Overview of Big Data

Unit I

Introduction to Big Data, Characteristics of Data, and Big Data Evolution of Big Data, Definition of Big Data, Challenges with big data, Why Big data? Data Warehouse environment, Traditional Business Intelligence versus Big Data. State of Practice in Analytics, Key roles for New Big Data Ecosystems, Examples of Big Data Analytics.

Big Data Analytics, Introduction to big data analytics, Classification of Analytics, Challenges of Big Data, Importance of Big Data, Big Data Technologies, Data Science, Responsibilities, Soft state eventual consistency. Data Analytics Life Cycle.

Topics

- Big Data Overview
  - What is Big Data?
  - Data -> Big Data
  - Definition of Big Data
  - Data Structure / Types of Big Data
  - Growth of Data - Tools and Techniques
- Characteristics of Data
- Evolution of Big Data
- Characteristics of Big Data
- Challenges With Big Data
- Advantages of Big Data
- Disadvantages of Big Data

How big is the Canvas of Big Data?

1000 (10^3) kB 1 kilobyte

1000 (10^6) MB 1 megabyte

1000 (10^9) GB 1 gigabyte

1000 (10^{12}) TB 1 terabyte

1000 (10^{15}) PB 1 petabyte

1000 (10^{18}) EB 1 exabyte

1000 (10^{21}) ZB 1 zettabyte

1000 (10^{24}) YB 1 yottabyte

We shall be soon reaching to a range which is known as
पद्म (*padma*)

Data Storage Units Chart: From Smallest to Largest

Unit	Shortened	Capacity
Bit	b	1 or 0 (on or off)
Byte	B	8 bits
Kilobyte	KB	1024 bytes
Megabyte	MB	1024 kilobytes
Gigabyte	GB	1024 megabytes
Terabyte	TB	1024 gigabytes
Petabyte	PB	1024 terabytes
Exabyte	EB	1024 petabytes
Zettabyte	ZB	1024 exabytes
Yottabyte	YB	1024 zettabytes

Types of Digital Data

Structured

Semi-Structured

Quesi Structured

Unstructured

Characteristics of Data

- Structure of the Data
- Data Sources?
- Granularity?
- Types?
- Nature: Static or Real Time Streaming?

Composition

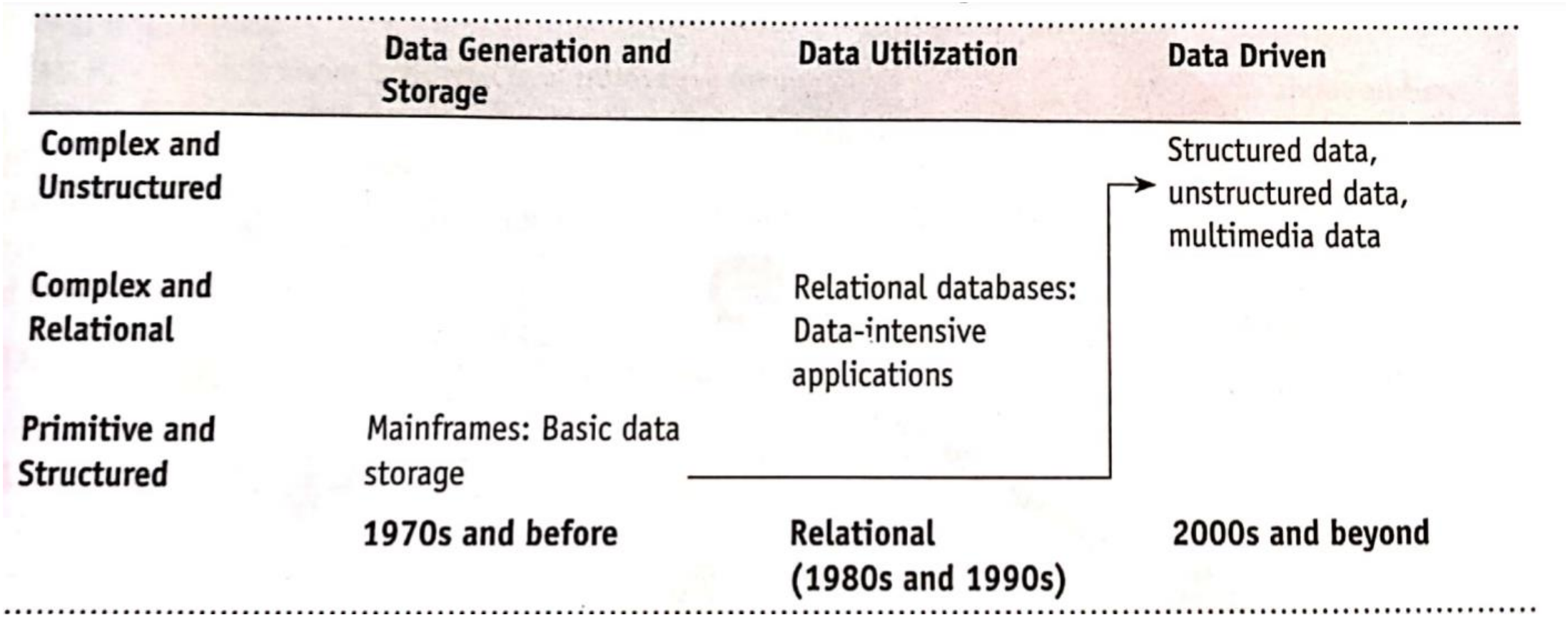
- State of the Data
- Does data requires cleansing
- Is the data useful for analysis?

Condition

- Where and why the Data has been generated?
- Consistency and usability of Data

Context

Evolution of Big Data



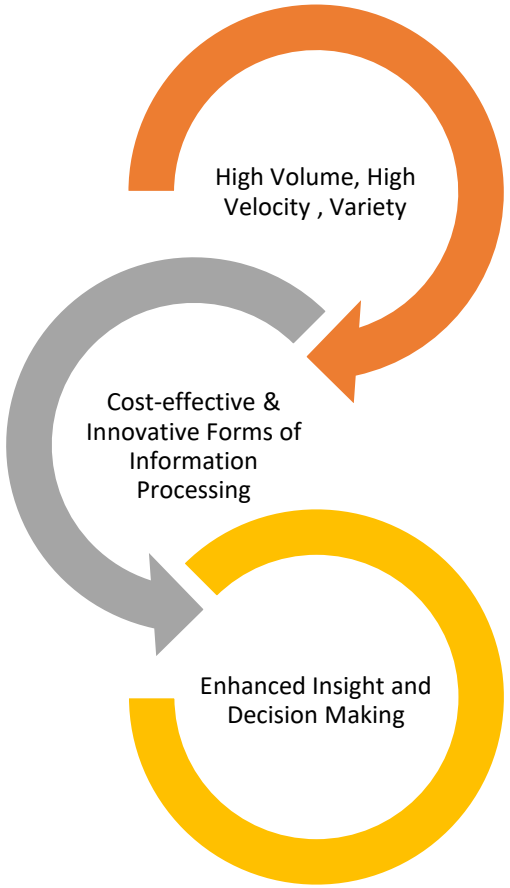
Big Data Definition

- No single standard definition…

"***Big Data***" is data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to enable insights that unlock new sources of business value.

Big data is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.

Definition of Big Data



Data

->

Information

->

Actionable Intelligence

->

Better Decisions

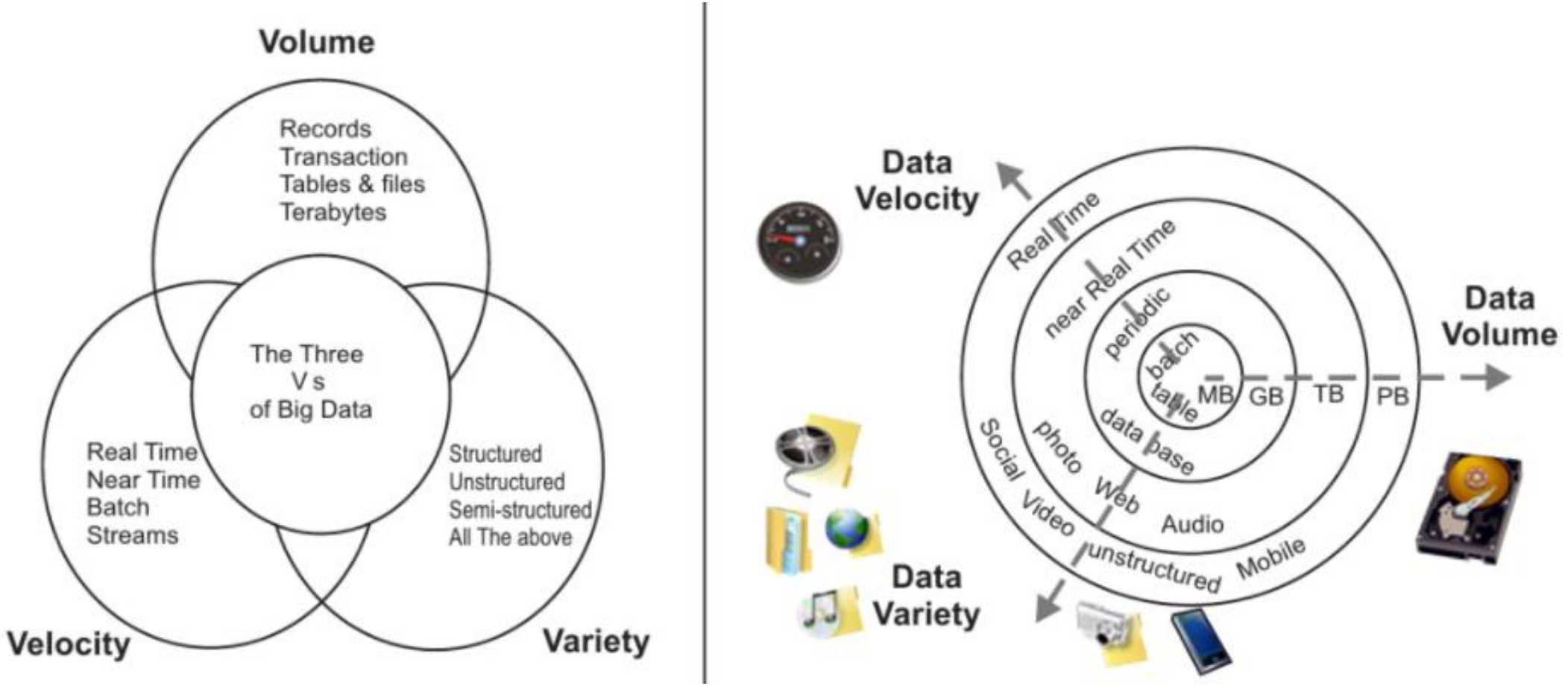
->

Enhanced Business Value

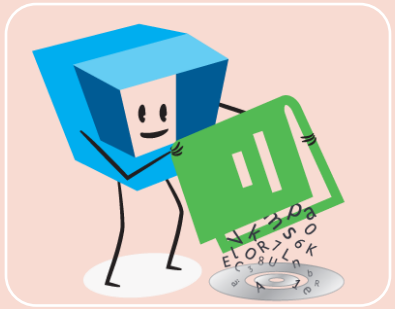
Challenges with big data

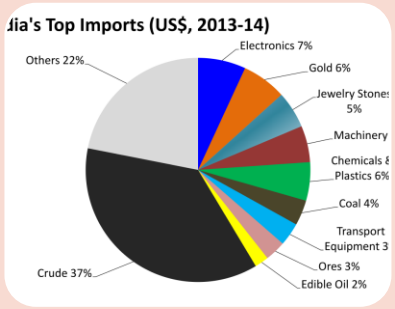
- Lack of proper understanding of bigdata
- Data Growth Issues
- Confusing variety of big data tools/ technologies
- Lack of data professionals
- Securing data
- Integrating data from variety of sources

3 V's of Big Data



Characteristics of Big Data (5 V's)





Volume

- Data at Rest
- Terabytes to Exabytes of existing data to process

Velocity

- Data in Motion
- Streaming Data , Milliseconds to seconds to respond

Varity

- Data in Many Forms
- Structures, Unstructured, Text, Multimedia

Veracity

- Data in Doubt
- Uncertainty due to data inconsistencss, ambiguities, latency, deception, model approximations

Value

- Extracting meaningful data
- Transform a tsunami of data into Business

Volume

- ***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.

  Volume refers to the unimaginable amounts of information generated every second from social media, cell phones, cars, credit cards, M2M sensors, images, video, and whatnot.

  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.

Velocity

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.

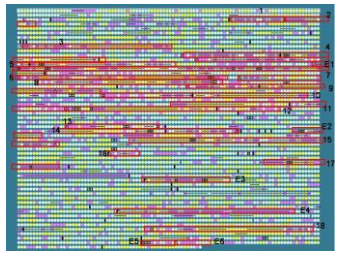
Big Data Velocity deals with the speed at which data flows in from sources like business processes, application logs, networks, and social media sites, sensors, Mobile devices, etc. The flow of data is massive and continuous.

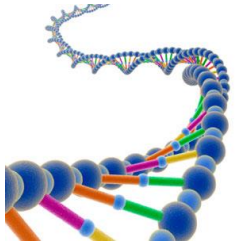
Hence Velocity plays a major role compared to the others, there is no point in investing so much to end up waiting for the data.

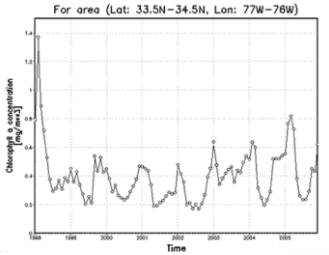
So, the major aspect of Big Data is to provide data on demand and at a faster pace.

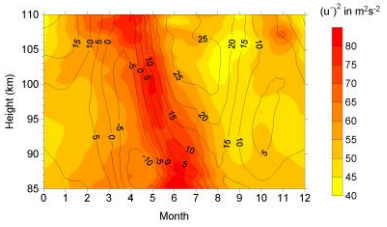
Verity

- Data comes in all types of formats – from structured datasets, numeric data in traditional databases to unstructured text documents, email, video, audio, stock ticker data and financial transactions.
- 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.
- Now days, data in the form of emails, photos, videos, monitoring devices, PDFs, audio, etc. is also being considered in the analysis applications.
- This variety of unstructured data poses certain issues for storage, mining and analysing data.









To extract knowledge➔ all these types of data need to linked together

Veracity

- Data in Doubt

Veracity basically means the degree of reliability that the data has to offer. Since a major part of the data is unstructured and irrelevant, Big Data needs to find an alternate way to filter them or to translate them out as the data is crucial in business developments

Uncertainty due to data inconsistences, ambiguities, latency, deception, model approximations

Value

- Extracting meaningful data / Transform a tsunami of data into Business

Value is the major issue that we need to concentrate on. It is not just the amount of data that we store or process. It is actually the amount of valuable, reliable data that needs to be stored, processed, analyzed to find insights.

New Dimension to Data

- Real – Time Data
- Shared data
- Linked data
- High-fidelity data

Advantages of Big Data

- Better Decision making
- Increased Productivity
- Reduce Costs
- Improved customer service
- Fraud Detection
- Increased Revenue
- Increased Agility
- Greater Innovation
- Faster speed to market

Disadvantages of Big Data

- Need for Talent
- Data Quality
- Need for Cultural change
- Compliance
- Cybersecurity risks
- Rapid change
- Hardware Needs
- Costs
- Difficulty integrity legacy systems

Thank You….

Revise the topics from Syllabus References…

Fill Your Attendance Form….!