SREENIVASA INSTITUTE OF TECHNOLOGY AND MANAGEMENT STUDIES
(AUTONOMOUS)
MCA DEPARTMENT

LECTURE NOTES

Subject Name: BIG DATA ANALYTICS

Year / Branch: II MCA – I SEMESTER

Regulation: R24

Prepared By: Mr. N P Gangadhar,

Assistant Professor

II MCA – I SEMESTER			
COURSE CODE:	**24MCA213**	**CREDITS:**	**4**
COURSE TITLE:	**BIG DATA ANALYTICS**	**L-T-P:**	**4-0-0**

PREREQUISITES: *Courses on "Database Management Systems" , "Object Oriented Programming through JAVA" and knowledge on Intelligence Techniques.*

COURSE EDUCATIONAL OBJECTIVES*:*

CEO1 *To explore the fundamental concepts of Big Data.*
CEO2 *To Learn Basic concepts of Hadoop.*
CEO3 *To Write Hadoop MapReduce Programs for analyzing Big data.*
CEO4 *To Explore Hadoop Ecosystem.*

UNIT - I: UNDERSTANDING BIG DATA — **Lecture Hrs:10**

What is Big Data, Concepts and Terminology - Datasets, Data Analysis, Data Analytics, Big Data Characteristics – volume, velocity, variety, veracity, value, Different Types of Data – Structured Data, Unstructured Data, Semi-Structured Data, Case Study Background.

UNIT - II : HADOOP BASICS — **Lecture Hrs:10**

Brief history of Hadoop, Apache hadoop and the hadoop ecosystem. A weather dataset, analyzing the data with unix tools, analyzing the data with hadoop , Understanding different Hadoop modes, understanding Hadoop Features-Understanding HDFS, Understanding MapReduce, Learning the HDFS and Mapreduce Architecture-Understanding the HDFS architecture, Understanding the MapReduce Architecture,
Understanding the HDFS and MapReduce architecture by plot.

UNIT - III: WORKING WITH PIG AND HIVE — **Lecture Hrs:12**

Pig -Execution Types, An Example, Pig Latin-Structure, Statements,Types, Schemas, Functions, Data Processing Operators.Hive – An example, Tables –Managed Tables and External tables, Partitions and Buckets, Importing data, Altering Data, Dropping Tables, Querying Data-Sorting and Aggragating,
Mapreduce Scripts.

UNIT - IV : INTRODUCTION TO APACHE SPARK — **Lecture Hrs:12**

What is Apache Spark, Characteristics – Speed, Easy of Use, Modularity, Extensibility, Apache Spark Components as a Unified Stack, Spark Basic Data Types, Spark Core and SQL- Dynamic Partition Pruning, SQL Operations, Schemas and Creating Data Frames

UNIT - V: HBASE, ZOOKEEPER, SQOOP — **Lecture Hrs:12**

HBase Overview – Limitations of Hadoop, what is HBase,HBase and HDFS,StorageMechansim in HBase,Features of HBase, Applications of HBase. ZooKeeper Overview – what is ZooKeeper, Distributed Application, Benefits of Distributed Applications,Challenges of Distributed Applications, What is Apache
Zookeeper meant for, Benefits of ZooKeeper

TEXT BOOKS:

1. *Big Data Fundamentals: Concepts, Drivers & Techniques", 1/e, 2016, Thomas Erl, Wajid Khattak, Paul Buhler, Prentice Hall.*
2. *"Hadoop:The Definitive Guide," 3/e, 2012, Tom White, O'REILLY Publications.*
3. *"Learning Spark"*

REFERENCE BOOKS:

1. *Taming the Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics , 2012, Bill Franks, John Wiley & Sons..*

2. *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data, 2012, Paul C. Zikopoulos, Chris Eaton, Dirk deRoos, Thomas Deutsch, George Lapis, McGraw-Hill.*

3. *Intelligent Data Analysis, 2007, Michael Berthhold, David J.Hand, Springer.*

4. *Understanding Big Data: Analytics for Enterprise Class Hadoop and streaming data, 2011, Paul Zikopoulos, Chris Eaton, Paul Zikopoulos, McGraw hill.*

5. *Big Data for Dummies, 2012, Hurwitz, Alan Nugent, Dr. Fern Halper, Marcia Kaufman,John Wiley & Sons*

COURSE OUTCOMES: *On successful completion of this course, students will be able to:*		POs related to COs
CO1	**Realize** characteristics of Big Data and various types of data like structured, unstructured and semistructured	PO1,PO2
CO2	**Understand** two major components of Hadoop	PO1,PO2,PO3,PO4, PO5
CO3	**Analyze** Big data using Hadoop Map Reduce programs	PO1, PO2,PO4, PO5
CO4	get **Acquainted** with two Data Access Components of Hadoop Ecosystem called pig and hive	PO1,PO3, PO4
CO5	Acquire **Knowledge** on Data Storage Component of Hadoop Ecosystem called Hbase, Data Integration Component of Hadoop Ecosystem like sqoop and Monitoring, Management and Orchestration of Hadoop Ecosystem component like zookeeper	PO1,PO2, PO3,PO4

CO-PO MAPPING(DETAILED; HIGH:3; MEDIUM:2; LOW:1)

Course	**POs** / **COs**	**PO1**	**PO2**	**PO3**	**PO4**	**PO5**	**PO6**	**PO7**	**PO8**
C303: Big Data Analytics	**C303.1**	**3**	**2**	-	-	-	-	-	-
	C303.2	**3**	**3**	**2**	**3**	**2**	-	-	-
	C303.3	**3**	**3**	-	**3**	**3**	-	-	-
	C303.4	**3**	-	**3**	**3**	-	-	-	-
	C303.5	**3**	**3**	**2**	**3**	-	-	-	-
	C303	**3**	**2.75**	**2.3**	**3**	**2.5**	-	-	-

UNIT I - UNDERSTANDING BIG DATA

"Data is a precious thing and will last longer than the systems themselves."

Tim Berners-Lee

Unit Overview

This unit introduces the fundamentals of Big Data, its concepts, terminology, and importance in modern computing. It explains datasets, data analysis, and data analytics, along with the characteristics of Big Data such as volume, velocity, variety, veracity, and value. The unit also discusses different types of data including structured, unstructured, and semi-structured data, and presents real-world case studies to understand Big Data applications.

Learning Outcomes

After completing this unit, students will be able to:

1. Define Big Data and explain its significance.
2. Describe datasets, data analysis, and data analytics.
3. Explain the characteristics of Big Data (5Vs).
4. Differentiate between structured, unstructured, and semi-structured data.
5. Identify real-world applications of Big Data.
6. Analyze case studies involving Big Data solutions.
7. Understand the role of Big Data in decision-making and business intelligence.

Importance of Studying this Unit

Big Data has become a critical component of modern organizations. Businesses, governments, and researchers generate massive amounts of data every day. Understanding Big Data helps in extracting meaningful insights, improving decision-making, predicting trends, and solving complex problems. Knowledge of Big Data is essential for careers in Data Science, Artificial Intelligence, Machine Learning, Business Analytics, and Cloud Computing.

Key Concepts

Big Data, Dataset, Data Analysis, Data Analytics, Volume, Velocity, Variety, Veracity, Value, Structured Data, Unstructured Data, Semi-Structured Data, Data Processing, Data Storage, Business Intelligence, Predictive Analytics, Case Study.

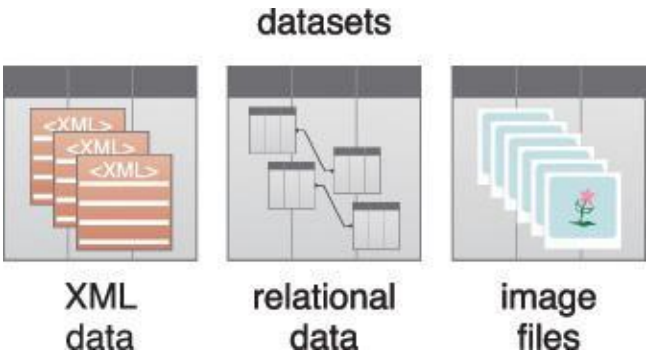
What is Big Data?

Big Data refers to extremely large and complex datasets that cannot be processed efficiently using traditional data processing tools and techniques. (or) Data that satisfies 5 V's is called Big Data

Concepts and Terminology

Datasets:

- Collection or group of related data are generally refer to as Dataset.
- Each group(datum) shares the same set of attributes as others in the same dataset.
- Some of the examples are:
  - tweets stored in a flat file
  - a collection of image files in a directory.
  - an extract of rows from a database table stored in a CSV formatted file
  - historical weather observations that are stored as XML files



Datasets can be found in many different formats

What is Data Analysis

- Data analysis is the process of examining data to find facts, relationships, patterns, insights and/or trends.
- The overall goal of data analysis is to support better decision-making.
- A simple data analysis example is the analysis of ice cream sales data in order to determine how the number of ice cream cones sold is related to the daily temperature.
- The results of such an analysis would support decisions related to how much ice cream a store should order in relation to weather forecast information.
- Following figure shows the symbol used to represent data analysis.

What is Data Analytics

- Data analytics is a broader term that encompasses data analysis.
- Data analytics is a discipline that includes the management of the complete data lifecycle, which encompasses collecting, cleansing, organizing, storing, analyzing and governing data.
- Following figure shows the symbol used to represent analytics.

There are four general categories of analytics that are distinguished by the results they produce:

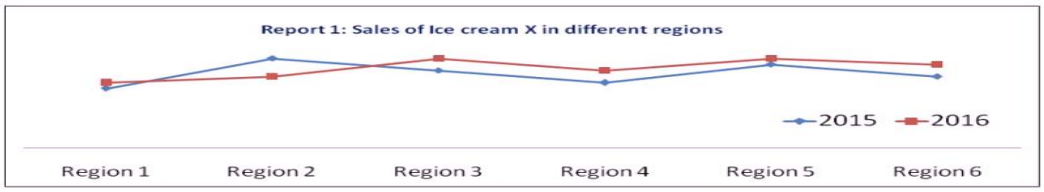
1. **Descriptive Analytics**
2. **Diagnostic Analytics**
3. **Predictive Analytics**
4. **Prescriptive Analytics**

Descriptive Analytics

Descriptive Analytics gives you information on what is happening. Let us see a particular Example

Descriptive Analytics

What is happening?



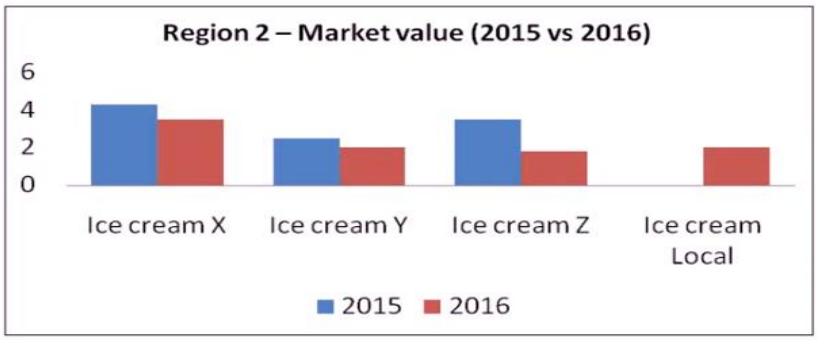
This report gives us information about sales of ice cream x in different regions and its also compared these sales for years 2015 and 16.This is a typical descriptive analytics support which gives you information on what is happening to sales of this ice cream so this chart could infer that except for the region 2 the sales have increased from 2015 to 16 so this is the level of information one could obtain from descriptive analytics now that we know the sales have dropped in region 2 .

Diagnostic Analytics

Information on why is it happening . so our job is to deep dive into region 2 and see what exactly is going wrong for us so this is a typical report for the market share of region 2

Diagnostic Analytics

Why is it happening?



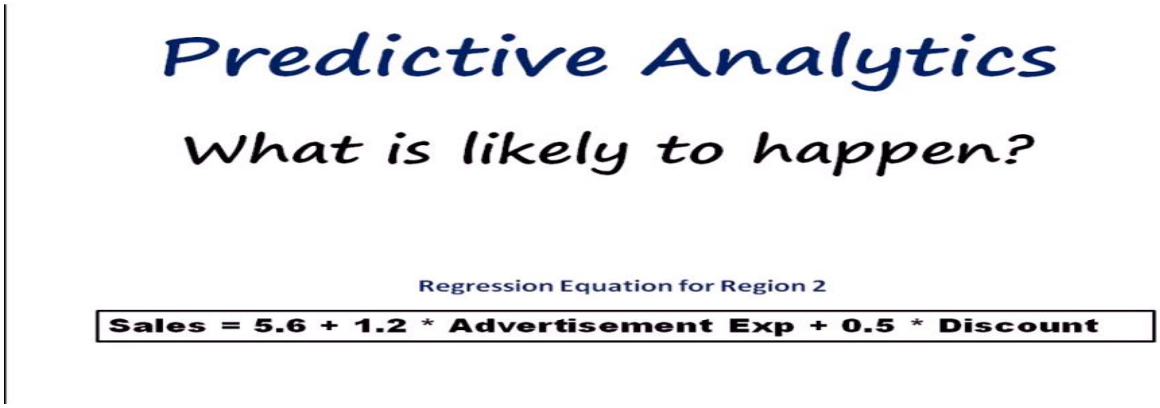
From the graph , it is clear that the sales of ice cream brands x, y and z is dropped from 2015 to 2016 and it is because the local brand ice cream has come up in 2016.In the first step we

found what is happening and here we found why it is happening. here we have diagnosed the reason of sales drop in region 2

Predictive Analytics

Once we identified the reason of happenings , then we can take measures to improve things i.e what should be done to make things happen correct. In our example, first we identified region 2 has less sales in 2016 next, we found the reason of why region 2 sales got down and now it is the time to find the solution to increase the sales and that is called predictive data analytics

For this example what we can do is to build a regression for region 2 sales the different factors that would affect the sales of region 2 this would help us change the different factors say for example the advertisement expense or discount and see what would be the impact of the changes on sales so these are some things that u can do so using predictive analytics which essentially tells you what is likely to happen in future now we know that what had happened and why it had happened and our next job is to find what is likely to happen i.e predictive analytics



Prescriptive Analytics

- Our next step is to move to next topic i. e Prescriptive analysis there you suggest the best course of action to solve the problem.
- for this particular example we can give solutions like increase the advertising expense by 10% or give 5 % discount for two months
- if we are able to find out using the regression equation that these factors like advertising expense or discount has an impact on sales
- so these are the four different levels of business analytics .
- Now lets summarize what we have just seen

Summary

Four levels of Analytics:

1. Descriptive – What is happening?
2. Diagnostic – Why is it happening?
3. Predictive – What is likely to happen?
4. Prescriptive – What is the best course of action?

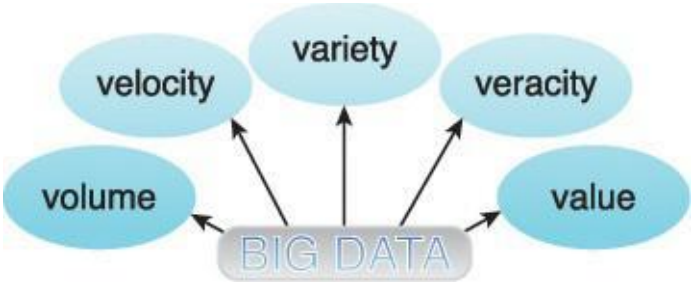
Big Data Characteristics

1) **Volume**
2) **Velocity**
3) **Variety**
4) **Veracity**
5) **Value**

- For a dataset to be considered Big Data, it must possess one or more characteristics like volume, velocity,variety,veracity and value

(or)

- We differentiate Big Data from traditional data by one or more of the five V's: Volume, Velocity,Variety, Veracity and value
- Five Big Data Characeristics that can be used to help differentiate data categorized as "Big" from other forms of data are



Volume

- Refers to the vast amounts of data generated every second.

- we are not talking about Terabytes but zettabytes.
- If we take all the data generated in the world between the beginning of time and 2008, the same amount of data will soon be generated every minute.
- This makes most data sets too large to store and analyse using Traditional database technology.
- Example: Amazon handles 15 million customer click stream user data per day to recommend products.

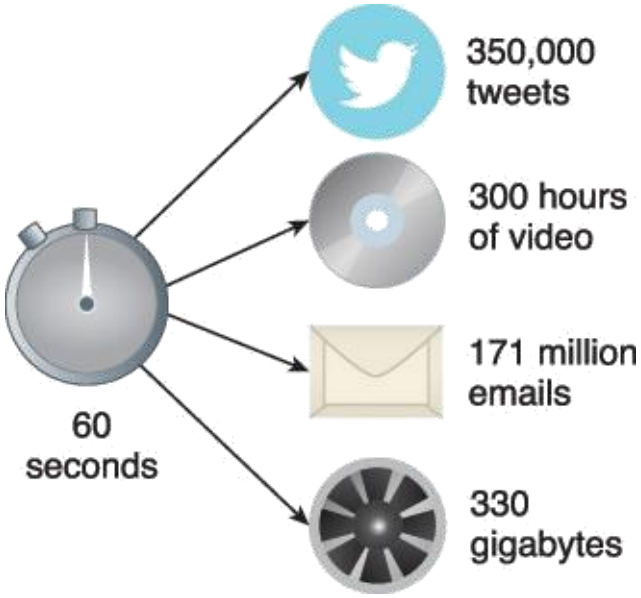
Typical data sources that are responsible for generating high data volumes can include:

- online transactions, such as point-of-sale and banking
- sensors, such as GPS sensors, RFIDs, smart meters and telematics social media, such as Facebook and Twitter

Velocity:

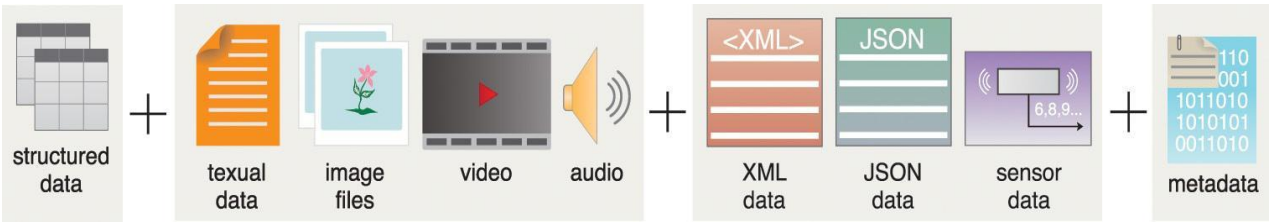
- Refers to the speed at which new data is generated and the speed at which data moves around. Just think of social media messages going viral in seconds

Example: 72 hours of video are uploaded to YouTube every minute this is the velocity. Extremely high velocity of data is another major characteristic of big data.



Variety

Data variety refers to the multiple formats and types of data that need to be supported by Big Data solutions. Data variety brings challenges for enterprises in terms of data integration, transformation, processing, and storage. Following figure provides a visual representation of data variety, which includes structured data in the form of financial transactions, semi-structured data in the form of emails and unstructured data in the form of images.

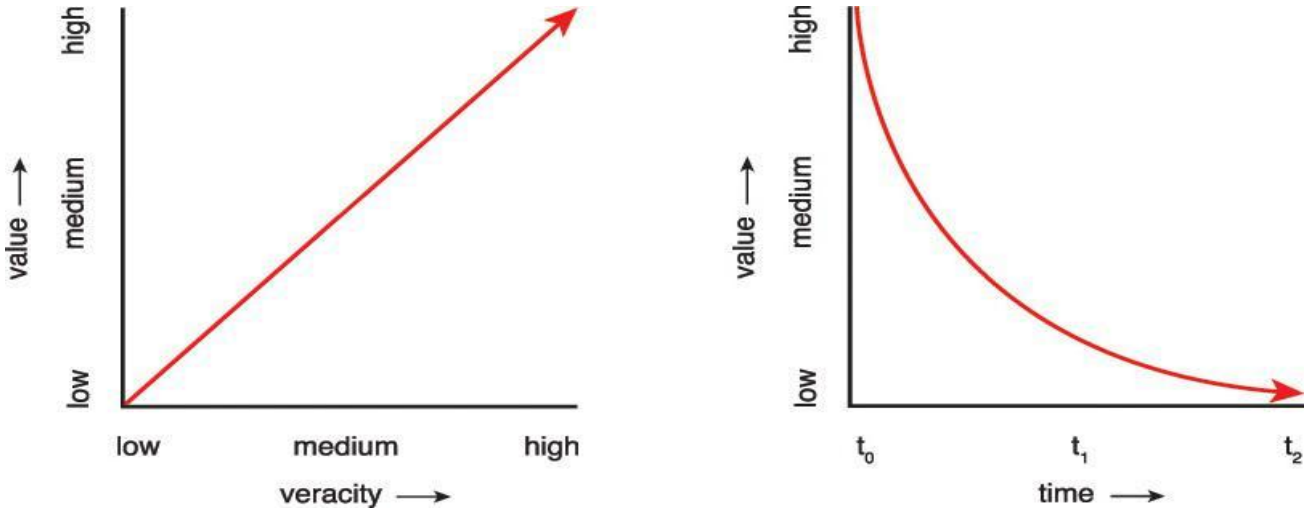


Veracity

Veracity refers to the quality or fidelity of data. Data that enters Big Data environments needs to be assessed for quality, which can lead to data processing activities to resolve invalid data and remove noise. In relation to veracity, data can be part of the signal or noise of a dataset. Noise is data that cannot be converted into information and thus has no value, whereas signals have value and lead to meaningful information. Data with a high signal-to-noise ratio has more veracity than data with a lower ratio. Thus the signal-to-noise ratio of data is dependent upon the source of the data and its type.

Value

Value is defined as the usefulness of data for an enterprise. The value characteristic is intuitively related to the veracity characteristic in that the higher the data fidelity, the more value it holds for the business.



Different Types of Data

The data processed by Big Data solutions can be human-generated or machine-generated, although it is ultimately the responsibility of machines to generate the analytic results. Human-generated data is the result of human interaction with systems, such as online services and digital devices. Following figure shows examples of human-generated data.

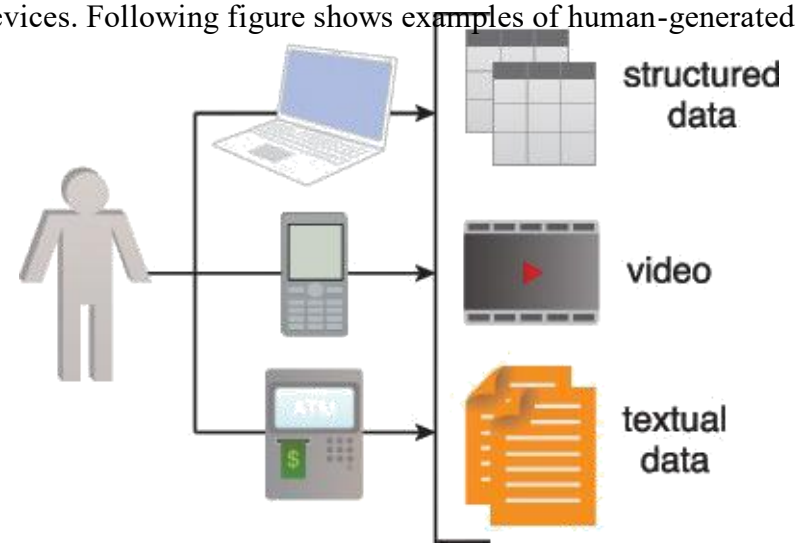


Fig: Examples of human-generated data include social media, blog posts, emails, photo sharing and messaging.

Machine-generated data is generated by software programs and hardware devices in response to real-world events.

As demonstrated, human-generated and machine-generated data can come from a variety of sources and be represented in various formats or types. This section examines the variety of data types that are processed by Big Data solutions. The primary types of data are:

Structured data

Unstructured data

Semi-structured data

Structured Data

Structured data conforms to a data model or schema and is often stored in tabular form. It is used to capture relationships between different entities and is therefore most often stored in a relational database. Structured data is frequently generated by enterprise applications and information systems like ERP and CRM systems. Due to the abundance of tools and databases that natively support structured data, it rarely requires special consideration in regards to processing or storage. Examples of this type of data include banking transactions, invoices, and customer records. Following figure shows the symbol used to represent structured data.

Fig: The symbol used to represent structured data stored in a tabular form.

Unstructured Data

Data that does not conform to a data model or data schema is known as unstructured data. It is estimated that unstructured data makes up 80% of the data within any given enterprise. Unstructured data has a faster growth rate than structured data. Following figure illustrates some common types of unstructured data. This form of data is either textual or binary and often conveyed via files that are self-contained and non-relational.

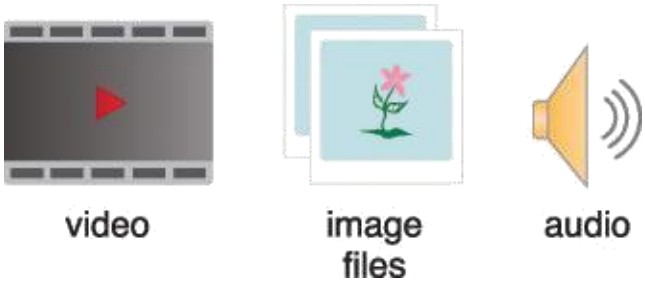


Fig: Video, image and audio files are all types of unstructured data.

Semi-structured Data

Semi-structured data has a defined level of structure and consistency, but is not relational in nature. Instead, semi-structured data is hierarchical or graph-based. This kind of data is commonly stored in files that contain text. For instance following figure shows that XML and JSON files are common forms of semi-structured data. Due to the textual nature of this data and its conformance to some level of structure, it is more easily processed than unstructured data.

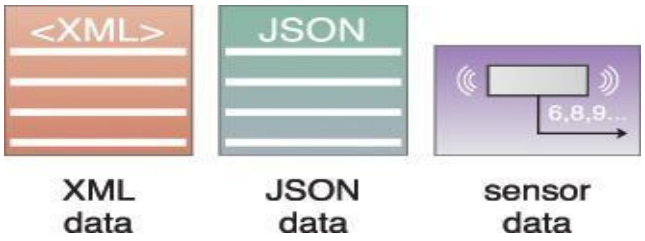


Fig: XML, JSON and sensor data are semi-structured.

Unit Highlights

- Big Data deals with large and complex datasets.
- Data analysis converts raw data into meaningful information.
- Data analytics supports data-driven decision-making.
- The 5Vs of Big Data are Volume, Velocity, Variety, Veracity, and Value.
- Data can be structured, unstructured, or semi-structured.
- Big Data technologies are widely used in healthcare, finance, e-commerce, and social media.
- Case studies demonstrate how organizations use Big Data to gain competitive advantages.

Case Study: Big Data in Amazon E-Commerce Platform

Title

Case Study: How Amazon Uses Big Data to Improve Customer Experience and Business Performance

TEXT BOOKS:

1. Big Data Fundamentals: Concepts, Drivers & Techniques", 1/e, 2016, Thomas Erl, Wajid Khattak, Paul Buhler, Prentice Hall.
2. "Hadoop:The Definitive Guide," 3/e, 2012, Tom White, O'REILLY Publications.
3. "Learning Spark"

REFERENCE BOOKS:

1. Taming the Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics, 2012, Bill Franks, John Wiley & Sons..
2. Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data, 2012, Paul C. Zikopoulos, Chris Eaton, Dirk deRoos, Thomas Deutsch, George Lapis, McGraw-Hill.
3. Intelligent Data Analysis, 2007, Michael Berthhold, David J.Hand, Springer.
4. Understanding Big Data: Analytics for Enterprise Class Hadoop and streaming data, 2011, Paul Zikopoulos, Chris Eaton, Paul Zikopoulos, McGraw hill.
5. Big Data for Dummies, 2012, Hurwitz, Alan Nugent, Dr. Fern Halper, Marcia Kaufman,John Wiley & Sons

S.No	Questions
UNIT I - Understanding Big Data	
PART-A (Two Marks Questions)	
1.	Define Dataset
2.	Distinguish between Data Analysis and Data Analytics
3.	Mention Different types of Data Analytics
4.	Explain Descriptive Analytics with an example
5.	Distinguish between Predictive and Prescriptive Analytics
6.	Define Big Data
7.	List out the Big Data Characteristics
8.	Mention different types of Data
9.	Mention some of Big Data Examples
10.	Give an example for structured and unstructured data
PART-B (Ten Marks Questions)	
1.	Explain Different types of Data Analytics with examples
2.	What is BigData and Discuss different characteristics of Big Data
3.	Illustrate different types of Data
4.	Differentiate Descriptive and Diagnostics analytics with example
5.	Differentiate Structured, unstructured and semi structured data in detail
6.	Explain Dataset, Data Analysis and Data analytics in detail