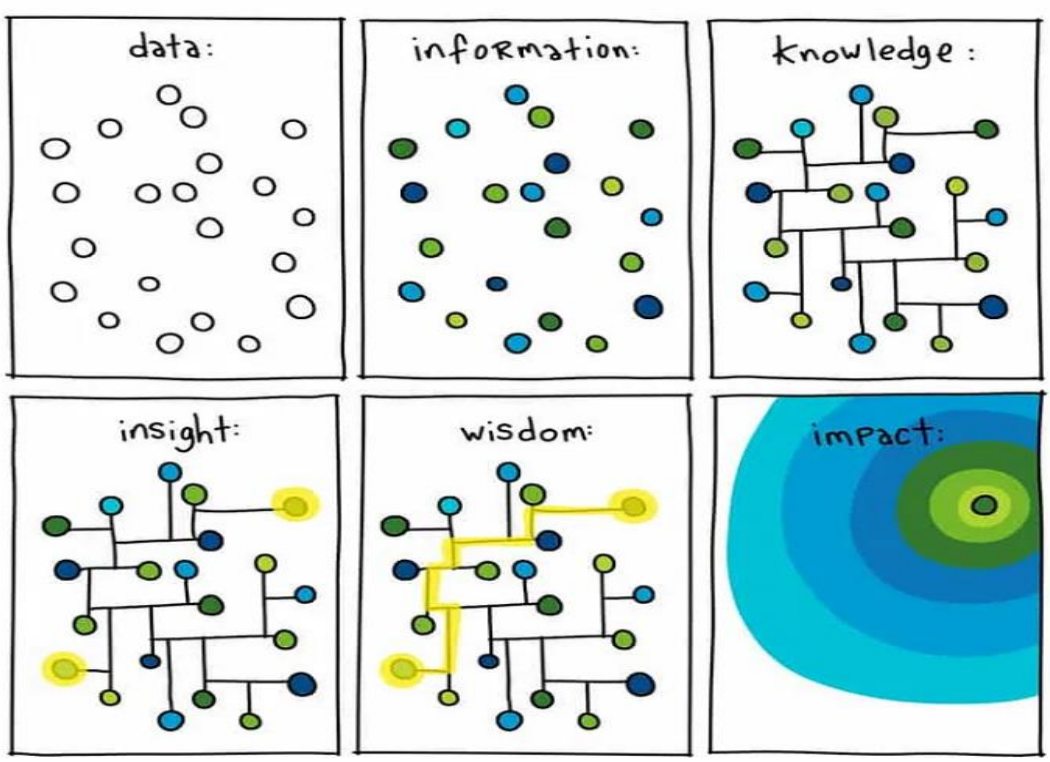
# INTRODUCTIONTO BIG DATA ANALYTICS

***Syllabus:***

***Types of Digital Data:*** *Classification of Digital data: Structured Data, Semi Structured Data, Unstructured Data; Introduction to Big Data: Characteristics of Data, Evolution of Big Data, Definition of Big Data, Challenges of Big Data, What is Big Data?, Why Big Data? Traditional Business Intelligence versus Big Data, A Typical Data Warehouse Environment, A Typical Hadoop Environment and Coexistence of Big Data and Data Warehouse.*

Today, **data** undoubtedly **is an invaluable asset** of any enterprise (big or small).Even though professionals work with data all the time, the understanding, management and analysis of data from heterogeneous sources remains a serious challenge. Data growth has seen exponential acceleration since the advent of the computer and Internet.




- Today, **data** is everywhere. It's generated by our devices, our interactions, and our activities. And it will only continue to grow in **volume, velocity, and variety**.
- The above image is visual representation of the progression from **raw data to impactful actions**, often described as the "**Data to Impact Journey**" or "**Information vs Knowledge**".
- This illustration represents the stages of transforming data into impact, visually showing the progression from raw data to impactful outcomes:
  1. **Data**: This panel shows scattered data points, representing raw data as it's initially collected. At this stage, the data lacks structure and context.
  2. **Information**: Here, some of the data points are connected, indicating that relationships are being established. This could represent the early stages of data organization and initial analysis, where data begins to have meaning but is still basic.
  3. **Knowledge**: Connections among the data points are more structured, forming complex networks. This suggests a deeper understanding and interpretation of the data, often involving methodologies to uncover patterns and relationships.

4. **Insight**: Strategic connections between specific data points are highlighted, suggesting that significant findings or valuable insights have been identified from the data. These insights often answer specific business or research questions.
5. **Wisdom**: Building on insights, wisdom involves applying the derived knowledge prudently to make informed decisions. This stage is about leveraging insights in practical, often innovative ways, and may also incorporate learning from past experiences.
6. **Impact**: The final panel shows a target-like design, symbolizing the strategic impact of actions taken based on the accumulated wisdom. It illustrates the achievement of desired outcomes, effectiveness in execution, and the realization of goals.

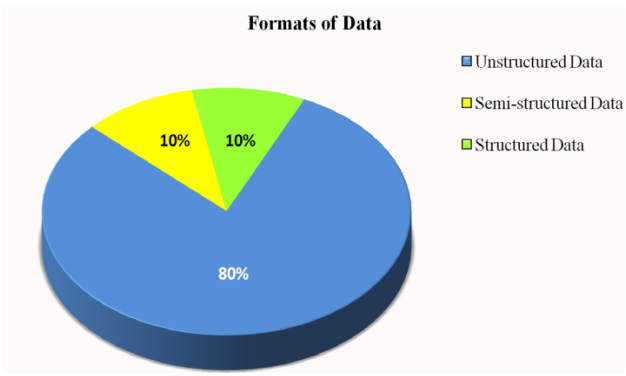
The progression through these panels emphasizes how raw data can be transformed into actionable, strategic decisions that lead to measurable impact, highlighting the importance of processing and analyzing data thoroughly.

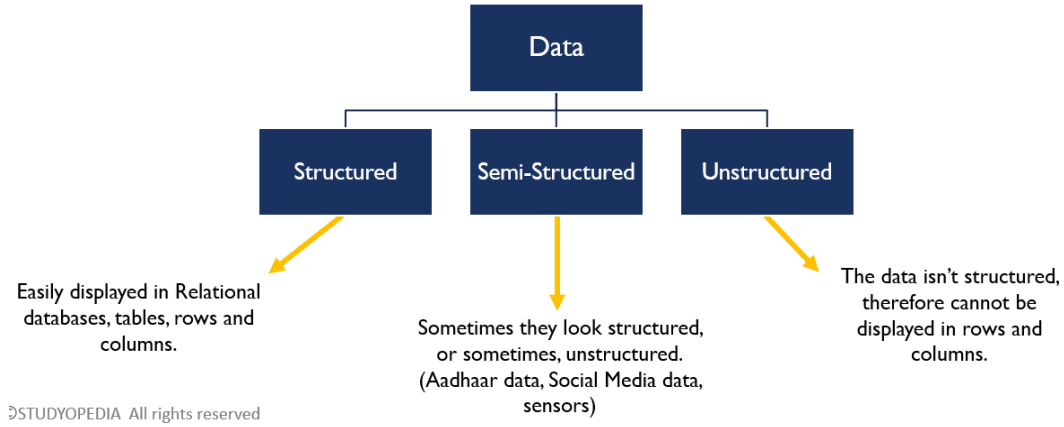
## Classification of Digital data

- Digital data can be classified into three forms:
  1. Unstructured
  2. Semi-structured
  3. Structured

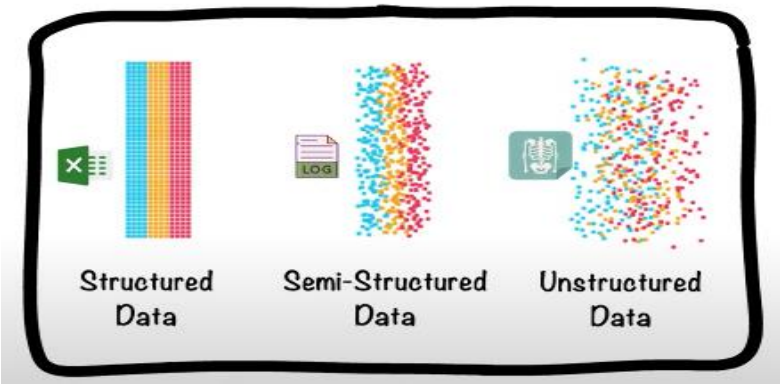
- Usually, data is in the unstructured format which makes extracting information from it difficult. According to Merrill Lynch, 80–90% of business data is either unstructured or semi-structured. Gartner also estimates that unstructured data constitutes 80% of the whole enterprise data.

Here is a percent distribution of the three forms of data -



**Example:**




## Structured Data

**Structured data: This is the data which is in an organized form** (e.g., in rows and columns) and can be easily used by a computer program. Relationships exist between entities of data, such as classes and their objects. Data stored in databases is an example of structured data.

- **Example:**
  - Most of the structured data is held in RDBMS.
  - An RDBMS conforms to the relational data model wherein the data is stored in rows/columns-Shown below :

| | Column 1 | Column 2 | Column 3 | Column 4 |
|---|---|---|---|---|
| Row 1 → | Student_USN | Student_Name | Student_Age | Student_address |
| Row 2 → | 1DA20CS001 | AAA | 21 | BANGALORE |
| Row 3 → | 1DA20CS002 | BBB | 21 | MANGALORE |
| Row 4 → | 1DA20CS005 | GGG | 21 | DELHI |

  - The number of rows/records/tuples in a relation is called the ***cardinality of a relation*** and the number of columns is referred to as ***the degree of a relation***.
- The **first step** is the design of a relation/table, the fields/columns to store the data, the type of data that will be stored [number (integer or real), alphabets, date, Boolean, etc.].
- **Next** we think of the constraints that we would like our data to conform to (constraints such as UNIQUE values in the column, NOT NULL values in the column).
- To explain further, let us design a table/relation structure to store the details of the employees of an enterprise.
- Table 1.2 shows the structure/schema of an "Employee" table in a RDBMS such as Oracle.
- Table 1.2 is an example of a good structured table (complete with table name, meaningful column names with data types, data length, and the relevant constraints) with absolute adherence to relational data model.

**Table 1.2** Schema of an "Employee" table in a RDBMS such as Oracle

| Column Name | Data Type | Constraints |
|---|---|---|
| EmpNo | Varchar(10) | PRIMARY KEY |
| EmpName | Varchar(50) | |
| Designation | Varchar(25) | NOT NULL |
| DeptNo | Varchar(5) | |
| ContactNo | Varchar(10) | NOT NULL |

- It goes without saying that each record in the table will have exactly the same structure. Let us take a look at a few records in Table 1.3.

Table 1.3 Sample records in the "Employee" table

| EmpNo | EmpName | Designation | DeptNo | ContactNo |
|---|---|---|---|---|
| E101 | Allen | Software Engineer | D1 | 0999999999 |
| E102 | Simon | Consultant | D1 | 0777777777 |

- The tables in an RDBMS can also be related. For example, the above "Employee" table is related to the "Department" table on the basis of the common column, "DeptNo". It is not mandatory for the two tables that are related to have exactly the same name for the common column. On the contrary, the two tables are related on the basis of values held within the column, "DeptNo". Given in Figure 1.3 is a depiction of referential integrity constraint (primary - foreign key) with the "Department" table being the referenced table and "Employee" table being the referencing table.

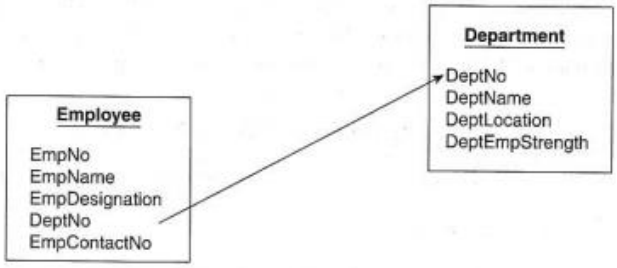
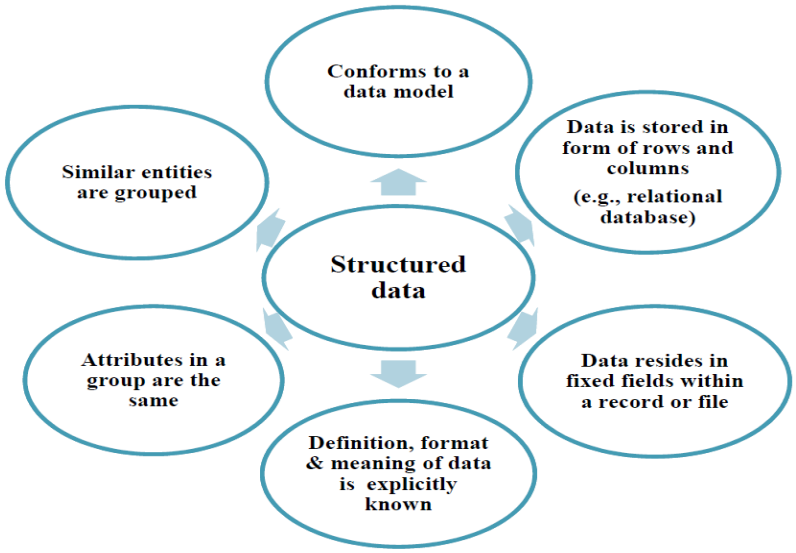



Figure 1.3 Relationship between "Employee" and "Department" tables.

## Charcteristics of Structured Data:




The characteristics of structured data include:

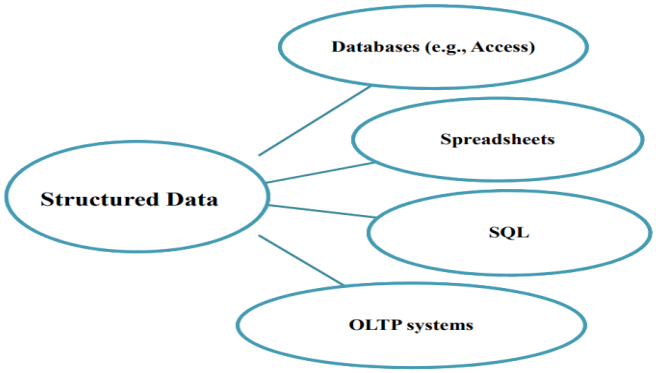
- The data conforms to a data model and has an easily identifiable structure
- The data is stored in rows and columns
- The data is well organized, which means that the definition, format, and the meaning of the data is known.
- The data lives in fixed fields inside a record or a file.
- Similar entities get grouped together to form relations or classes.
- Entities that belong to the same group will have the same attributes.

- The data is easy to access and query. Because of this, it is easy for other programs to make use of the data.
- The data elements are addressable, which means that analyzing and processing them becomes a rather efficient task.

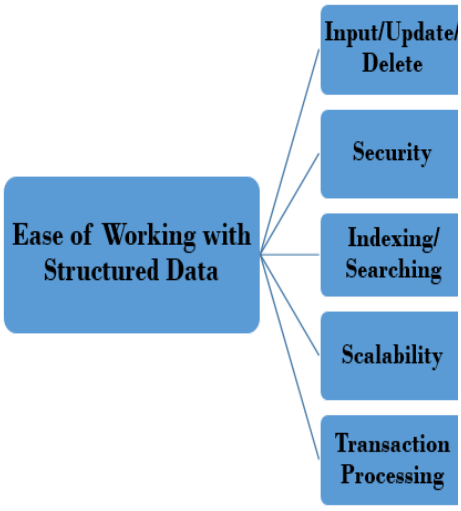
## Sources of Structured Data:

The sources of unstructured data include:

- SQL Databases
- Spreadsheets like Excel
- OLTP Systems
- Online forms
- Sensors like GPS or RFID tags
- Network and Web server logs
- Medical devices




## Ease with Structured Data-Storage:




Structured data provides the ease of working with it. Refer Figure. The ease is with respect to the following:

**1. Insert/update/delete:** The Data Manipulation Language (DML) operations provide the required ease with data input, storage, access, process, analysis, etc.

**2. Security:** Organizations are able to retain control and maintain compliance adherence by ensuring that only authorized individuals are able to decrypt and view sensitive information.

**3. Indexing:** An index is a data structure that speeds up the data retrieval operations at the cost of additional writes and storage space, but the benefits that ensue in search operation are worth the additional writes and storage space.

**4. Scalability:** The storage and processing capabilities of the traditional RDBMS can be easily scaled up by increasing the horsepower of the database server.

**5. Transaction processing:** RDBMS has support for Atomicity, Consistency, Isolation, and Durability (ACID) properties of transaction.

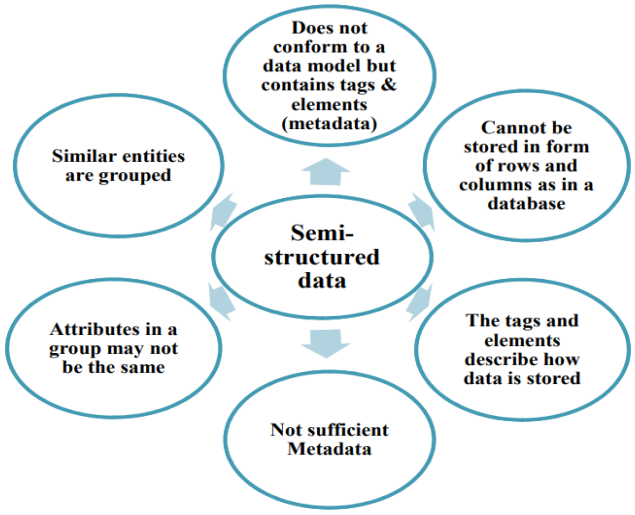
## Semi-Structured Data

- **Semi-structured data is the data which does not conform to a data model but has some structure**. However, it is not in a form which can be used easily by a computer program; for example, emails, XML, markup languages like HTML, etc. Metadata for this data is available but is not sufficient.
- Semi-structured uses tags and markers to segregate semantic elements.
  - which help to group data and describe how data is stored, giving some metadata but it is not sufficient for management and automation of data.
- Similar entities in the data are grouped and organized in a hierarchy. The attributes or the properties within a group may or may not be the same.

- For example two addresses may or may not contain the same number of properties as in
  Address 1
  <house number><street name><area name><city>
  Address 2
  <house number><street name><city>
- For example an e-mail follows a standard format
  To: <Name>
  From: <Name>
  Subject: <Text>
  CC: <Name>
  Body: <Text, Graphics, Images etc. >

- The tags give us some metadata but the body of the e-mail contains no format neither is such which conveys meaning of the data it contains.
- There is very fine line between unstructured and semi-structured data.

## Characteristics of Semi-Structured Data:



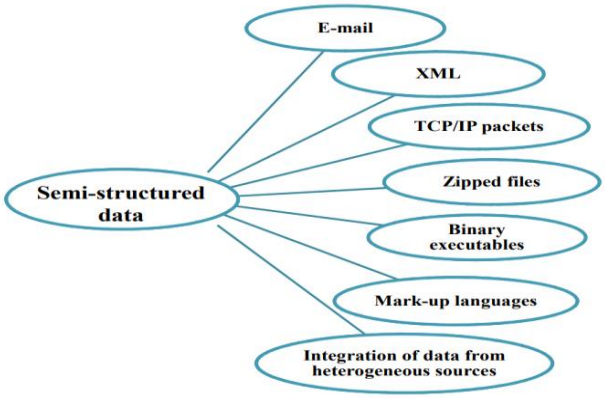

Characteristics of semi-structured data include:

- Data does not conform to a data model but has some structure.
- Data cannot be stored in the form of rows and columns as in Databases.
- Semi-structured data contains tags and elements (Metadata) which is used to group data and describe how the data is stored.
- Similar entities are grouped together and organized in a hierarchy.
- Entities in the same group may or may not have the same attributes or properties.
- Does not contain sufficient metadata which makes automation and management of data difficult.
- Size and type of the same attributes in a group may differ.
- Due to lack of a well-defined structure, it cannot used by computer programs easily.

## Sources of Semi-Structured Data:

The sources of unstructured data include:

- E-mails
- XML and other markup languages
- Binary executables
- TCP/IP packets
- Zipped files
- Integration of data from different sources
- Web pages




## Unstructured Data

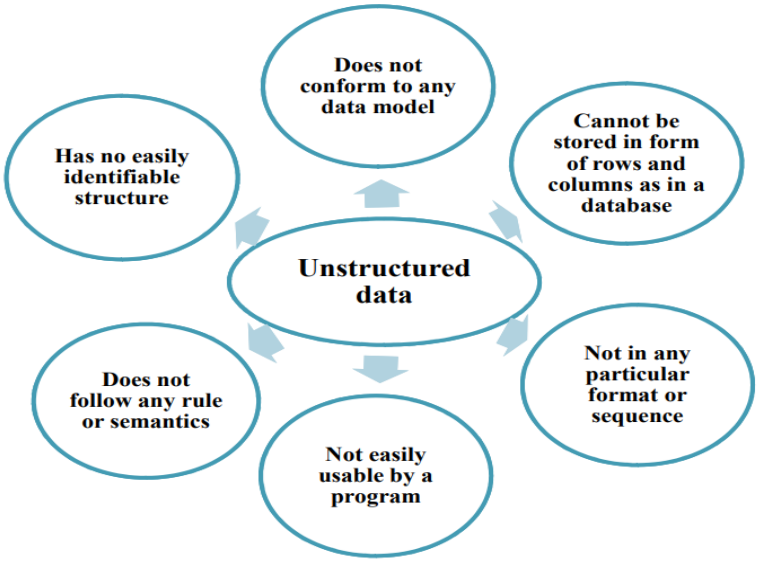
**Unstructured data: This is the data which does not conform to a data model or is not in a form which can be used easily by a computer program.** About 80—90% data of an organization is in this format; for example, memos, chat rooms, PowerPoint presentations, images, videos, letters, researches, white papers, body of an email, etc.

In fact, to explain things a little more, let us take a closer look at the various kinds of text available and the possible structure associated with it. As can be seen from the examples quoted in Table 1.4, the structure is quite unpredictable.

**Table 1.4** Few examples of disparate unstructured data

| | |
|---|---|
| Twitter message | Feeling miffed ☹. Victim of twishing. |
| Facebook post | LOL. C ya. BFN |
| Log files | 127.0.0.1 - frank [10/Oct/2000:13:55:36 -0700] "GET /apache_pb.gif HTTP/1.0" 200 2326 "http://www.example.com/start.html" "Mozilla/4.08 [en] (Win98; I; Nav)" |
| Email | Hey Joan, possible to send across the first cut on the Hadoop chapter by Friday EOD or maybe we can meet up over a cup of coffee. Best regards, Tom |

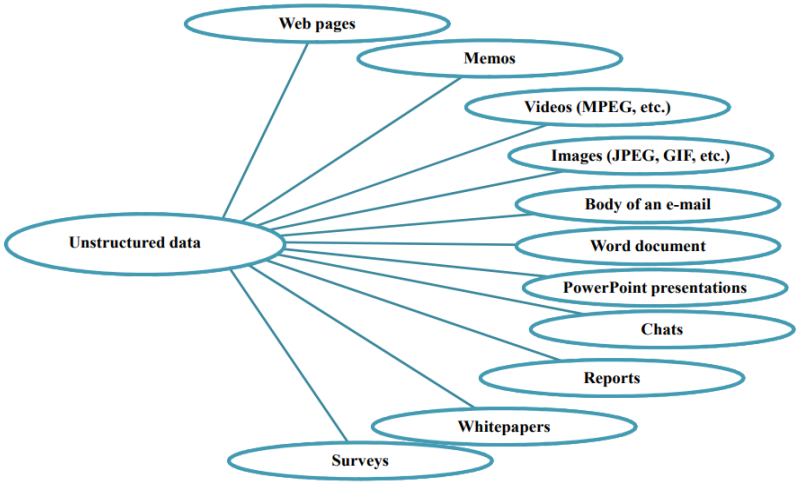
## Characteristics of Unstructured Data:



Unstructured data has these characteristics:

- Data neither conforms to a data model nor has any structure.
- Data cannot be stored in the form of rows and columns as in Databases
- Data does not follows any semantic or rules
- Data lacks any particular format or sequence
- Data has no easily identifiable structure
- Due to lack of identifiable structure, it cannot used by computer programs easily

## Sources of Unstructured Data:




**Text documents.** You may encounter unstructured data in the form of text documents, which can be plain text files (.txt), Microsoft Word documents (.doc, .docx), PDF files (.pdf), HTML files (.html), and other word processing formats. They primarily contain written content and may include elements like text, tables, and images.

**Emails.** As a form of electronic communication, emails often contain unstructured text data and various file attachments, such as images, documents, or spreadsheets.

**Images.** Image files come in various formats, such as JPEG (.jpg, .jpeg), PNG (.png), GIF (.gif), TIFF (.tiff), and more. These files store visual information and require specialized techniques such as computer vision to analyze and extract data.

**Audio files.** Audio data is usually presented in formats such as MP3 (.mp3), WAV (.wav), and FLAC (.flac), to name a few. These files contain sonic information that requires audio processing techniques to extract meaningful insights.

**Video files.** Video data comes in popular formats such as MP4 (.mp4), AVI (.avi), MOV (.mov), and others. Analyzing videos requires combining computer vision and audio processing techniques since they contain visual and auditory information.

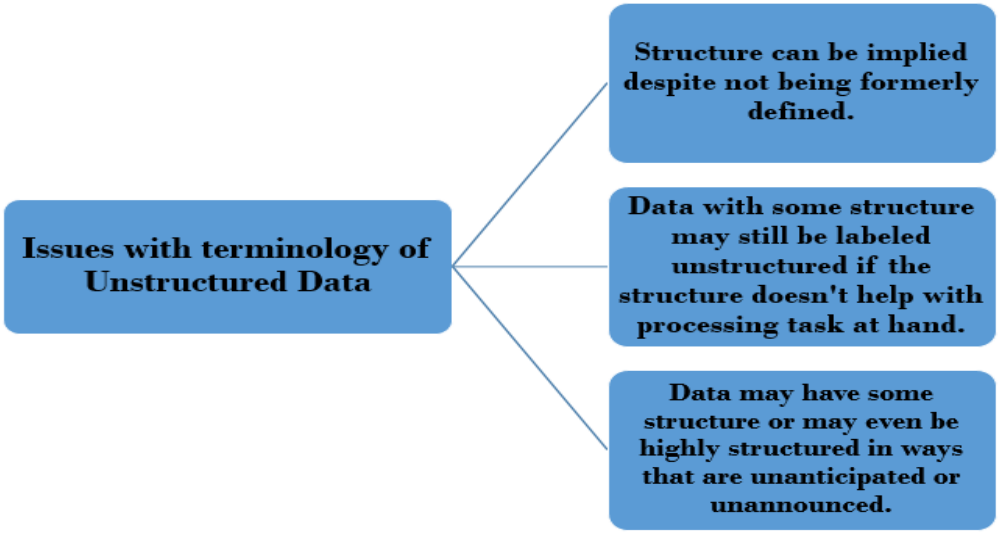
**Log files.** Generated by various systems or applications, log files usually contain unstructured text data that can provide insights into system performance, security, and user behavior.

**Sensor data.** Information from sensors embedded in wearable, industrial, and other IoT devices can also be unstructured, including temperature readings, GPS coordinates, etc.

**Social media posts.** Data from social media platforms, such as Twitter, Facebook, or messaging apps, contains text, images, and other multimedia content with no predefined structure to it.

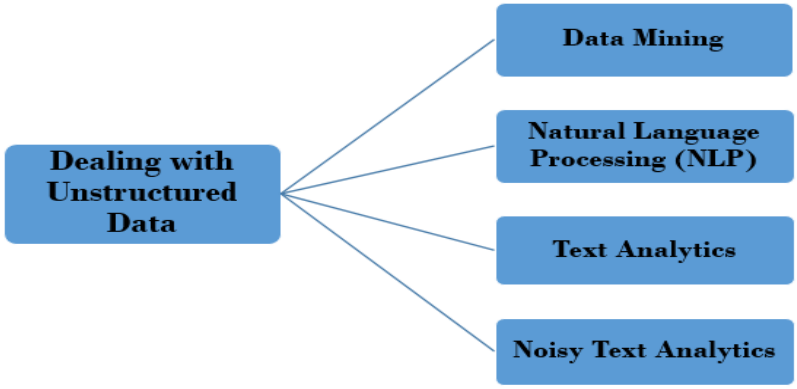
These are just a few examples of unstructured data formats. As the data world evolves, more formats may emerge, and existing formats may be adapted to accommodate new unstructured data types.

## Issues with Unstructured Data:




- The term "unstructured data" can be imprecise because data can have some structure, even if it's not formally defined. For example, data can be semi-structured or highly structured in ways that aren't obvious.
- The term is imprecise for several reasons:
  1. **Structure**, while not formally defined, can still be implied.
  2. Data with some form of structure may still be characterized as unstructured if its structure is not helpful for the processing task at hand.
  3. Unstructured information might have some structure (semi-structured) or even be highly structured but in ways that are unanticipated or unannounced.

## Dealing with unstructured data:




The following techniques are used to find patterns in or interpret unstructured data:

1. **Data mining:**
   - First, we deal with large data sets.
   - Second, we use methods at the intersection of artificial intelligence, machine learning, statistics, and database systems to unearth consistent patterns in large data sets and/or systematic relationships between variables.
   - It is the analysis step of the "knowledge discovery in databases" process.
   - Few popular data mining algorithms are as follows:
     - **Association rule mining:** It is also called "market basket analysis" or "affinity analysis". It is used to determine "What goes with what?" It is about

when you buy a product, what is the other product that you are likely to purchase with it. For example, if you pick up bread from the grocery, are you likely to pick eggs or cheese to go with it.

- **Regression analysis:** It helps to predict the relationship between two variables. The variable whose value needs to be predicted is called the dependent variable and the variables which are used to predict the value are referred to as the independent variables.
- **Collaborative filtering:** It is about predicting a user's preference or preferences based on the prefer ences of a group of users. For example, take a look at Table 1.5.
- We are looking at predicting whether User 4 will prefer to learn using videos or is a textual learner depending on one or a couple of his or her known preferences. We analyze the preferences of similar user profiles and on the basis of it, predict that User 4 will also like to learn using videos and is not a textual learner.

2. **Text analytics or text mining:**
   - Compared to the structured data stored in relational databases, text is largely unstructured, amorphous, and difficult to deal with algorithmically.
   - Text mining is the process of gleaning high quality and meaningful information (through devising of patterns and trends by means of statistical pattern learning) from text.
   - It includes tasks such as text categorization, text clustering, sentiment analysis, concept/entity extraction, etc.
3. **Natural language processing (NLP):**
   - It is related to the area of human computer interaction.
   - It is about enabling computers to understand human or natural language input.
4. **Noisy text analytics:**
   - It is the process of extracting structured or semi-structured information from noisy unstructured data such as chats, blogs, wikis, emails, message-boards, text messages, etc.
   - The noisy unstructured data usually comprises one or more of the following:
     - Spelling mistakes, abbreviations, acro- nyms, non-standard words, missing punctuation, missing letter case, filler words such as "uh"; "um", etc.
5. **Manual tagging with metadata:**
   - This is about tagging manually with adequate metadata to provide the requisite semantics to understand unstructured data.
6. **Part-of-speech tagging:**
   - It is also called POS or POST or grammatical tagging.
   - It is the process of reading text and tagging each word in the sentence as belonging to a particular part of speech such as "noun", "verb", "adjective", etc.
7. **Unstructured Information Management Architecture (UIMA):**
   - It is an open source platform from IBM.
   - It is used for real-time content analytics.

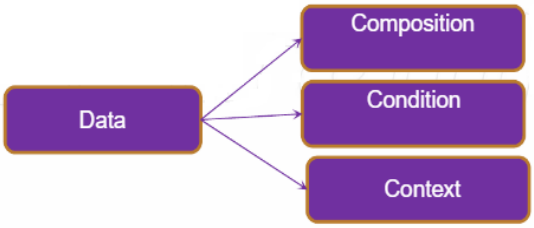
It is about processing text and other unstructured data to find latent meaning and relevant relationship buried therein.

# Introduction to Big Data

- **Data:** The quantities, characters, or symbols on which operations are performed by a computer, which may be stored and transmitted in the form of electrical signals and recorded on magnetic, optical, or mechanical recording media.
- **Small data:** (data as it existed prior to the big data revolution) is about certainty. It is about known data sources; it is about no major changes to the composition or context of data.
- **Big data:** is about complexity -complexity in terms of multiple and unknown data set, in terms of exploding volume, in terms of speed at which the data is been generated at the speed at which it needs to be processed, and in terms of variety of data (internal or external, behavioral or social ) that is been generated.

## Characteristics Of Data:

Data has 3 characteristics:




**1. Composition:** The composition of data deals with the structure of data, that is, the sources of data, the granularity, the types, and the nature of data as to whether it is static or real-time streaming.

**2. Condition:** The condition of data deals with the state of data, that is, "Can one use this data as is for analysis?" or "Does it require cleansing for further enhancement and enrichment?"

**3. Context:** The context of data deals with "Where has this data been generated?" "Why was this data generated?" How sensitive is this Data?", "What are the events associated with this data?" and so on.

## Evolution of Big Data

- 1970s and before was the era of mainframes. The data was essentially primitive and structured.
- Relational database evolved in 1980s and 1990s. the era was of data intensive applications.
- The World Wide Web (WWW) and the Internet of Things (IoT) have led to an onslaught of structured, unstructured and multimedia data.

| | Data Generation and Storage | Data Utilization | Data Driven |
|---|---|---|---|
| Complex and Unstructured | | | Structured, unstructured and multimedia data |
| Complex and Relational | | Relational databases: Data-intensive applications | |
| Primitive and Structured | Mainframes: Basic data storage | | |
| | 1970s and before | Relational (1980s and 1990s) | 2000s and beyond |

Table: The Evolution of Big Data

## What is Big Data?

- **Big Data** is a collection of data that is huge in volume, yet growing exponentially with time. It is a data with so large size and complexity that none of traditional data management tools can store it or process it efficiently.
- Big data is also a data but with huge size.
- "***Big data is high volume, high velocity and high variety*** information assets that demand cost efficient innovation forms of information processing for enhance insight and decision making".
- We will look definition in three parts:

  **Part I-** ***"Big data is high volume, high velocity and high variety information assets***" talks about volumnation data that may have great variety and will required a good speed/pace for storage preparation, the processing and analysis.

  **Part II-** "***Cost efficient innovation forms of information processing***" talks about embracing new techniques and technologies to capture, store , process, persist, integrate and visualize the high volume, high velocity and high variety data.

  **Part III-** "***Enhance insight and decision making***" talks about deriving deeper, richer and meaningful insights and then using this insights to make faster and better decision to gain business value.

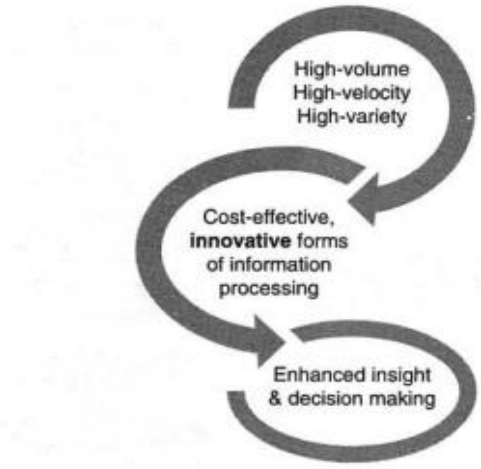



**Figure 2.3** Definition of big data – Gartner.

## Characteristics Of Big Data

**Big data can be described by the following characteristics:**

1. **Volume**
2. **Variety**
3. **Velocity**
4. **Veracity**
5. **Value**
6. **Validity**
7. ***Volatility***
8. **Variability**


### *(1) Volume*

- The name Big Data itself is related to a size which is enormous.
- Size of data plays a very crucial role in determining value out of data.
- Also, whether a particular data can actually be considered as a Big Data or not, is dependent upon the volume of data.
- Hence, **'Volume'** is one characteristic which needs to be considered while dealing with Big Data solutions.
- **Example:** In the year 2016, the estimated global mobile traffic was 6.2 Exabytes(6.2 billion GB) per month. Also, by the year 2020 we will have almost 40000 ExaBytes of data.

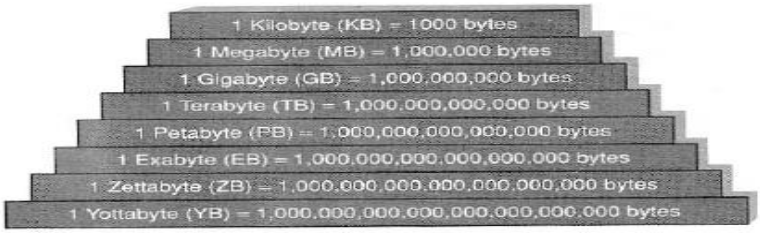



**Figure 2.6** A mountain of data.


**Figure 2.7** Sources of big data.

### *(2) Velocity*

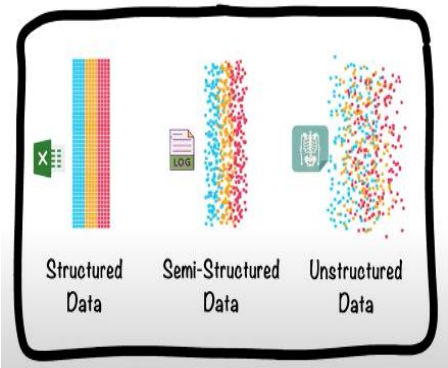
- The term **'velocity'** refers to the speed of generation of data.
- How fast the data is generated and processed to meet the demands, determines real potential in the data.
- Big Data Velocity deals with the speed at which data flows in from sources like business processes, application logs, networks, and social

media sites, sensors, Mobile devices, etc.

- The flow of data is massive and continuous.

Batch → Periodic → Near real time → Real-time processing

***(3) Variety***




- The next aspect of Big Data is its **variety**.
- Variety refers to heterogeneous sources and the nature of data, both structured and unstructured.
- During earlier days, spreadsheets and databases were the only sources of data considered by most of the applications.
- Nowadays, data in the form of emails, photos, videos, monitoring devices, PDFs, audio, etc. are also being considered in the analysis applications.
- This variety of unstructured data poses certain issues for storage, mining and analyzing data.


***(4) Veracity:*** Veracity refers more to the provenance or reliability of the data source, its context, and how meaningful it is to the analysis based on it. Big data can be messy, noisy, and error-prone, which makes it difficult to control the quality and accuracy of the data. Large datasets can be unwieldy and confusing, while smaller datasets could present an incomplete picture. The higher the veracity of the data, the more trustworthy it is.

***(5) Validity***: Similar to veracity, validity refers to how accurate and correct the data is for its intended use.


***(6) Value***: This refers to the value that big data can provide, and it relates directly to what organizations can do with that collected data.


***(7) Variability:*** Variability can refer to the inconsistent speed at which big data is loaded into your database. Data variability also known as spread or dispersion, refers to how spread out a set of data is. Variability gives users a way to describe how much data sets vary and allows users to use statistics to compare their data to other sets of data.

***(8) Volatility***: Volatility refers to the rate of change and lifetime of data. Organizations need to understand how long a specific type of data is valid.

## Challenges with big data

1. **Data Volume: Managing and Storing Massive Amounts of Data**
   - The most apparent challenge with Big Data is the sheer volume of data being generated. Organizations are now dealing with petabytes or even exabytes of data, making traditional storage solutions inadequate. This vast amount of data requires advanced storage infrastructure, which can be costly and complex to maintain.

**Solution:** Adopting scalable cloud storage solutions, such as ***Amazon S3, Google Cloud Storage, or Microsoft Azure.***

2. **Data Variety: Handling Diverse Data Types**
   - Big Data encompasses a wide variety of data types, including structured data (e.g., databases), semi-structured data (e.g., XML, JSON), and unstructured data (e.g., text, images, videos). The diversity of data types can make it difficult to integrate, analyze, and extract meaningful insights.

**Solution:** organizations can employ data integration platforms and tools like ***Apache Nifi, Talend, or Informatica.***

3. **Data Velocity: Processing Data in Real-Time**
   - The speed at which data is generated and needs to be processed is another significant challenge. For instance, IoT devices, social media platforms, and financial markets produce data streams that require real-time or near-real-time processing. Delays in processing can lead to missed opportunities and inefficiencies.

**Solution:** organizations can implement real-time data processing frameworks such as ***Apache Kafka, Apache Flink****, or* ***Apache Storm****.* Additionally, leveraging ***edge computing*** can help process data closer to its source, reducing latency and improving real-time decision-making.

4. **Data Veracity: Ensuring Data Quality and Accuracy**
   - With Big Data, ensuring the quality, accuracy, and reliability of data—referred to as data veracity—becomes increasingly difficult. Inaccurate or low-quality data can lead to misleading insights and poor decision-making. Data veracity issues can arise from various sources, including data entry errors, inconsistencies, and incomplete data.

**Solution:** Implementing robust data governance frameworks is crucial for maintaining data veracity. This includes establishing data quality standards, performing regular data audits, and employing data cleansing techniques. Tools like ***Trifacta, Talend Data Quality, and Apache Griffin*** can help automate and streamline data quality management processes.

5. **Data Security and Privacy: Protecting Sensitive Information**
   - As organizations collect and store more data, they face increasing risks related to data security and privacy. High-profile data breaches and growing concerns over data privacy regulations highlight the importance of safeguarding sensitive information.

**Solution:** To mitigate security and privacy risks, organizations must adopt comprehensive data protection strategies. This includes implementing encryption, access controls, and regular security audits.

6. **Data Integration: Combining Data from Multiple Sources**
   - Integrating data from various sources, especially when dealing with legacy systems, can be a daunting task. Data silos, where data is stored in separate systems without easy access, further complicate the integration process, leading to inefficiencies and incomplete analysis.

**Solution:** Data integration platforms like ***Apache Camel, MuleSoft, and IBM DataStage*** can help

streamline the process of integrating data from multiple sources.

**7. Data Analytics: Extracting Valuable Insights**

- The ultimate goal of Big Data is to derive actionable insights, but the complexity of analyzing large, diverse datasets can be overwhelming. Traditional analytical tools may struggle to scale, and the lack of skilled data scientists can further hinder the ability to extract meaningful insights.

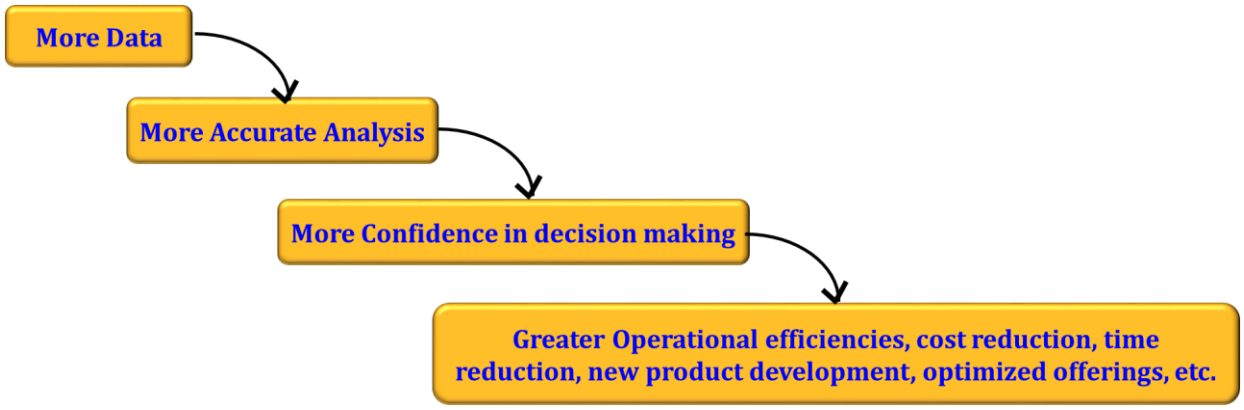
**Solution:** Organizations should invest in advanced analytics platforms like ***Apache Spark, Hadoop, or Google BigQuery,*** which are designed to handle large-scale data processing and analysis.

**8. Data Governance: Establishing Policies and Standards**

- As data becomes a critical asset, establishing effective data governance becomes essential. However, many organizations struggle with creating and enforcing policies and standards for data management, leading to issues with data consistency, quality, and compliance.

**Solution:** Implementing a formal data governance framework is key to overcoming this challenge. Tools like ***Collibra, Alation, and Informatica's*** data governance suite can assist in creating and maintaining a robust data governance strategy.

## Why Big data?/ Importance of Big data




The importance of big data does not revolve around how much data a company has but how a company utilizes the collected data. Every company uses data in its own way; the more efficiently a company uses its data, the more potential it has to grow. The company can take data from any source and analyze it to find answers which will enable:

**1. Cost Savings:** Some tools of Big Data like Hadoop and Cloud-Based Analytics can bring cost advantages to business when large amounts of data are to be stored and these tools also help in identifying more efficient ways of doing business.

**2. Time Reductions:** The high speed of tools like Hadoop and in-memory analytics can easily identify new sources of data which helps businesses analyzing data immediately and make quick decisions based on the learning.

**3. Understand the market conditions:** By analyzing big data you can get a better understanding of current market conditions. For example, by analyzing customers' purchasing behaviors, a company can find out the products that are sold the most and produce products according to this trend. By this, it can get ahead of its competitors.

**4. Control online reputation:** Big data tools can do sentiment analysis. Therefore, you can get

feedback about who is saying what about your company. If you want to monitor and improve the online presence of your business, then, big data tools can help in all this.

**5. Using Big Data Analytics to Boost Customer Acquisition and Retention:** The customer is the most important asset any business depends on. There is no single business that can claim success without first having to establish a solid customer base. However, even with a customer base, a business cannot afford to disregard the high competition it faces. If a business is slow to learn what customers are looking for, then it is very easy to begin offering poor quality products. In the end, loss of clientele will result, and this creates an adverse overall effect on business success. The use of big data allows businesses to observe various customer related patterns and trends. Observing customer behavior is important to trigger loyalty.

**6. Using Big Data Analytics to Solve Advertisers Problem and Offer Marketing Insights:**
Big data analytics can help change all business operations. This includes the ability to match customer expectation, changing company's product line and of course ensuring that the marketing campaigns are powerful.

## Business Intelligence (BI)

**What is Business Intelligence?**

- BI(Business Intelligence) is a set of processes, architectures, and technologies that convert raw data into meaningful information that drives profitable business actions.
- It is a suite of software and services to transform data into actionable intelligence and knowledge.
- BI has a direct impact on organization's strategic, tactical and operational business decisions.
- BI supports fact-based decision making using historical data.
- BI tools perform data analysis and create reports, summaries, dashboards, maps, graphs, and charts to provide users with detailed intelligence about the nature of the business.

**How Business Intelligence systems are implemented?**

Here are the steps:

**Step 1)** Raw Data from corporate databases is extracted. The data could be spread across multiple systems heterogeneous systems.

**Step 2)** The data is cleaned and transformed into the data warehouse. The table can be linked, and data cubes are formed.

**Step 3)** Using BI system the user can ask quires, request ad-hoc reports or conduct any other analysis.

## Traditional BI environment Vs from Big Data Environment

1. **Data sources:** Traditional BI typically relies on structured data from internal systems, while big data can come from a wide range of external and internal sources, including social media, IoT devices, and sensors.
2. **Data volume:** Traditional BI deals with relatively small amounts of data, while big data deals with extremely large and complex sets of data.
3. **Data variety:** Traditional BI typically deals with structured data, while big data can include structured, semi- structured, and unstructured data from various formats such as text, images, videos, and audio.
4. **Data processing:** Traditional Bl relies on traditional data processing techniques such as SQL,

while big data uses newer technologies such as Hadoop, NoSQL databases, and machine learning to process and analyze the data.

5. **Time frame:** Traditional BI focuses on historical data, while big data can also include real-time data streams.
6. **Scale:** Traditional BI is typically implemented in a small scale, while big data is implemented in large scale and distributed environments.

- While traditional BI and big data have different focuses and technologies, they are not mutually exclusive.
- In fact, big data can be used to enhance traditional BI by providing additional data sources and insights. Many organizations are now using big data technologies to complement and enhance their traditional Bl systems.

## A TYPICAL DATA WAREHOUSE ENVIRONMENT

- **Data warehouses** serve as a central repository for storing and analyzing information to make better informed decisions.
- In other words, **A data warehouse** is a centralized storage system that allows for the storing, analyzing, and interpreting of data in order to facilitate better decision-making.
- A typical **data warehouse environment** facilitates the storage and analysis of data drawn from various sources.
- By integrating various **data sources**, the data warehouse becomes a powerful tool for gaining insights across the entire spectrum of organizational data. The **outputs**, such as reporting, OLAP, and modeling, allow businesses to leverage this integrated data for comprehensive analysis, reporting, and predictive analytics, which are critical for maintaining competitive advantage.
- Here's a general overview of the key components and processes that you might find in a typical data warehouse environment:

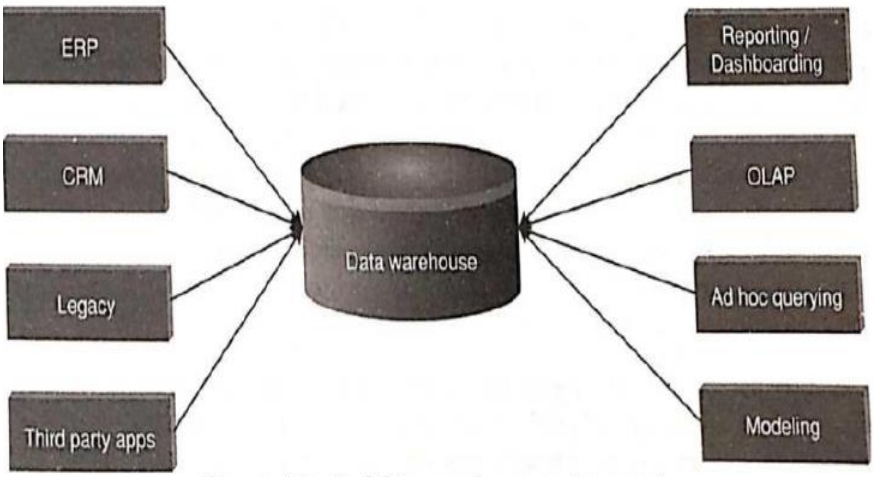



*Figure- A typical data warehouse environment.*

**Data Inputs:** Data comes from various sources such as Operational or transactional or day-to-day business data, which is gathered from:

1. **ERP (Enterprise Resource Planning):**
   - **Description:** Systems that integrate core business processes, such as finance, HR, manufacturing, supply chain, services, procurement, and others.
   - **Role in DW:** Provides comprehensive data from across the organization to support a unified view in the warehouse.

2. **CRM (Customer Relationship Management):**
   - **Description:** Systems used to manage a company's interactions with current and potential customers.
   - **Role in DW:** Supplies customer-related data, helping in analyses like customer segmentation, retention rates, and sales trends.

**3. Legacy Systems:**
   - **Description:** Older software or hardware systems that are still in use.
   - **Role in DW:** Often contains valuable historical data that is crucial for trend analysis and long-term reporting.

**4. Third Party Apps:**
   - **Description:** External applications that provide specialized services or data, such as market data, social media insights, etc.
   - **Role in DW:** Enhances the data warehouse with external perspectives, useful for comprehensive analytics.

**Data Outputs:**

**1. Reporting/Dashboarding:**

**Tools Used:** Business intelligence tools that provide visualizations and dashboards.

**Purpose:** To present data in an easily digestible format for decision-makers, often displaying key performance indicators (KPIs) and other critical metrics.

**2. OLAP (Online Analytical Processing):**

**Tools Used:** Specialized databases and tools designed for complex queries and analysis.

**Purpose:** Supports multi-dimensional queries, allowing users to drill down into data cubes to analyze data from multiple perspectives.

**3. Ad Hoc Querying:**

**Tools Used:** SQL tools and other query software that allow users to write and execute their queries.

**Purpose:** Enables users to perform spontaneous and one-time data analysis to answer specific business questions.

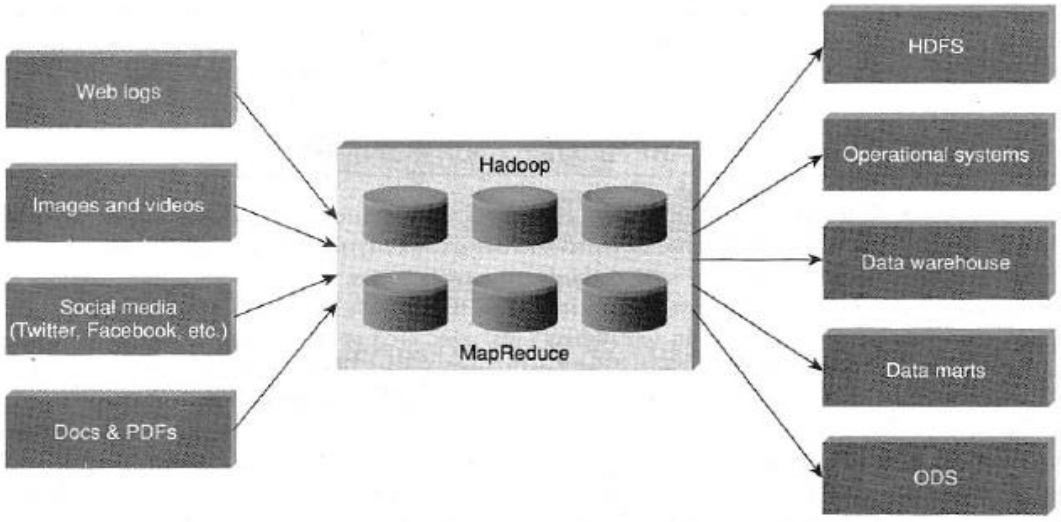
**4. Modeling:**

**Tools Used:** Advanced analytics tools including statistical and predictive modeling software.

**Purpose:** To build models that predict future trends or outcomes based on historical data, which can inform strategic planning and operational adjustments.

## A TYPICAL HADOOP ENVIRONMENT

- **Hadoop** is an open source framework from Apache and is used to store process and analyze data which are very huge in volume.
- Hadoop is written in Java and is not OLAP (online analytical processing). It is used for batch/offline processing. It is being used by Facebook, Yahoo, Google, Twitter, LinkedIn and many more. Moreover it can be scaled up just by adding nodes in the cluster.
- **A data warehouse** is a centralized repository for structured data that allows for business intelligence and analytics. On the other hand, **Hadoop** is an open-source framework that can process massive amounts of structured and unstructured data across distributed computing clusters.




- In a typical Hadoop environment, various components like web logs, images, videos, social media data, and documents (docs & PDFs) are processed, stored, and analyzed using Hadoop technologies.

**1. Web Logs**

- Web Logs are generated by web servers that track user activity, page requests, errors, and other interaction data.
- Web logs are ingested into Hadoop for large-scale log analysis, used for monitoring, troubleshooting, and understanding user behavior.
- Hadoop's distributed computing allows for the storage of large volumes of logs, and tools like MapReduce can process and analyze this data.

**2. Images and Videos**

- These are media files, such as pictures, videos, or any multimedia content.
- Hadoop can be used to store and index large-scale images and video files for use cases like recommendation systems, facial recognition, and video analytics.
- Hadoop Distributed File System (HDFS) allows storing these large files efficiently across multiple nodes. Specialized tools (like Apache Mahout or deep learning libraries) are used to analyze image/video content.

**3. Social Media (Twitter, Facebook, etc.)**

- Data generated from social media platforms, including text posts, comments, likes, shares, hashtags, etc.
- Social media data is harvested for sentiment analysis, trend detection, and customer behavior understanding.
- This data, often unstructured and massive in volume, is ingested into HDFS. MapReduce or tools like Apache Spark can process and analyze this data.

**4. Docs & PDFs**

- Documents like Word files, PDFs, etc.
- Hadoop stores these documents and can be used for text mining, search, and content analysis applications.
- HDFS stores the documents while tools like Apache Tika can be used to extract content and metadata, which can then be analyzed using MapReduce or other frameworks.

**5. Hadoop**

- A framework for distributed storage and processing of large datasets using commodity hardware.

- Hadoop is the central platform that integrates the storage (HDFS) and processing (MapReduce) for all types of data.
- Acts as the core of the environment, coordinating how data is distributed and processed across the cluster.

**6. MapReduce**

- A programming model used for processing large datasets in parallel across a Hadoop cluster.
- MapReduce is employed for tasks like filtering, aggregation, and analysis of large datasets.
- Executes the logic to process large-scale datasets in a distributed manner, typically leveraging the distributed nature of HDFS.

**7. Hadoop Distributed File System (HDFS)**

- A distributed storage system that breaks large files into blocks and stores them across different nodes in the cluster.
- All types of data (structured, unstructured, and semi-structured) are stored in HDFS in a reliable and scalable way.
- Ensures redundancy and fault tolerance by replicating data blocks across nodes, allowing large datasets to be stored and accessed efficiently.

**8. Operational Systems**

- Traditional systems (like relational databases or transactional systems) that run day-to-day operations of a business.
- Data from these systems can be transferred to Hadoop for analysis. This data often serves as input to the data warehouse or ODS.
- Hadoop ingests data from operational systems to be further processed, analyzed, or stored.

**9. Data Warehouse**

- A system used for reporting and data analysis, typically storing historical data in an organized structure.
- Hadoop complements data warehouses by providing storage and processing for unstructured data, while structured data can still be loaded into traditional warehouses.
- Hadoop can preprocess large, unstructured datasets and send structured or aggregated results to a data warehouse.

**10. Data Marts**

- Subsets of a data warehouse, often focused on a specific business area or function (e.g., sales, marketing).
- After processing in Hadoop, refined data can be moved into data marts for more specialized and faster access by specific departments.
- Hadoop helps in the preparation of data that gets loaded into data marts by performing complex transformations and aggregations.

**11. Operational Data Store (ODS)**

- A database that integrates data from multiple sources to provide a near-real-time, consolidated view for operational reporting.
- ODS is used for short-term data storage, often providing the raw data that Hadoop ingests for further processing.
- Acts as a storage layer that feeds into the ODS, enabling near-real-time processing of operational data.

## Coexistence of Big Data and Data Warehouse

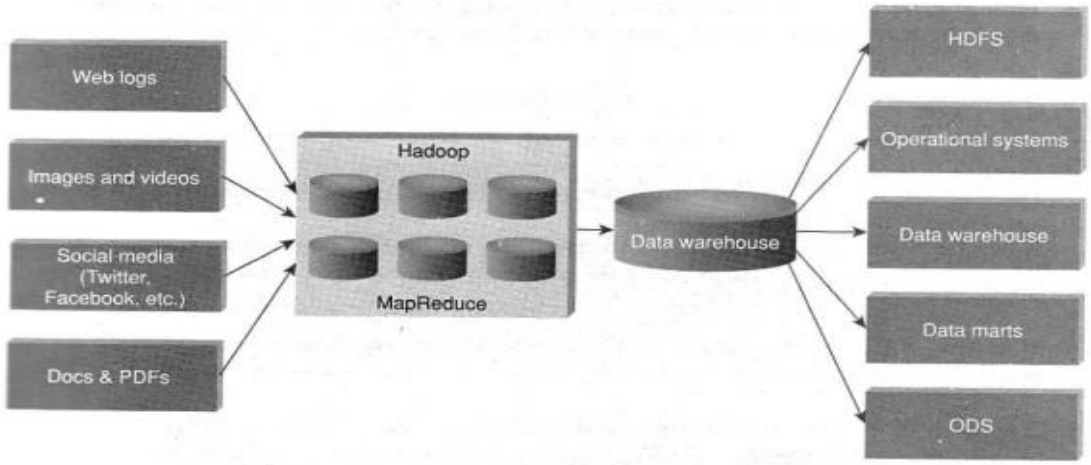



**Figure 2.11** Big data and data warehouse coexistence.

The above diagram illustrates the integration of big data with traditional data warehouse systems, commonly referred to as the coexistence of big data and data warehouses. Here's a breakdown:

**Data Sources (Left Side)**

- **Web logs**, **images and videos**, **social media** (e.g., Twitter, Facebook), and **documents** (e.g., PDFs) represent unstructured or semi-structured data sources.
- These types of data are typical in big data environments, coming in large volumes, varying formats, and are often difficult to store and process in traditional systems.

**Hadoop and MapReduce**

- In the middle section, **Hadoop** with **MapReduce** is used to process these large and diverse data sets.
    - **Hadoop** is a framework designed to store and process big data across clusters of computers.
    - **MapReduce** is a programming model used for distributed processing of large data sets by dividing tasks into smaller sub-tasks.

**Data Warehouse**

- The output from Hadoop/MapReduce is fed into a traditional **data warehouse**.
    - Data warehouses are designed to store structured data and are typically used for analytics and reporting purposes.

**Data Destinations (Right Side)**

- From the data warehouse, data flows to various systems:
    - **HDFS (Hadoop Distributed File System)**: a storage system used by Hadoop.
    - **Operational systems**: systems that support day-to-day operations in a business.
    - **Data warehouses**: a traditional storage system for structured data, used for reporting.
    - **Data marts**: smaller, more focused subsets of data warehouses, usually tailored for specific business needs.
    - **ODS (Operational Data Store)**: a type of database used for operational reporting and data integration from multiple sources.

The diagram represents how big data from various sources is processed using Hadoop/MapReduce and then integrated into traditional data warehouse environments for further use in various operational systems.

# Questions on Unit-1

1. Define Big Data. Explain 8 characteristics of Big Data.
2. Explain with examples the different types of Digital/Big Data.
3. Differentiate between Structured, Semi-structured and Unstructured data with examples.
4. Explain the Challenges with Big data.
5. How is traditional BI environment different from big data environment? Explain.
6. What is Data warehouse? Explain the key components in typical data warehouse environment.
7. What is Hadoop? Explain the key components in typical Hadoop environment.
8. What is Business Intelligence? How it is different from big data.
9. Discuss the importance of big data.
10. Explain the stages involved in transforming raw data into impactful actions.
11. Differentiate between Data Warehouse and Hadoop.
12. Differentiate between Traditional BI and Big Data.
13. Illustrate with a neat diagram the coexistence of big data and data warehouses.

| Data Warehouse Vs Hadoop | | |
|---|---|---|
| **S.No.** | **Hadoop** | **Data Warehouse** |
| 1. | An open-source software framework for the distributed storage and processing of huge datasets. | A central database of structured, ordered data. |
| 2. | It uses Distributed file system (HDFS) for data storage. | It uses a Relational database or structured storage system for data storage. |
| 3. | MapReduce programming model and ecosystem are used for data processing. | SQL-based queries are used for data processing. |
| 4. | Designed to scale horizontally. | Designed to scale vertically. |
| 5. | It can handle variety of data like structured, unstructured and semi structured data. | It can mainly handle structured data. |
| 6. | It offers high scalability and is capable of handling petabytes of data. | The scalability offered by a data ware house is limited depending on hardware resources. |
| 7. | The speed of processing data is very slow. | The data processing speed is faster in the data warehouse. |
| 8. | It is ideal for complex data transformations. | It has limited capability to handle complex data transformations. |
| 9. | It is affordable and has quite a lower cost. | It is highly expensive. |
| 10. | It provides direct access to raw data. | It provides aggregated data for analysis purposes. |
| 11. | It uses the "Schema-on-Read" Data schema. | It uses the "Schema-on-Write" Data schema. |
| 12. | It is mainly used for big data analysis and processing. | It is mainly used for reporting and business intelligence. |