What is Data?

Data is the foundation of Analytics. Before starting any analysis, you need to understand the characteristics of data, its source of origination, and the transformation it has gone through.

Agenda

- Types of Data
- Sources of Data
- Data Quality and Changes

What is Data?

Data is a **set of values** of **qualitative** or **quantitative** variables.

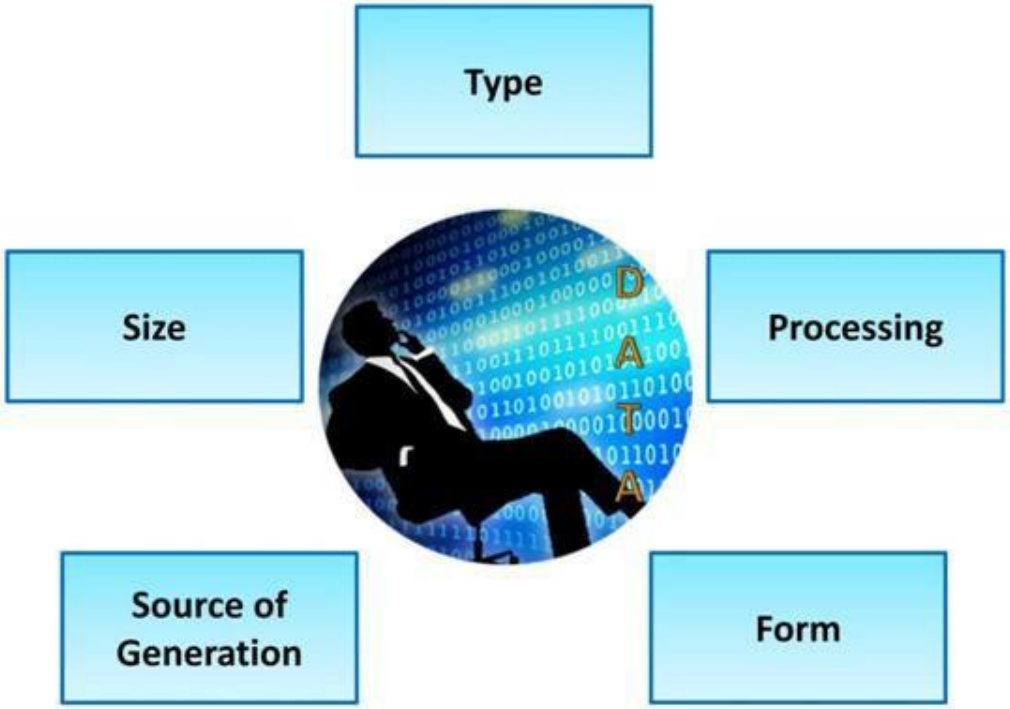
- Data is descriptive in nature; it describes an attribute that can be observed, but not measured
- Examples:
  - Flavors of ice cream = {"Vanilla", "Butterscotch", "Chocolate" }
  - Hair color = { "Blonde", "Brunette", "Black" }
  - Profession type = { "Engineer", "Tailor", "Consultant" }

Qualitative

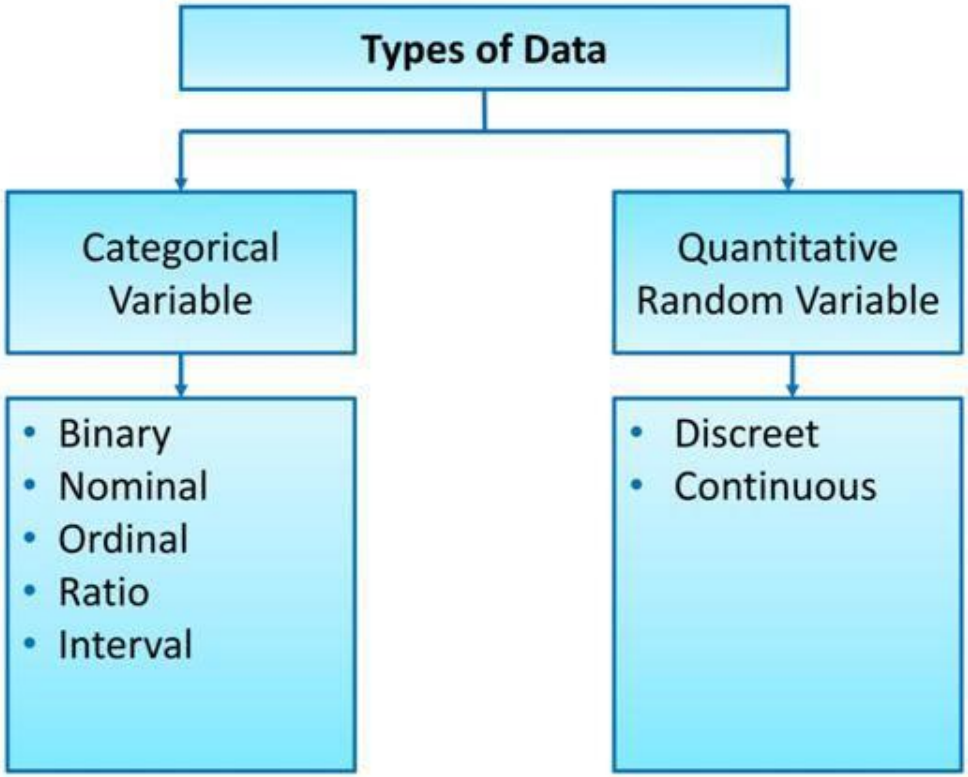
- Data is a numeric measure; it captures the measure of an attribute
- Examples:
  - Heights of students = {5'6", 5'9", 5'3", 5'5" }
  - Cost = {120.5, 130.2, 111.6, 90.8}
  - Age = {34,26,67, 53}

Quantitative

Basis of Data Categorization



Types of Data



Binary Data

Binary data has only two possible states:

- 0 or 1
- Toss of a coin
- Switch On or Off
- Dot and dash of telegraph

Nominal Data

- Categorical data where the data is coded in a manner that it represents a label
- You can only count but cannot order or measure nominal data

Examples: Names of cars, book titles in a library, and marital status

Ordinal Data

- Data is ordered
- It has a natural hierarchy
- The intervals between the ranks may not be necessarily equal (distance between groups can be different)

Examples: Customer satisfaction score and medal tally

1	RUS	13	11	9	33
2	NOR	11	5	10	26
3	CAN	10	10	5	25
4	USA	9	7	12	28
5	GER	8	6	5	19
6	NED	8	7	9	24
7	SUI	6	3	2	11
8	BLR	5	0	1	6
9	AUT	4	8	5	17
10	FRA	4	4	7	15

Discreet and Continuous Data

- Numerical data
- Finite number of possible values
- Examples:
  - Number of people in a room
  - Number of items in a basket
  - Numbers of hours in a day

Discrete Data

- Numerical data
- Infinite number of possible values
- Usually is in decimals
- Examples:
  - Height
  - Weight
  - Sales
  - Account balance

Continuous Data

Raw Data

RAW Data Definition:

- Data from the source
- Input to the data processing process
- Raw data may:
  - Have errors
  - Not validated
  - Multiple forms
  - Unformatted
  - Dubious, requiring confirmation or citation

RAW Data Example:

- If the correct format is not specified in an application form, the date of birth data can take many forms, such as "31st January 1990", "31/01/1990", "31/1/90", and "31 Jan 90". This raw data needs to be processed to a common format for further use by systems/humans

Processed Data

Processed Data Definition:

- Data after processing for issues in the raw data
- Analysis ready data
- Processing includes scrubbing, cleansing, merging, formatting, transforming, and so on
- All data processing steps documented

Processed Data Example:

- Recoding: "Number of children" field in a survey form may be left blank by people who don't have children. This has to be coded as "0", which is a valid value for this variable
- Deriving: End of day sale amount for a store can be calculated by summing up all the transactions in a day

Data Collection Types

Census

- Systematic collection of data about all members of population

Observational study

- Collection of data to draw inference of outcome of a treatment on subjects. It is not in control of the investigator to assign the subjects either to the test or the control groups

Convenience sample

- Collection of data from a sample where the subjects are selected because of their convenient accessibility and proximity to the researcher

Randomized trial

- Collection of data to draw inference of outcome of a treatment on subjects. The investigator randomly allocates the subjects to either the test or the control group

Forms of Data

Structured Data:

- Data can be organized in well-defined structures
- Structures include arrays, vectors, or tables
- Relationships in data defined within the structure

Student Data								
ID	First Name	Last Name	Date of joining	Batch Number	Course Name	Address1	Address2	City
1001	Santosh	Kumar	10/6/2014	2	Analytics	100/A, 2nd Cross	Indiranagar	Bangalore

Forms of Data (Cont'd)

- Data is organized in an arbitrary manner with no pre-defined structure
- The types of content includes free text, documents, images, and videos
- Example: Resume document of a student with free text and images

Unstructured Data

- Semi-structured data does not conform to a formal defined structure, but entities belong to classes with attributes
- Data cannot be processed as effectively as the structured data
- Example: Information stored as XML

Semi-structured Data

Forms of Data (Cont'd)

- Data is collected in a batch mode at periodic intervals
- There is a delay in the availability of data as a certain periodicity is maintained for its collection

Batch Data

- Data is collected in real time; the data is delivered as it gets generated
- There is no delay in timeliness of data provided

Real-time Data

Sources of Data

User generated	• Blogs • Documents
System/Application generated	• Web logs • Network event logs
Device generated	• Surveillance cameras capturing traffic patterns • Point Of Sale (POS) system
Internal	• Generated internally in an organization across business processes; Sales data, logistics data, finance data, HR data, and so on
External	• Generated by external bodies or data aggregators or credit bureaus

Two Views of the Same Individual

- Woman
- Single
- Age 25-30
- Personal income $80K+

Demographic Overview

- Loves to travel
- Has subscription for a fashion magazine
- Annual membership at the local health club
- Spends $400/- month on personal care products
- Active on social network everyday for 1-2 hours

Psychographic Overview

Data Quality Issues

Missing data	• Input process not capturing all the data or not mandatory in the input process
Junk values	• Lack of validations
Definitions	• Inaccurate or incomplete definition • Multiple definitions for the same measures across organization
Completeness	• Incomplete data, left blank
Validity	• Invalid data (does not follow the expected structure)
Accuracy	• Inaccurate data due to problems with either the measurement system or the operator
Timeliness	• Delay in availability
Consistency	• Inconsistent definition across systems • Difference in scales of measurements

Data Quality Horror Story

PREGNANT MEN!

GREAT BRITAIN: 17,000 men were pregnant between 2009 and 2010. The hospital data reveals that these men went to hospitals for "pregnancy-related services" and had obstetric exams performed among other prenatal treatments.

Data Quality Horror Story (Cont'd)

Read the Fine Print

These men had gone to the doctor for procedures that had medical codes close to the medical code for obstetric services.

The employees working at the hospitals had **incorrectly coded** the procedures by carelessly entering the numbers.

"Big Data" Landscape

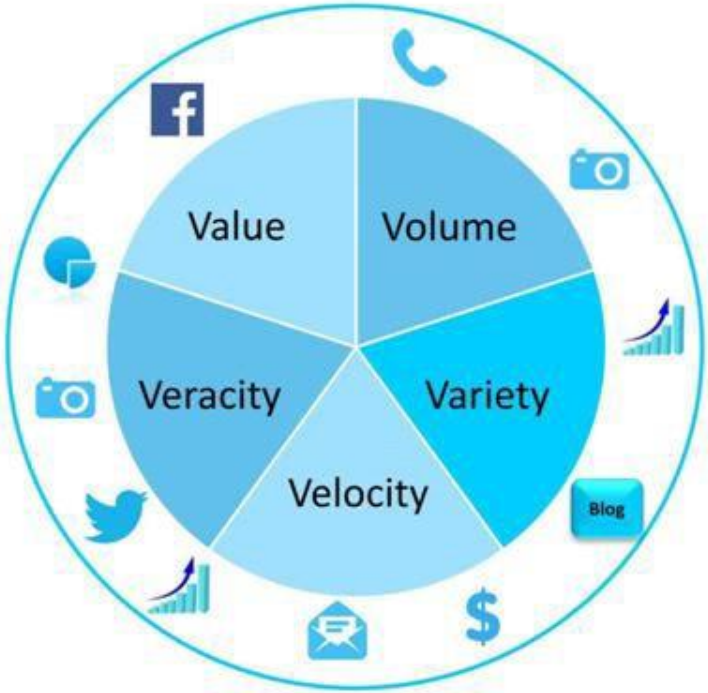
"There was 5 Exabytes of information created between the dawn of civilization through 2003, but that much information is now created every 2 days"

Eric Schmidt

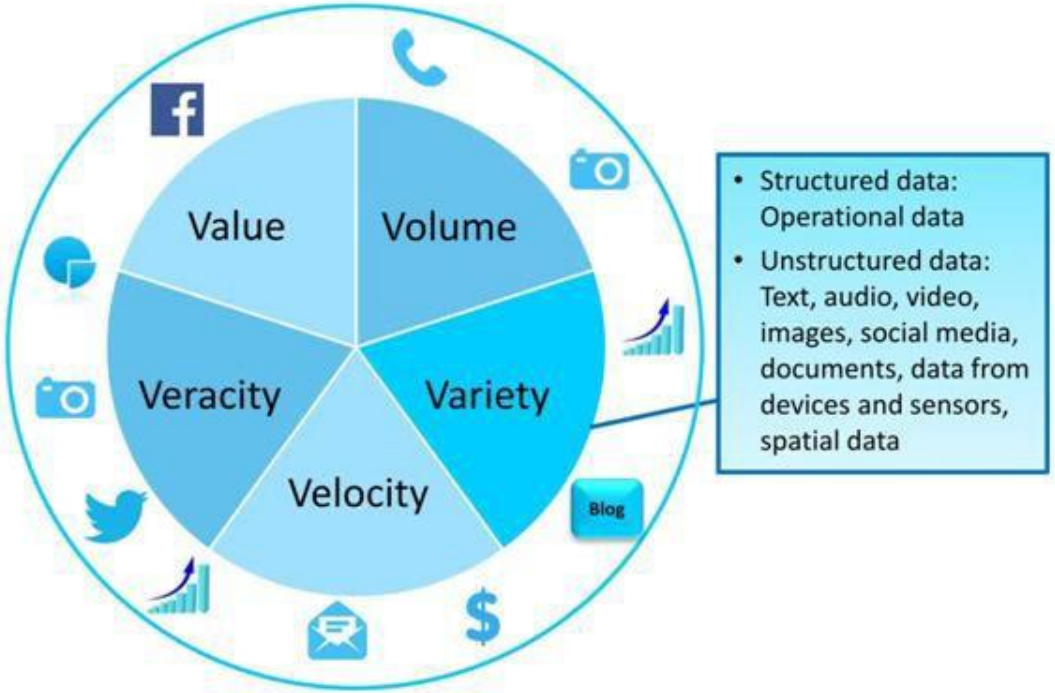
- Twitter processes 340 million messages weekly (Data generation in last one year is equivalent to data generated in last 15 years)
- Facebook users generate 2.7 billion comments and likes
- Amazon S3 storage adds more than one billion objects bi-weekly
- eBay stores 90 Petabytes of data about customer transactions
- Enterprise information measure is no more Tera and Peta bytes, but Exa and Zetta bytes

5 Vs of Big Data

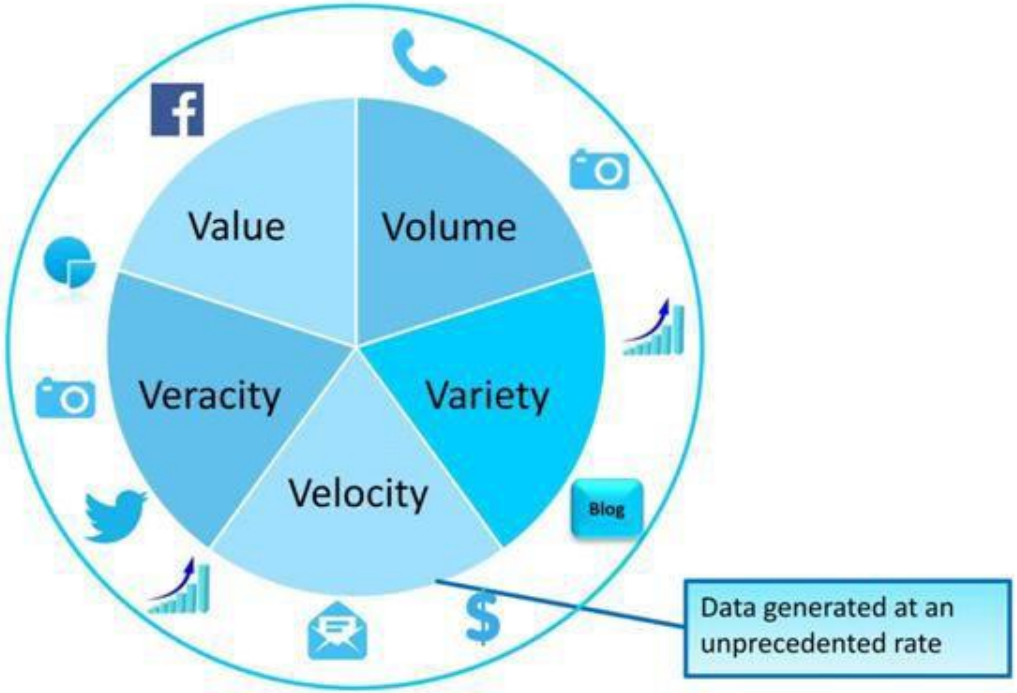
- Big Data is often described using the five Vs: Volume, Variety, Velocity, Veracity, and Value



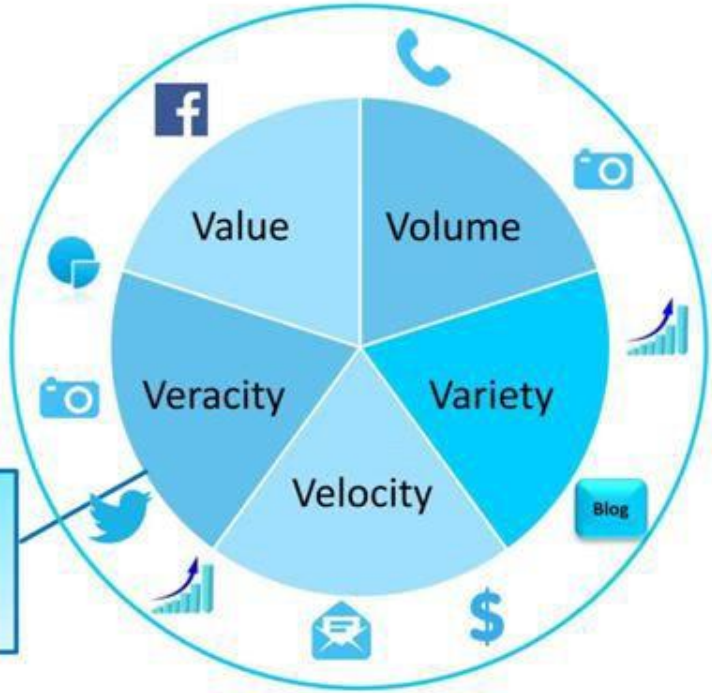
5 Vs of Big Data



5 Vs of Big Data



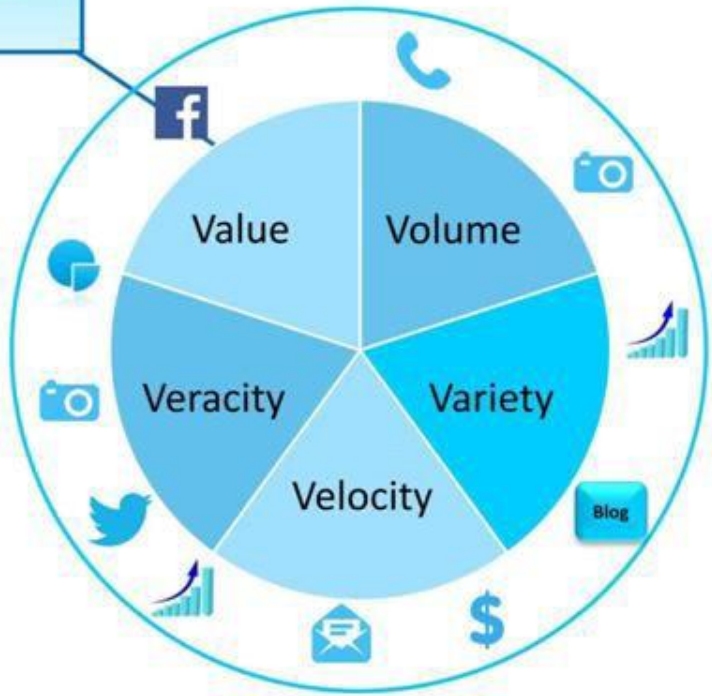
5 Vs of Big Data



- Huge variability in data being collected
- Quality and accuracy less controllable

5 Vs of Big Data

- Noise vs. Signal
- Quantifiable business value a challenge



Big Data: The Challenge

The challenge with the large volume of data is:

- Large scale data storage and retrieval
- Large scale processing for data analysis
- Storage and retrieval for the variety of data types, which are largely unstructured