fashion count of every person in the empire.

Source:https://www-01.ibm.com/software/data/infosphere/hadoop/mapreduce/

Example: JP Morgan Chase Used Hadoop for Big Data Analytics

JP Morgan Chase dealt with 150 petabytes of data stored in 30000 databases to serve some 3.5 billion users. The bank used big data analytics to analyse unstructured and structured data using the Hadoop. Hadoop's could store huge volumes of unstructured data like web logs, transaction data and social media data. The data was analysed using data mining tools to derive insights to provide better customer service.

Source: How JPMorgan uses Hadoop to leverage Big Data Analytics? (projectpro.io) (28 June 2022), Accessed on 16/08/22

9.4 Computing Platforms that can Handle Big Data

There are three parallel computing options to handle big data:

- Clusters or grids
- Massively parallel processing (MPP)
- High performance computing (HPC)

Clusters or Grids

Clusters and grids are either homogeneous or heterogeneous hardware environments where the servers are loosely coupled for distributed workloads. Grid environment provides lower-cost storage environment but not that much significant compared to other parallel alternatives.

Cluster and grid are public cloud environments. Here, space and computing power are sold to customers through vendors like Amazon.

The cloud environment has become popular because of its flexibility. The customer can pay for the space that is needed at a particular point in time, looking at the increase or decrease in the data needs.

Table 9.1 shares a summary of cluster, grid, and cloud computing paradigms.

Table 9.1: Summary of Cluster, Grid, and Cloud Computing Paradigms

	Cluster Computing	Grid Computing	Cloud Computing
Basic Idea	Aggregation of resources	Segregation of reources	Consolidation of resources
Running Processes	Same processes run on all computers over the cluster at the same time	Job is divided into sub-jobs each is assigned to an idle CPU so they all run concurrently	Depends on service provisioning. Which computer offers a service and provisions it to the requesting clients

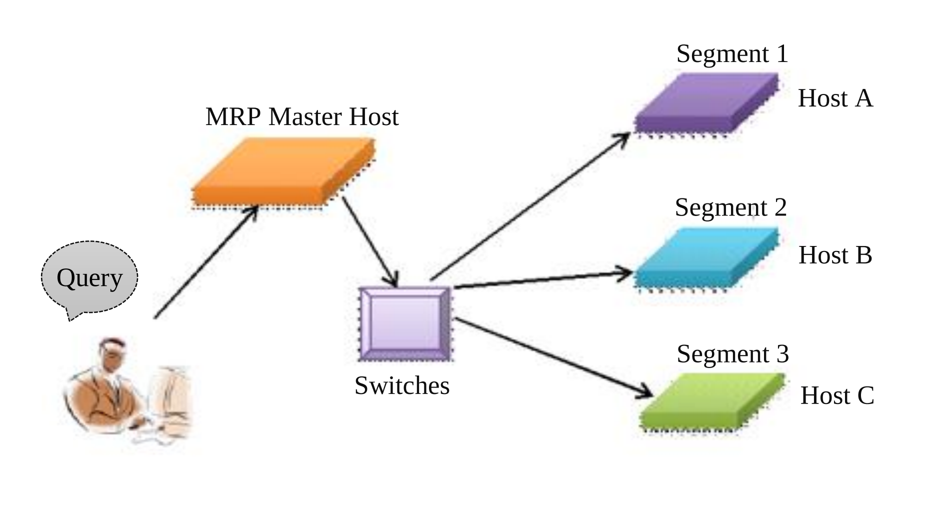
Contd….

Operating System	All nodes must run the same operating system	No restriction is made on the operating system	No restriction is made on the operating system
Job Exeuction	Exeuction depends on job scheduling. So, jobs wait until it's assigned a runtime	Execution is scalable in a way that moves the execution of a job to an idle processor (node)	Self-Managed
Suitable for Apps	Cascading tasks. If one tasks depends on another one.	Not suitable for cascading tasks.	On-demand service provisioning.
Location of nodes	Physically in the same location	Distributed geographicaly al over the globe	Location doesn't matter
Homo/Heterogeneity	Homogeneous	Heterogeneous	Heterogeneous
Virtualization	None	None	Virtualization is a key
Transparency	Yes	Yes	Yes
Security	High	High, but doesn't reach the level of cluter computing	Lower than both types
Interoperability	Yes	Yes	No
Application Domains	Industrial sector, research centers, health care, and centers that offer services on the nation-wide level	Industrial sector, research centers, health care, and centers that offer services on the nation-wide level	Banking, Insurance, Weather Forecasting, Space Exploration, Business, IaaS, PaaS, SaaS
Implementation	Easy	Difficult	Difficult – need to be done by the host.
Management	Easy	Difficult	Difficult
Resource Management	Centralized (locally)	Distributed	Both centralized and distributed.
Internet	No internet access is required	Required	Required

Source: Al Etawi. (2018). A Comparison between Cluster, Grid, and Cloud Computing. International Journal of Computer Applications, 179(32), 41. https://docslib.org/doc/6714517/a-comparison-between-cluster-grid-and-cloud-computing

A massively parallel processing platform (MPP) combines storage and memory. MPP is used for known high-value use cases and allows the software to maximize the storage and computational capabilities. Further, MPP computes the storage and memory into a single machine to optimize performance and scalability. The MPP appliances are designed for easier analytics if the enterprises have used MPP for 10+ years. Figure 9.1 presents data distribution in MPP.

Figure 9.1: Data Distribution in MPP



Source: ICFAI Research Center

High Performance Computing (HPC)

HPC is designed for high-speed processing. In this, the calculations are done in memory for the highest computational performance. HPC environments are mostly used by research organizations and business units which require very high computational performance. IBM and Cray Computers are examples of HPC environments.

Example: Pfizer used High Performance Computing Platforms for Faster Vaccine and Drug Discovery and Development

Pfizer was well known for bringing in vaccine and anti-viral drugs (for Covid virus) to the market with record speed.

The whole process of drug discovery and development was hastened up several times due to use of High-Performance computing (Super computers), HPC facilitated sophisticated modelling and simulation techniques which enabled the scientist to test molecular compounds in virtual lab in place of a physical lab. Millions of compounds could be tested in a fraction of time taken for real lab tests. The company could easily short list potential compounds for drug development.

Source : Pfizer Discusses Use of Supercomputing and AI for Covid Drug Development (hpcwire.com), (March 24, 2022) Accessed on 16/08/2022

9.5 Big Data Computation

Big data computing platforms are parallel computing environments. When the workload is distributed to the parallel processors, the results are combined to produce the correct result and that is called a parallel computing environment. Some environments provide software tools for parallel programing and others leave it to the skill of the programer.

The parallel programing environment includes:

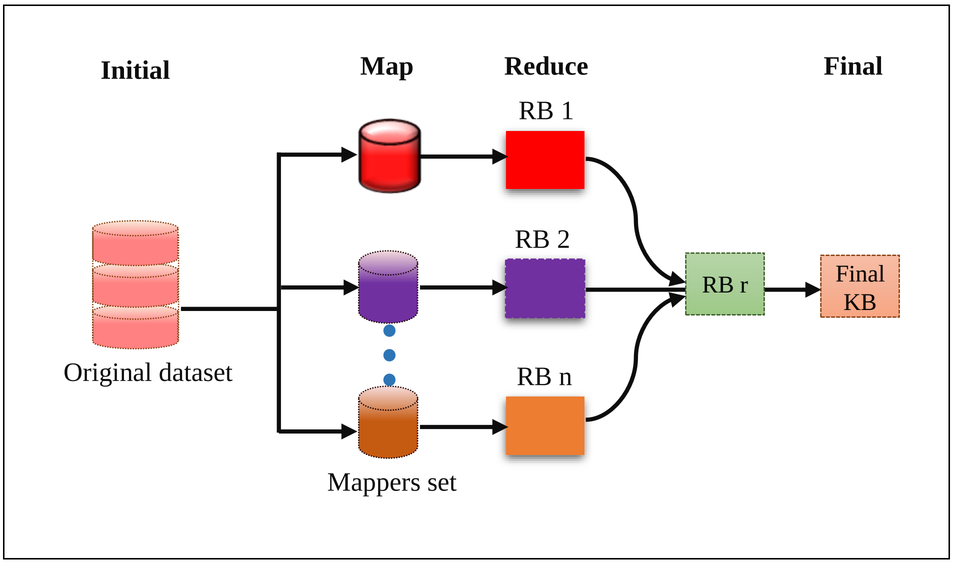
- MapReduce
- In-database Analytics
- Message Passing Interface (MPI)

MapReduce

MapReduce is a common parallelization technique used in big data computation.

The MapReduce framework distributes the workload to all the available processes through the map function. Each processor performs the reduce task individually, and the result is fed into reducer task. The following Figure 9.2 shows the working of MapReduce framework.

Figure 9.2: MapReduce Framework



Source: ICFAI Research Center

In-Database Analytics

In-database analytics refers to the integration of data analytics into data warehousing functionality. It follows two methods. The first one is a collection of user-defined-extensions (SQL extensions) to execute the code in database. The second one is parallelized analytics that is executed in the database.

There are three types of user-defined extensions

- **User-defined function (UDF):** It performs a task and returns a value.
- **User-defined aggregate (UDA):** It summarizes a group of data and returns a value.
- **User-defined table function (UDTF):** It reshapes the table format of data. Ex: 10 columns and 100 rows are transformed into 2 columns and 500 rows.

Message Passing Interface (MPI)

MPI is a software process which allows sharing of information between the processes in a distributed environment. It is used to share interim processing results between the processes while attempting to find the middle number.

Example: Tesco (the retail giant) Used Big Data Analytics with in Data Base Analytics to Forecast Sales

Tesco used data modelling and mining techniques on huge customer data for sales forecasting. The retailer could get answers to such questions as "how various products are purchased" each day of the week in various stores. The pattern regarding online purchases and in store purchases was also obtained. As the company dealt with huge volume of data, it had to move data from data bases to do analysis and back. This was time consuming. So, the company adapted "in base data analytics" so that the analytics was done on the data in the data base.

Source: 5 ways Tesco uses Big data Analytics | Analytics Steps (June 20, 2020), Accessed on 17/08/2022

Activity 9.1

Get to know MPI in detail. How sharing of information happens between processes and how it is useful for a business manager, finance manager, production manager.

Check Your Progress - 1

1. HDFS has become very popular because it stores the data automatically without which of the following operations?

   a. Run

   b. Compile

   c. Set up

   d. Edit

   e. Save

2. Which of the following is a parallel programming framework designed for distributed processing?

   a. HDFS

   b. MapReduce

   c. Database Analytics

   d. In-database Analytics

   e. Message Passing Interface

3. Clusters or Grids provide which type of storage environment?

   a. Very expensive

   b. High-cost

   c. Medium-cost

   d. Low-cost

   e. Very low-cost

4. Which of the following platform combines storage and memory and computes to create a platform?

   a. Clusters

   b. Grids

   c. Massively Parallel Processing

   d. High Performance Computing

   e. Hadoop file system

5. HPC environments are mostly used by which of the following which require very high computational performance?

   a. Enterprises

   b. Manufacturing units

   c. Government agencies

   d. Research organizations

   e. Online shopping

9.6 Big Data Storage

The most common and popular storage for big data is data warehouse. In the data warehouse, the data is stored in the form of relational database that is rows and columns. The data is accessed either via rows or via columns in a relational database.

Big data storage should have the following characteristics-

- It should be scalable.
- It should provide a tiered architecture for storage.
- It should be self-managing.
- It should make the content highly available.
- It should make the content widely accessible.
- It should support both analytical and content applications.
- It should support workflow automation.
- It should be integrated with legacy applications.
- It should enable integration with private, public and hybrid cloud ecosystems
- It should be self-healing.

Exhibit 9.2 discusses big data storage.

Exhibit 9.2: Big Data Storage

Let us understand the storage approach-

- When it comes to storing and retrieving data, the first question is how long need I to wait to get the data I wanted? However, you also need to factor in the cost of storing the data on different tiers. An optimal big data storage architecture has processes to store the needed data and archive the rest at the lowest possible cost. The best systems support a life cycle from creation through archiving. That provides a home for data flows.
- Modern storage needs to be shared by multiple users and applications. Some intelligent applications like media asset managers, automatically migrate data between tiers of storage. Most applications do not have this capability. If a particular application’s priorities conflict with another set of applications or users, the storage administrators cannot use the functionality. Thus storage system itself manages the tiering of data across various range of media types like fast disk, slower disk, flash and tape.
- The largest scale websites and applications, created by Twitter, Facebook, Google and Amazon have the built-in ability to handle failure. When a server in this cloud environment fails, work is automatically redirected to another resource, with the failing server automatically taken offline.

Contd….

A well-designed big data storage system must work exactly in this same model. It must accommodate component failures and heal itself without customer intervention.

Source: http://searchstorage.techtarget.com/feature/Big-data-storage-and-analytics

Example: How Alaska USA Used Self-Healing Data Storage System

Alaska USA Federal Credit Union, was the 18th largest credit union in the United States. It is a cooperative with around 7 lac members. It operated from seventy-two locations in many states in USA. The company used cloud infrastructure for its IT infrastructure to handle big data analytics. The company was finding it a challenge to get a unified visibility of its operations as different hardware and software systems had different tools to monitor. So, if a failure occurred, it took time to react. It went for a solution whereby it had a comprehensive visibility across servers, storage, and applications. This included self-healing storage systems.

Source: Customer Stories | Alaska USA | Riverbed, 2022, Accessed on 17/08/2022

9.7 Big Data Computational Limitations

The typical computing limitations of big data are:

- Disk Bound
- I/O Bound
- Memory Bound
- CPU Bound

Disk Bound

Disk bound means not having enough storage capacity. For example, a disc that has 2 TB of data and only 500 TB of storage capacity.

I/O Bound

I/O bound means not having enough bandwidth to meet the needs of the business. For example, trying to squeeze 50 TB of big data in five minutes through a 10 inch diameter pipeline. In order to provide service level agreements (SLAs), a 100 inch pipeline is required.

Memory Bound

Memory is fairly expensive, therefore, it is limited on various computing platforms. Big data requires more memory for analytics. It uses memory and CPU for processing the data. It loads all data into memory then it uses CPU to perform the calculations. Memory is like a scrap of paper which is used for rough work while doing a mathematical problem.

CPU Bound

Big data computations use more capacity of the CPU. If there are many computational intensive calculations, then it can simply exhaust the CPU capacity. For example, Monte Carlo simulation in which a set of calculations are being performed with different starting conditions.

Example: How Xilinx Released High Bandwidth Versal HBM Series and Eliminated Processing, Memory Bottlenecks

Organizations in the finance, retail and finance sectors engaged in machine learning for Big Data Analytics were facing challenges for both bandwidth and memory bound processing. Machine learning based analytics projects in areas like fraud detection, Recommendation systems were facing problems with processing speed and bandwidth issues. The Xilinx platform would eliminate these challenges.

Source: Versal HBM Series (xilinx.com), 2022, Accessed on 17/08/2022

9.8 Emerging Technologies Related to Big Data

Today we are using two emerging technologies for big data computation. They are:

- SSD (solid state drives)
- GPGPUs (general purpose graphical processing Units)

Solid State Drives (SSD)

Today computing platforms are using SSDs in place of hard drives. SSD drives are memory devises which will allow to store data persistantly. SSD is also called as flash memory, which represents the underlying chips used in this technology.

SSD drives have many advantages like:

- The start-up or spin-up time is almost instantaneous.
- The random access time to get any information is 10 times faster than hard drive.
- SSD is not a mechanical drive so that the data is reduced which is already read.
- It is more affordable and has longevity.
- It is low-latency-oriented big data processing.

General Purpose Graphical Processing Units (GPGPU)

GPUs (graphical processing units) are used for video processing in PCs whereas GPGPUs (general purpose graphical processing units) are used for big data computing platforms.

Advantages of GPGPUs

- Mostly used for vector processing in big data analytics.
- It is a commodity based technology whose hardware provides similar computational capabilities to the floating points.
- Useful for large-scale computational workloads which take too long in a traditional environment.
- Performs the big data computations and sends the results to the traditional environment for persistent storage.
- Requires low level programming in either C or CUDA (Compute Unified Device Architecture).

Example: How Seoul Robotics Used Deep Learning Models to Automate Valet Services for Parking

Seoul Robotics was a software company developing autonomous solutions. The company was using Nvidia's Graphics Processing Units (GPUs) to develop an autonomous valet service that could be used at malls, airports, parking garages and more.

Source: Video: Using infrastructure and AI to create autonomous valets | Electronics360 (globalspec.com) (17 August 2022), Accessed on 17/08/2022

Activity 9.2

Discuss any popular big data storage methods used by Amazon and Walmart

Check Your Progress - 2

6. Which of the following user defined function performs the task and returns the value?

   a. SQL

   b. UDF

   c. UDA

   d. UDTF

   e. NoSQL

7. Which of the following is not a characteristic of big data storage?
    a. Scalable
    b. Tired architecture
    c. Self-managing
    d. Less content available
    e. Widely accessible
8. Which of the following means not having enough bandwidth to meet the needs of the business?
    a. Disk bound
    b. I/O bound
    c. Memory bound
    d. CPU bound
    e. Process bound
9. SSD (Solid state drives) is also called which of the following?
    a. Primary memory
    b. Secondary memory
    c. Cache memory
    d. Random memory
    e. Flash memory
10. Which of the following emerging technology is mostly used for vector processing in big data analytics?
    a. SDD
    b. GPU
    c. GPGPU
    d. HDFS
    e. CUDA

9.9 Summary

- Today there are many computing techniques that perform parallel computing.
- Big data computing platforms are parallel computing environments that work on parallel programming.
- Some environments provide software layers or tools for parallel computing.
- There are higher-level tools and analytics available which use MapReduce.
- MapReduce uses Java programming for manipulations and calculations on big data.

- Different computing platforms are available to handle the big data like clusters or grids or massively parallel processing (MPP) etc.
- In this unit we also discussed the big data emerging technologies like SSD and GPGPU.

9.10 Glossary

Cluster or Grid: Cluster or grid is a public cloud environment where it sells space and computing power to customers.

CUDA - Compute Unified Device Architecture: It is an extension of C programming language, developed by American technology company NVDIA.

GPGPUs - General Purpose Graphical Processing Units: GPGPUs are mostly used for vector processing in big data analytics.

HPC - High Performance Computing: HPC are designed for high-speed processing for highest computational performance.

In-database Analytics: In-database analytics refers to the integration of data analytics into data warehousing functionality.

MapReduce: MapReduce is a parallel programming framework designed for distributed processing.

MPI - Message Passing Interface: MPI is a software process which allows sharing of information between the processes in a distributed environment.

MPP - Massively Parallel Processing: MPP allows the software to maximize the storage and computational capabilities.

SSD - Solid State Drives: SSD drives are memory devises which will allow to store data persistantly.

9.11 Self-Assessment Test

1. What is big data foundation and how HDFS has become popular?
2. Explain in detail about parallel processing computing techniques.
3. What are the components included in big data computation?
4. What are the different user-defined extensions available in SQL language?
5. Explain in detail about the computational limitations of big data analytics.
6. What are the different emerging technologies we are using today in big data analytics?

9.12 Suggested Readings/Reference Material

1. Maleh, Yassine. Shojafar, Mohammad. Alazab, Mamoun. Baddi, Youssef. Machine Intelligence and Big Data Analytics for Cybersecurity Applications (Studies in Computational Intelligence, 919) 1st ed. 2021 Edition.

2. Ahmed, Syed Thouheed. Basha, Syed Muzamil. Arumugam, Sanjeev Ram. Patil, Kiran Kumari. Big Data Analytics and Cloud Computing: A Beginner's Guide, 2021.

3. Saleem, Tausifa Jan. Chishti, Mohammad Ahsan. Big Data Analytics for Internet of Things 1st Edition.

4. Jones, Herbert. Data Science: The Ultimate Guide to Data Analytics, Data Mining, Data Warehousing, Data Visualization, Regression Analysis, Database Querying, Big Data for Business and Machine Learning for Beginners Hardcover – 10 January 2020

5. Maheshwari, Anil. Data Analytics Made Accessible: 2023 edition Kindle Edition

6. Mayer-Schönberger, Viktor. Cukier, Kenneth. Big Data: A Revolution That Will Transform How We Live, Work, and Think Paperback – October 26, 2021

9.13 Answers to Check Your Progress Questions

1. (c) Set up

HDFS has become very popular because it stores the data automatically without any ‘set up’ whereas in traditional database we need to do ‘set up’ in order to store the data.

2. (b) MapReduce

MapReduce is a parallel programming framework designed for distributed processing. It divides the workload into smaller workloads which are distributed.

3. (d) Low-cost

Grid environment provides lower-cost storage environment but not that much significant comparative to other parallel alternatives.

4. (c) Massively Parallel Processing

Massively parallel processing platform combines storage, memory and computes to create a platform. MPP is used for known high-value use cases.

5. (d) Research organizations

HPC is designed for high-speed processing. HPC environments are mostly used by research organizations and business units which require very high computational performance.

6. (b) UDF

In-database analytics follow two methods. The first one is collection of user-defined-extensions (SQL extensions) to execute the code in database. The second one is parallelized analytics that execute in the data base. There are three types of user-defined extensions in which UDF performs a task and returns a value.

7. (d) Less content available

The most common and popular storage for big data is data warehouse. It makes the content highly available.

8. (b) I/O bound

I/O bound means not having enough bandwidth to meet the needs of the business. For example, trying to squeeze 50 TB of big data in five minutes through a 10 inch diameter pipeline.

9. (e) Flash memory

SDD is also called as flash memory, which represents the underlying chips used in this technology.

10. (c) GPGPU

GPUs (Graphical processing units) are used for video processing in PCs whereas GPGPUs (General purpose graphical processing unit) are mostly used for vector processing in big data analytics.

Big Data, Cloud and Analytics

Course Structure

Block 1: Introduction and Applications of Big Data	
Unit 1	What is Big Data?
Unit 2	Why Big Data is Important?
Unit 3	Big Data in Marketing & Advertising
Unit 4	Big Data in Healthcare
Block 2: Cloud Computing and Big Data Technologies	
Unit 5	Big Data and Cloud Technologies
Unit 6	Big Data Technologies and Terminologies
Unit 7	Cloud Computing and Big Data Management for Decision Making
Unit 8	Handling Unstructured Data
Unit 9	Information Management
Block 3: Business Analytics	
Unit 10	Analytics
Unit 11	Business Analytics-I
Unit 12	Business Analytics-II
Block 4: Managing Talent for Big Data Analytics	
Unit 13	Talent Management-I
Unit 14	Talent Management-II
Block 5: Data Privacy and Analytics in Various Business Areas	
Unit 15	HR Analytics in HR Planning
Unit 16	Data Analytics for Top Management Decision Making
Unit 17	Business and Marketing Intelligence Using Analytics
Unit 18	Data Privacy and Ethics