What is data?

- Dictionary Definition:

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 (audio tape), optical (CD), or mechanical recording media (Phonographic disc)

What is big data?

- It is a collection of data that is huge in volume and 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.

Definition

- Big data is high-volume, and or / high velocity information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.
  - - Gartner IT Glossary
- The huge data is of the order of tera (10^12)bytes, Peta (10^15) bytes or Zeta (10^21) bytes.

Explanation of Big Data definition

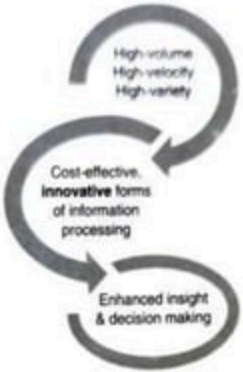


Figure 2.3 Definition of big data – Gartner.

Part I of the definition "big data is high-volume, high-velocity, and high-variety information assets" talks about voluminous data (humongous data) that may have great variety (a good mix of structured, semi-structured, and unstructured data) and will require a good speed/pace for storage, preparation, processing, and analysis.

Part II of the definition "cost effective, innovative forms of information processing" talks about embracing new techniques and technologies to capture (ingest), store, process, persist, integrate, and visualize the high-volume, high-velocity, and high-variety data.

Part III of the definition "enhanced insight and decision making" talks about deriving deeper, richer, and meaningful insights and then using these insights to make faster and better decisions to gain business value and thus a competitive edge.

Data → Information → Actionable intelligence → Better decisions → Enhanced business value

Big Data Definition

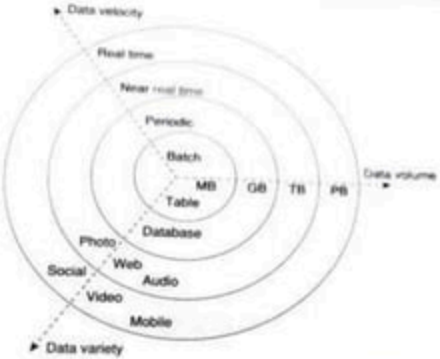


Figure 2.5 Data: Big in volume, variety, and velocity.

Table 2.2 Growth of data

Bits	0 or 1
Bytes	8 bits
Kilobytes	1024 bytes
Megabytes	1024^2 bytes
Gigabytes	1024^3 bytes
Terabytes	1024^4 bytes
Petabytes	1024^5 bytes
Exabytes	1024^6 bytes
Zettabytes	1024^7 bytes
Yottabytes	1024^8 bytes

Why is Big Data important?

- Using the data from any source and analyzing it, we can find answers that
  - Streamline resource management
  - Improve operational efficiencies
  - Optimize product development
  - Drive new revenue and growth opportunities
  - Enable smart decision making

Big data enables to accomplish business related tasks

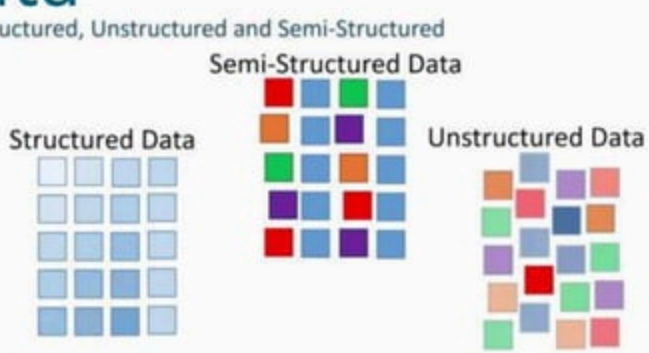
- Determine the root causes of failures, issues and defects in near- real time (industrial usage)
- Spotting anomalies faster and more accurately than human eye (healthcare usage)
- Recalculating entire risk portfolios in minutes (investment / finance sector)
- Detect fraudulent behavior before it affects your organization.

Some examples of big data

- The NYSE generates about one terabyte of new trade data per day.
- 55 billion messages and 4.5 billion photos are exchanged on wattsapp every day
- 300 hours of video is uploaded every minute
- Every minute user sends 31.25 million messages and watch 2.77 million videos
- There are around 40,000 search queries googled each second.

Types of Big Data

- Structured
- Semi-Structured
- Unstructured



Unstructured data

The university has 5600 students.
John's ID is number 1, he is 18 years old and already holds a B.Sc. degree.
David's ID is number 2, he is 31 years old and holds a Ph.D. degree. Robert's ID is number 3, he is 51 years old and also holds the same degree as David, a Ph.D. degree.

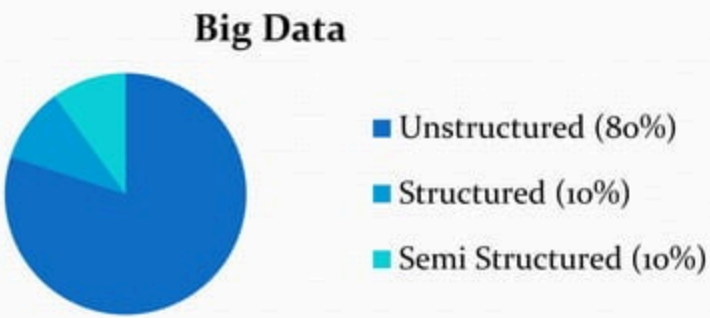
Semi-structured data

```
<University>
  <Student ID="1">
    <Name>John</Name>
    <Age>18</Age>
    <Degree>B.Sc.</Degree>
  </Student>
  <Student ID="2">
    <Name>David</Name>
    <Age>31</Age>
    <Degree>Ph.D. </Degree>
  </Student>
  ....
</University>
```

Structured data

ID	Name	Age	Degree
1	John	18	B.Sc.
2	David	31	Ph.D.
3	Robert	51	Ph.D.
4	Rick	26	M.Sc.
5	Michael	19	B.Sc.

Types of Big Data



	Structured Data	Semi-Structured Data	Unstructured Data
Characteristics	Defined data model	No defined data model or inherent structure	Partially organized
Storage	Relational databases and data warehouses	Data warehouses and data lakes	Relational databases
	Stored in rows/columns	Numerous formats	Tagged-text formats
Examples	Transactional information, Names , Dates, Addresses	XML, HTML, JSON, Emails, Web pages	PDFs, Images, Text files, Videos, Audio files

Un-structured Data

Figure 1.8 Sources of unstructured data.

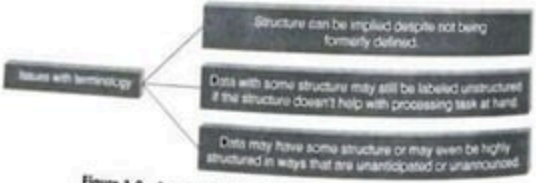
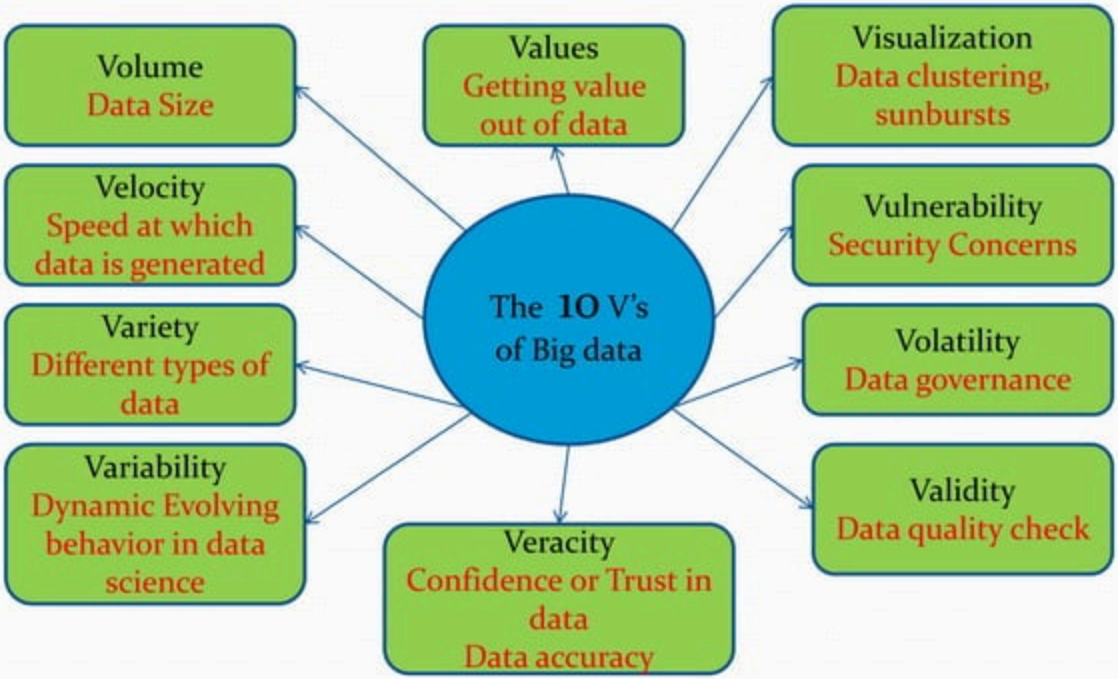
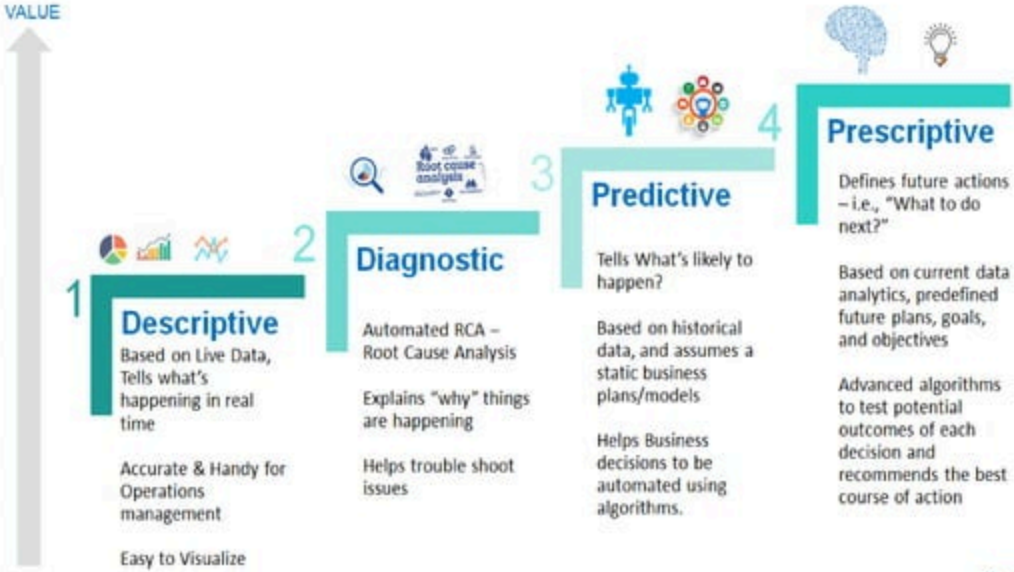


Figure 1.9 Issues with terminology of unstructured data.

10 Characteristics of Big Data

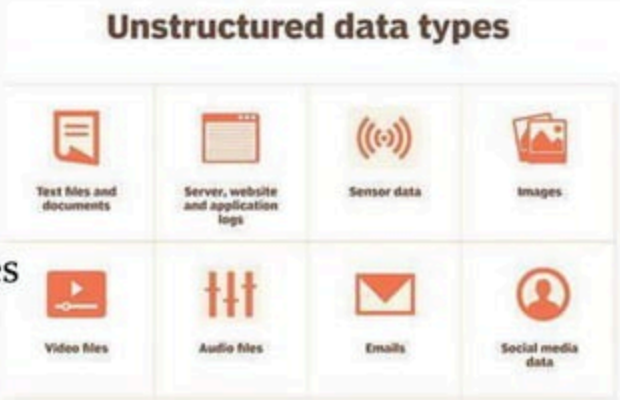


4 Types of Data Analytics



Unstructured data for Analytics

- Business Documents
- Emails
- Social Media
- Customer feedback
- Webpages
- Open-ended survey responses
- Images, Audio and Video

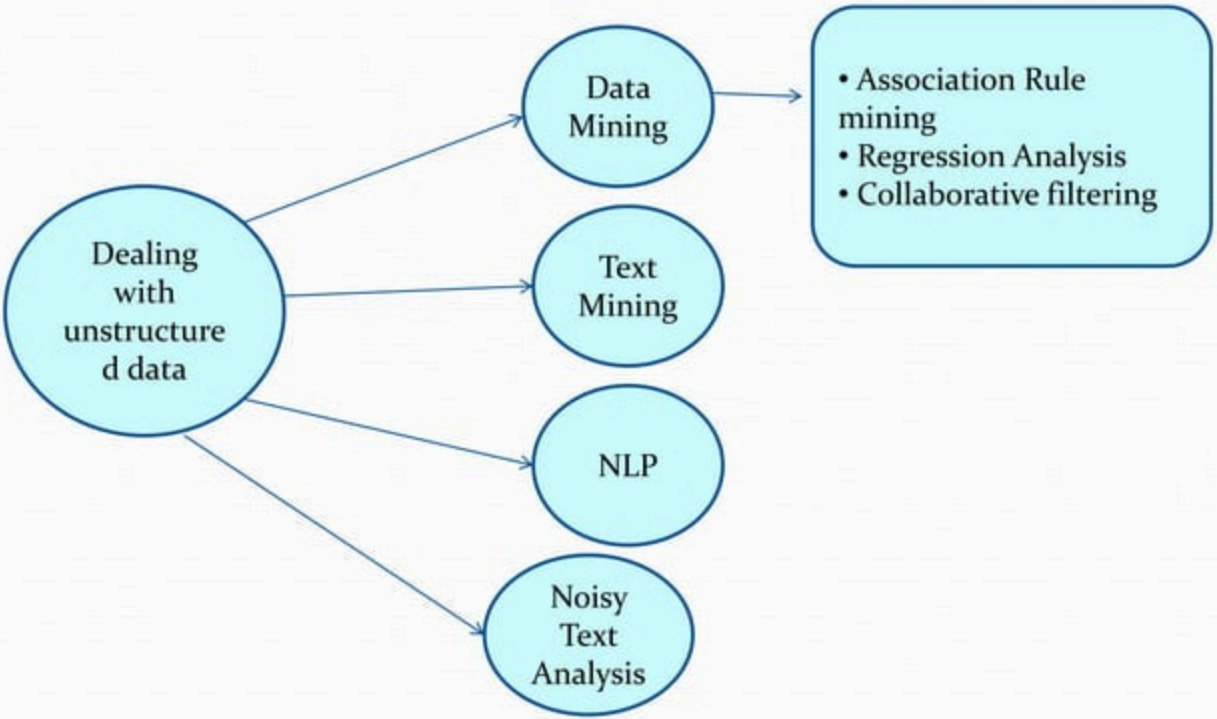


- *Importance of unstructured data analysis for businesses*:
  - Improve the customer experience
  - Discover gaps in the market and innovate
  - Listen to your customers

How to Analyze unstructured Data

- **Choose the End Goal**
  - Define a clear set of measurable goals.
- **Collect Relevant Data**
  - Focus on the source of data
- **Clean Data**
  - Reduce noise
  - Eliminate unwanted information
- **Implement Technology**
  - NoSQL databases
  - Data visualization using Tableau, Google data studio

Unstructured Data Analytics



Unstructured Data Analytics Tools

- MonkeyLearn – Used for Text Analytics.
  - This tool makes it simple to clean, label and visualize customer feedback
- Word Clouds - **textual data visualization** which allows anyone to see in a single glance the words which have the highest frequency within a given body of text
- Listen to customer's voice – open surveys and emails.
  - Aspect is based on sentiment analysis
- Amazon AWS
- Microsoft Azure
- IBM Cloud

The Advantages of Deploying Big Data

- Better Decision making
- Cost Reduction
- Newer Products and redevelopment of the old
- Risk Analysis
- Collection of Data

Industries using Big Data

- CA technology have done a global study in which clearly the benefits of Big data outweigh the obstacles in implementation
- The percentage of organizations that plan to and already have implemented a big data project is 84%
- Acquision has increased to 54%, revenue has improved by 88%.

- hiQ: It specializes in 'people analytics'.
- SumAI: Helps businesses optimize their social media campaigns with the help of one single chart.
- Splunk: Visual analytics
- Alteryx: Combines structured and unstructured data from a number of sources and stores it in one database. Spacial, predictive and statistical analysis tasks are done on this data.

How big data is used in different industries

- Media and entertainment:
  - Companies like Hulu and Netflix work with big data to analyze user tendencies, preferred content, trends in consumption.
  - Lot of services like spotify are coming up with new revenue models to increase profits
  - Ads are targeted more strategically thanks to big data analytics software.

Finance

- Shift from Manual trading to trading backed by technology

- These models analyze big data to make
  - accurate enter / exit trade decisions,
  - minimize risk using machine learning and
  - guage market sentiment using opinion mining

Healthcare

- With predictive analytics , big data can predict negative health events that senior citizens would experience from home care.
- This reduced visits by 73% and 64% amongst chronically ill patients
- Big data can identify disease trends based on demographics, geographies, socio economics and other factors

Education

- Improve learning management. Tracking how much time learners spend on tasks, tests, and exams helps to customize curricula efficiently.
- Improve students' performance. Leveraging data about learners' performance helps educators develop personalized learning paths.
- Provide data-driven decision-making.
- Predict learning outcomes.
- Use big data to reduce dropout rates

Retail

- Enhance Service Quality
- Optimize Price
- Manage Supply Chain
- Identify Potential Risks
- Forecast demand

Manufacturing

- Quality Assurance
- Supply Chain Optimization
- Improving Throughput and Yield
- Less downtime
- Greater Customer Service

Big Data Challenges

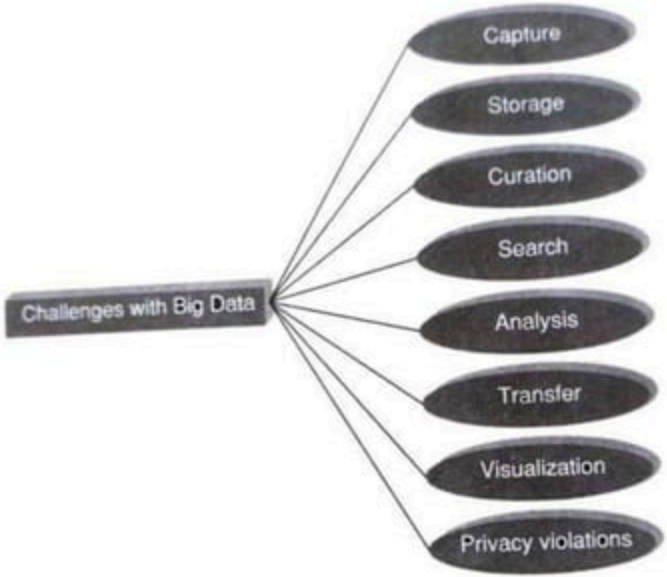


Figure 2.4 Challenges with big data.

Big Data Challenges

- 1. Lack of Knowledge Professionals
  - To run these large data tools, companies need skilled data professionals. (data scientists, data analysts and data engineers)
    - **Solution** : Big data tools are used by professionals who are not data science experts but have the basic knowledge.
    - This saves a lot of money for the companies.
- 2. Lack of proper understanding of Massive Data
  - Employees not knowing how to store sensitive data.
    - **Solution**: Data workshops and seminars must be held at companies for everybody.

Big Data Challenges

- 3. Data Growth Issues:
  - One of the biggest challenges is to store the huge data.
    - Solution: Compression is used to reduce the size of data stored.
    - De-duplication removes the duplicate and unwanted data
    - Data Tiering stores data in different data tiers.(public clouds, private cloud and flash storage)
- 4. Confusion while Big data tool selection
  - Companies are confused on which tool to select for Data analysis and storage? Hbase, Cassandra etc.
    - Solution: Hire experts who know which tools to use.

Big Data Challenges

- 5. Integrating data from a spread of sources
  - Data in corporation comes from various sources like social media pages, ERP applications, customer logs etc.
    - Solution: Data integration problems are solved by purchasing proper tools.
- 6. Securing Data
  - Companies can lost a lot of revenue due to a stolen record or a knowledge breach.
    - Solution: cyber-security professionals guard their data. Other steps include encryption, identity and access control, implementation of end point security real-time security monitoring

Assignment Questions

1. What is Big Data ?
2. Explain the types of data. Also briefly mention the sources of each types of data along with examples.
3. Why is big data important. How does it help businesses and briefly describe its usage across various domains(industry, retail, healthcare, manufacturing, education ...)
4. Briefly describe the characteristics of big data.
5. Describe the types of analytics and mention the sources of unstructured data used in analytics.
6. Mention some tools used in analytics
7. Discuss the Big Data challenges briefly