SNS COLLEGE OF TECHNOLOGY

Coimbatore-35
An Autonomous Institution

Department of Information Technology

19ITE305 – BIG DATA ANALYTICS

III B.Tech. IT/ VI SEMESTER

UNIT I : INTRODUCTION TO BIG DATA AND ANALYTICS

Topic 2 : Introduction to Big Data: Characteristics – Evolution – Definition

Classification of Digital Data, Structured and Unstructured Data - Introduction to Big Data: Characteristics – Evolution – Definition - Challenges with Big Data - Other Characteristics of Data - Why Big Data - Traditional Business Intelligence versus Big Data - Data Warehouse and Hadoop Environment

Big Data Analytics: Classification of Analytics – Challenges - Big Data Analytics important - Data Science - Data Scientist - Terminologies used in Big Data Environments .

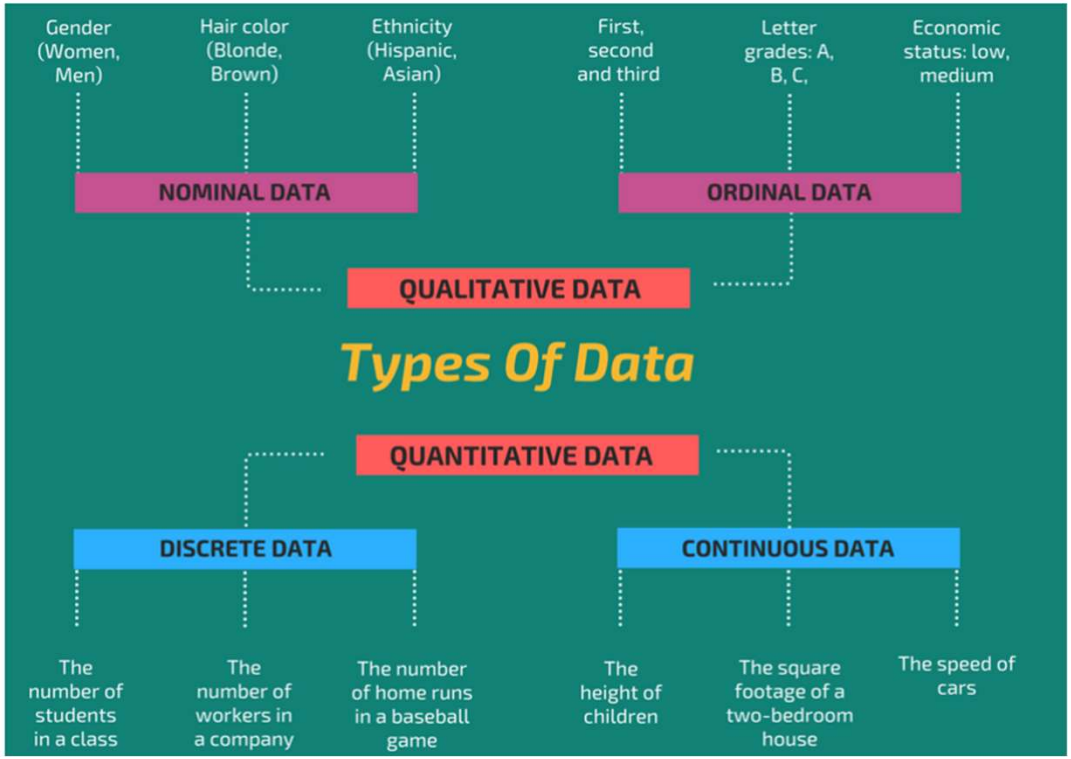
What is data and information?

Data is raw, unorganized, unprocessed information. E.g., the information collected for writing a research paper is data until it is presented in an organized manner.

Data generates information and from information we can draw valuable insight.

Information is the processed, organized data that is beneficial in providing useful knowledge. For eg., the data compiled in an organized way in a research paper provides information about a particular concept/ topic.

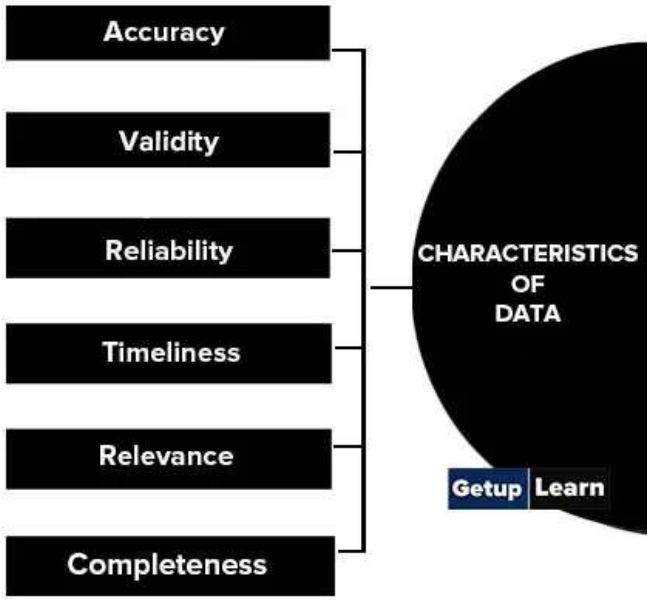
Types of data



Characteristics of Data

The following are six key characteristics of data

- Accuracy
- Validity
- Reliability
- Timeliness
- Relevance
- Completeness



Big Data, what is it?

traditional computer science

data that will not fit in main memory.

For example…

- busy web server access logs
- graph of the entire Web
- all of Wikipedia
- daily satellite imagery over a year

Big Data, what is it?

traditional computer science

data that will not fit in main memory.

data with a *large* number of observations and/or features.

statistics

Big Data, what is it?

Tall data: edge list of a large graph RGB values per pixel location in large images

data with a *large* number of observations and/or features.

statistics

Wide data: mobile app usage statistics of 100 people

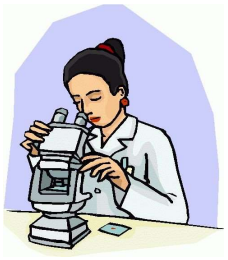
Big Data, what is it?

traditional computer science

data that will not fit in main memory.

data with a *large* number of observations and/or features.

statistics



other fields

non-traditional sample size (i.e. > 300 subjects); can't analyze in stats tools (Excel).

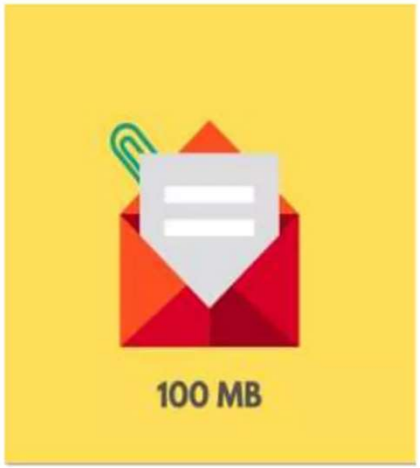
What is big data?

Big Data refers to a huge volume of data, that cannot be stored and processed using the traditional computing approach within a given time frame.

Example of big data?

For example, if we try to attach a document that is of 100 megabytes in size to an email we would not be able to do so. As the email system would not support an attachment of this size.

Therefore this 100 megabytes of attachment with respect to email can be referred to as Big Data.



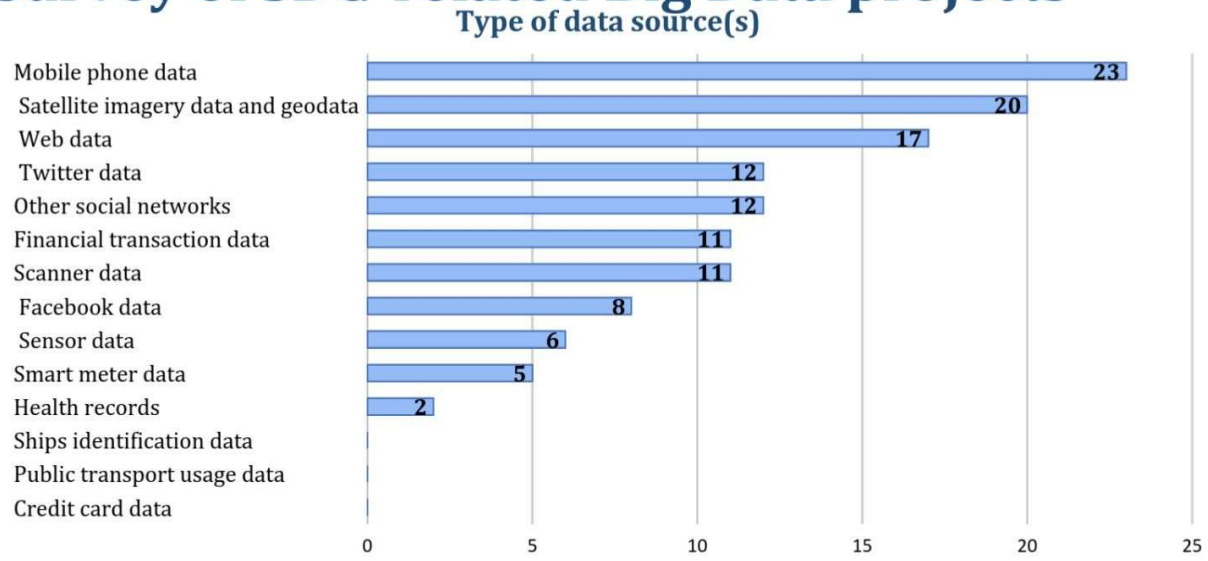
Big Data, what is it?

Government View

THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP

BigData
UN Global Working Group

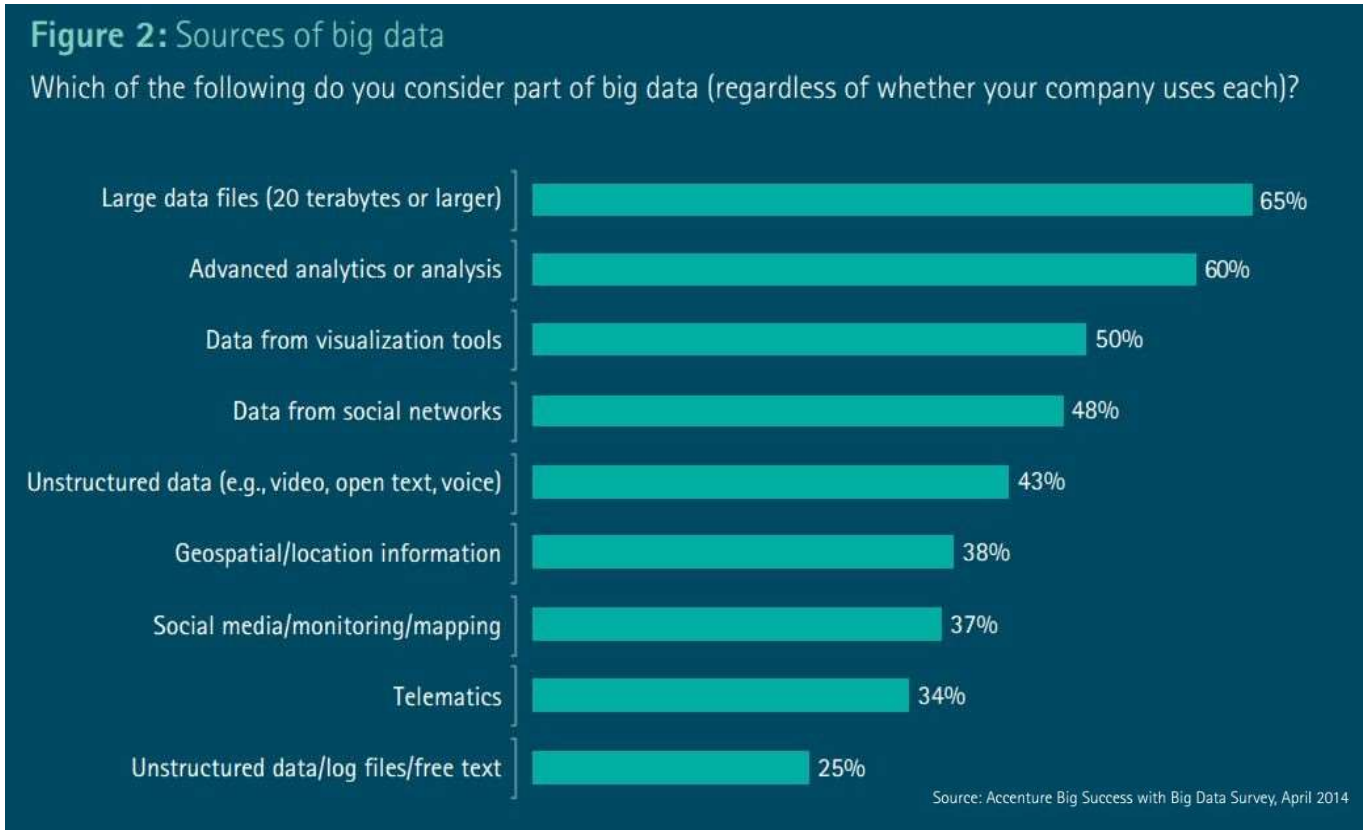
1. Survey of SDG-related Big Data projects



- Mobile (23), Satellite imagery (20) and social media (12+12+8) are the most prominent sources

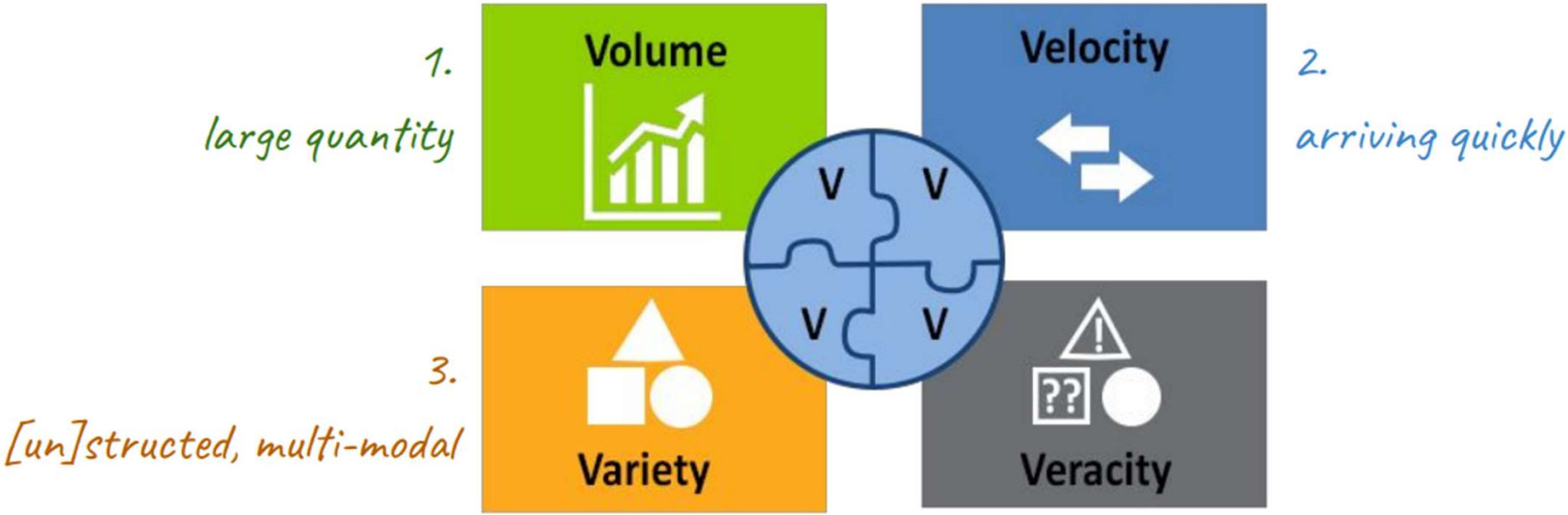
Big Data, what is it?

Industry View



Big Data, what is it?

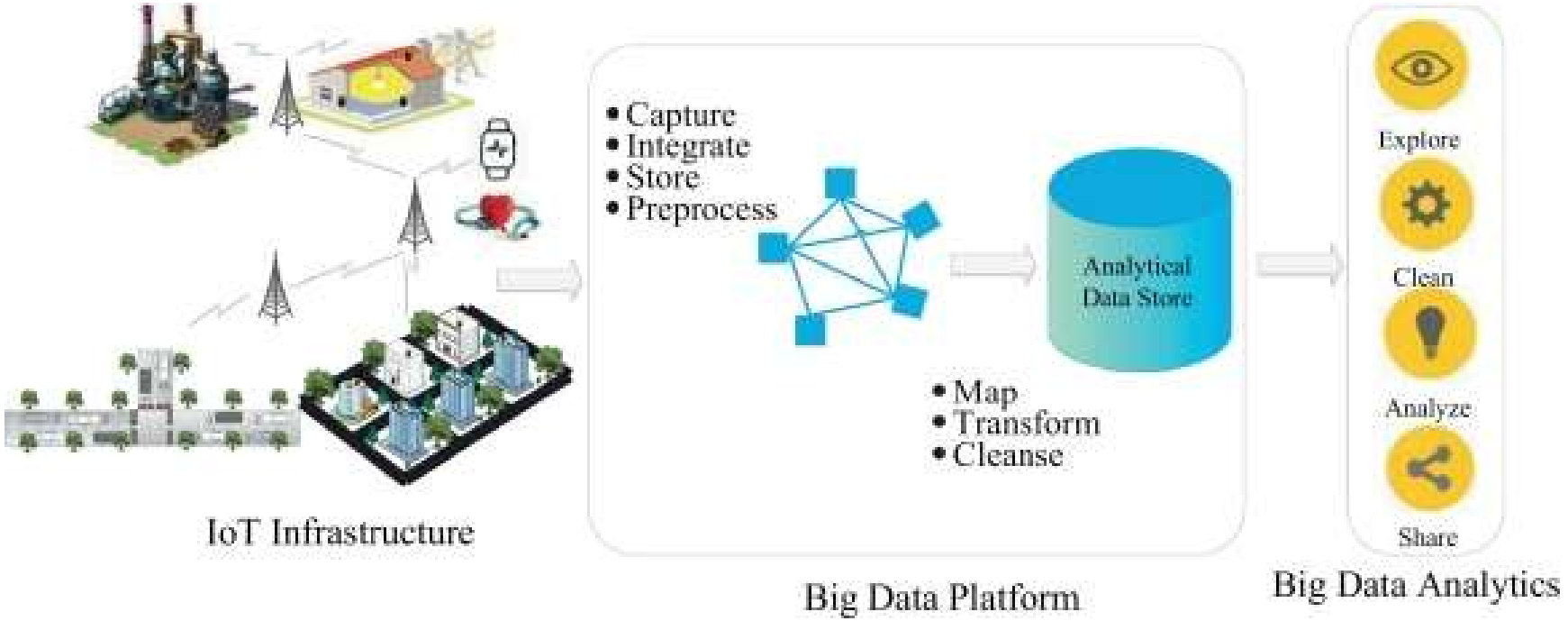
Analyses which can handle the 3 Vs and do it with quality (veracity)



What is big data Analytics?

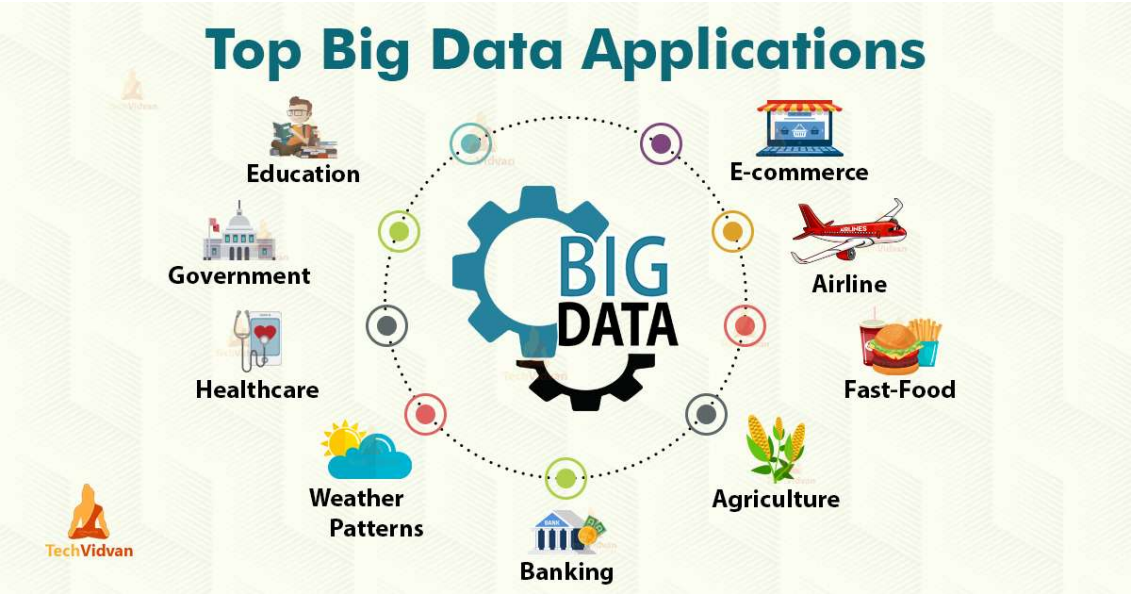
Big data Analytics is a process to extract meaningful insight from big such as hidden patterns, unknown correlations, market trends and customer preferences

Introduction to Bigdata



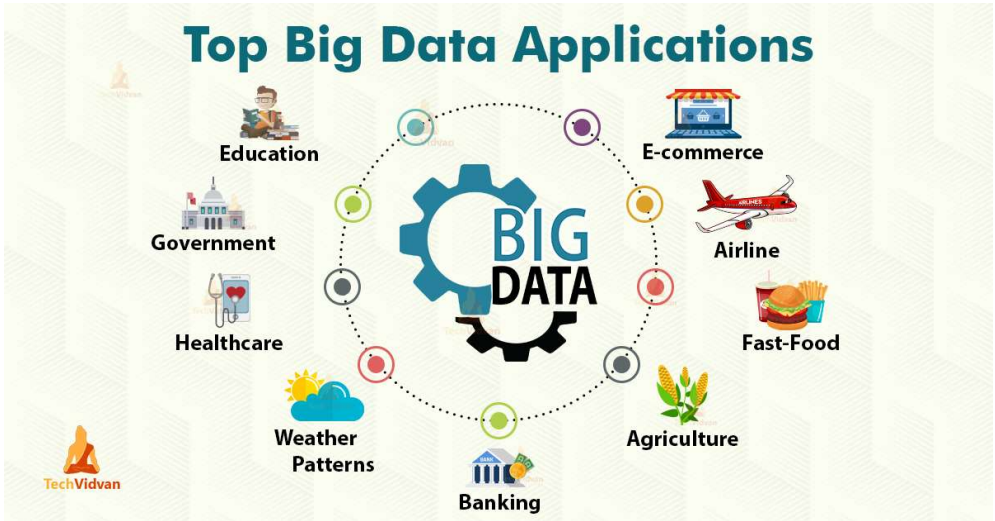
Example for Big Data

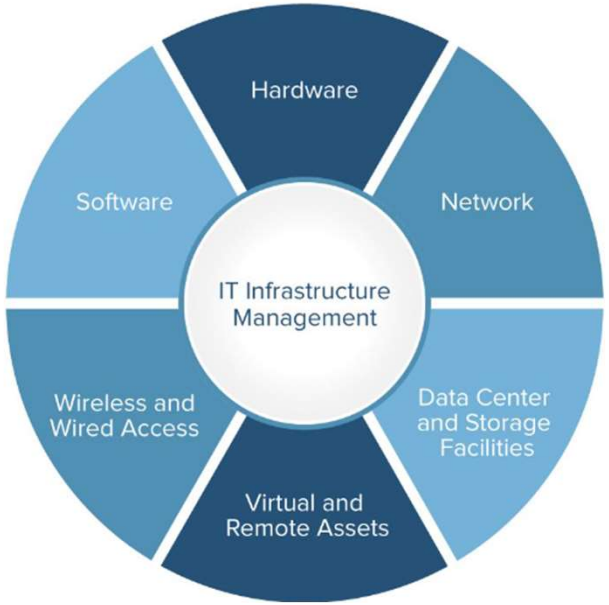
<u>Retail,</u> IT Infrastructure, and Social Media

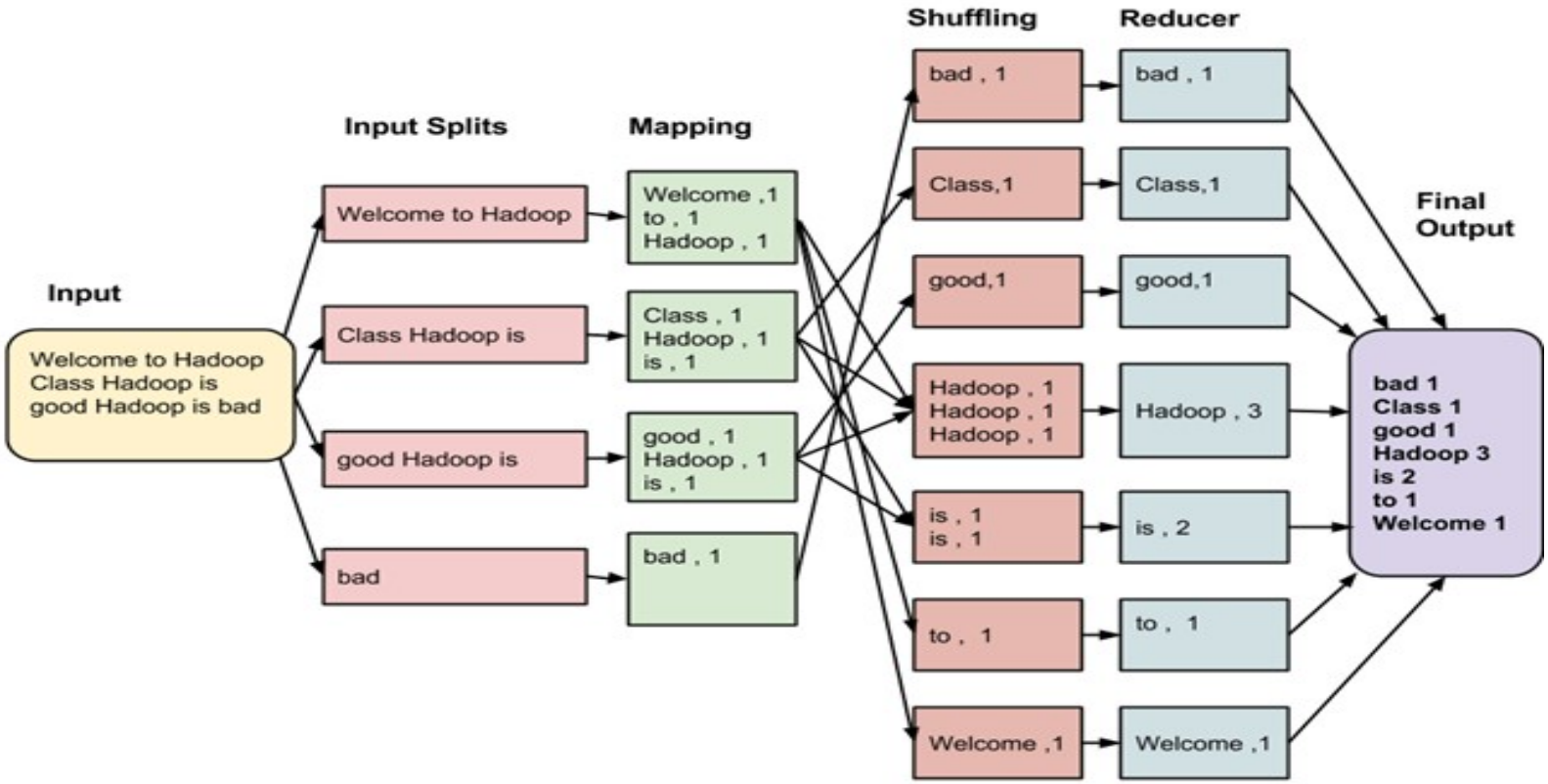


Example for Big Data

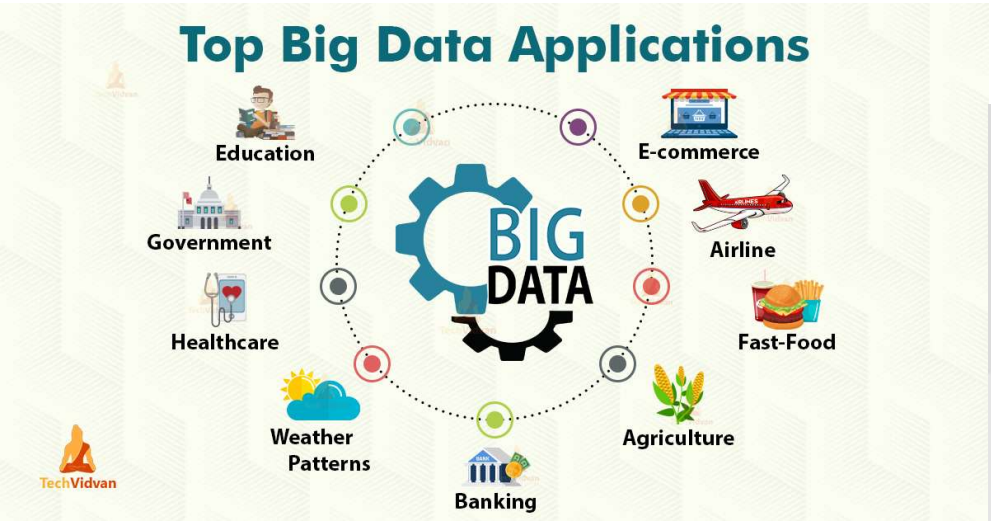
Retail, IT Infrastructure, and Social Media





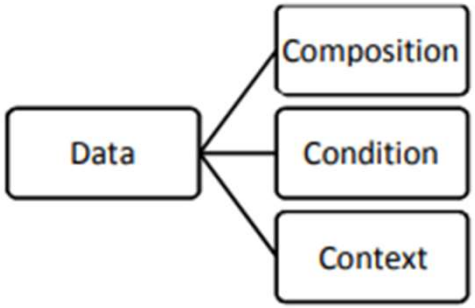

Input
Welcome to Hadoop
Class Hadoop is
good Hadoop is bad
Input Splits
Welcome to Hadoop
Class Hadoop is
good Hadoop is
bad
Mapping
Welcome ,1
to , 1
Hadoop , 1
Class , 1
Hadoop , 1
is , 1
good , 1
Hadoop , 1
is , 1
bad , 1
Shuffling
bad , 1
Class,1
good,1
Hadoop , 1
Hadoop , 1
Hadoop , 1
is , 1
is , 1
to , 1
Welcome ,1
Reducer
bad , 1
Class,1
good,1
Hadoop , 3
is , 2
to , 1
Welcome ,1
Final Output
bad 1
Class 1
good 1
Hadoop 3
is 2
to 1
Welcome 1

Example for Big Data

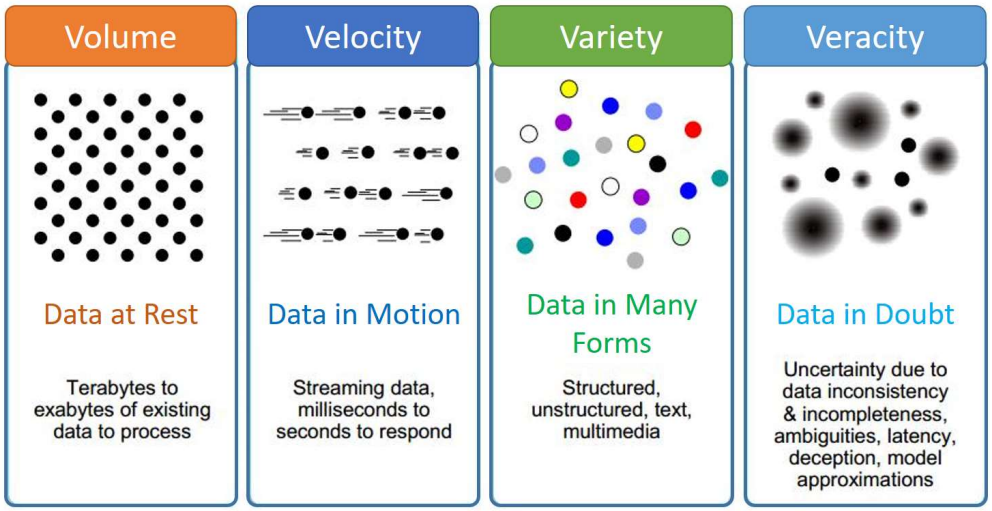


Retail, IT Infrastructure, and Social Media

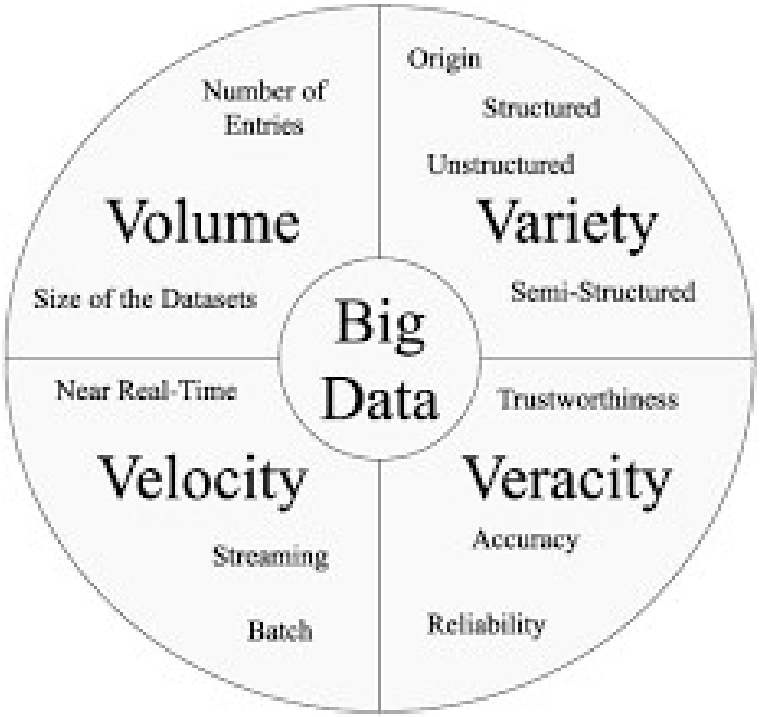
Characteristics of data



Composition	Condition	Context
Structure of data	State of data	
the sources of data, the granularity, the types, and the nature of data	Can one use this data as is for analysis?	Where has this data been generated?
Static or Real Time Streaming	Does it require cleansing for further enhancement and enrichment?	Why was this data generated? How sensitive is this data?"



Characteristics of Big Data



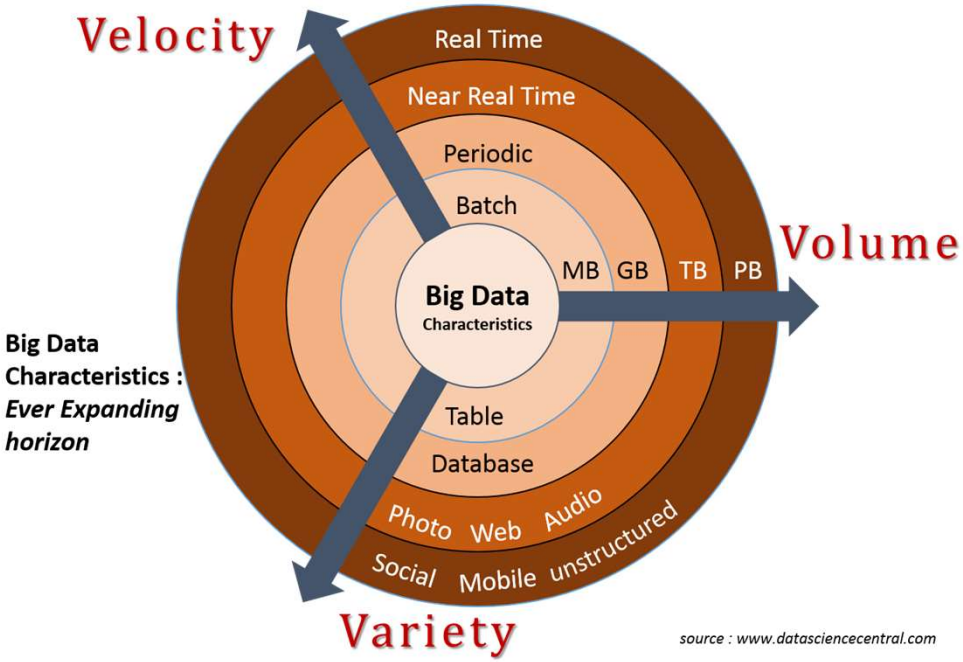
Evolution of Big Data

	DATA GENERATIONAND STORAGE	DATA UTILZATION	DATA DRIVEN
COMPLEX AND UNSTRUCTURED			Structured data, Unstructured data, Multimedia data
COMPLEX AND RELATIONAL		databases: Data-intensive applications	
PRIMITIVE AND STRUCTURED	Mainframes: Basic data storage		
	1970 and before	**Relational (1980 and 1990s)**	**2000 and beyond**

Definition of Big Data

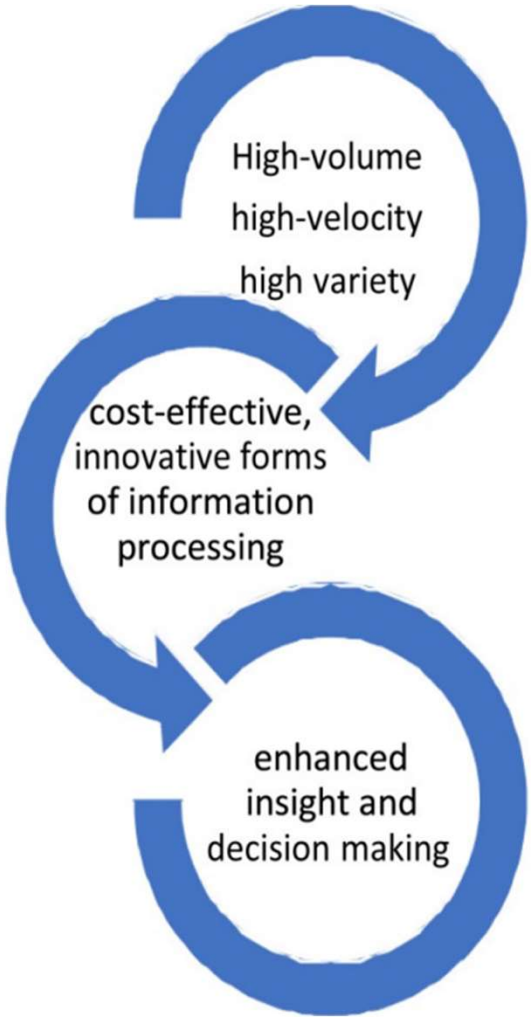
1. Big data is high-velocity and high-variety information assets that demand cost effective, innovative forms of information processing for enhanced insight and decision making.
2. Big data refers to datasets whose size is typically beyond the storage capacity of and also complex for traditional database software tools
3. Big data is anything beyond the human & technical infrastructure needed to support storage, processing and analysis.

- It is data that is big in volume, velocity and variety.



Big Data Characteristics : ***Ever Expanding horizon***

source : www.datasciencecentral.com



TEXT BOOKS

Seema Acharya, Subhashini Chellappan, “Big Data and Analytics”, Wiley Publications, First Edition,2015

REFERENCES

1. Judith Huruwitz, Alan Nugent, Fern Halper, Marcia Kaufman, “Big data for dummies”, John Wiley & Sons, Inc. (2013)
2. Tom White, “Hadoop The Definitive Guide”, O’Reilly Publications, Fourth Edition, 2015
3. Dirk Deroos, Paul C.Zikopoulos, Roman B.Melnky, Bruce Brown, Rafael Coss, “Hadoop For Dummies”, Wiley Publications, 2014
4. Robert D.Schneider, “Hadoop For Dummies”, John Wiley & Sons, Inc. (2012)
5. Paul Zikopoulos, “Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data, McGraw Hill, 2012