BIG DATA TOOLS FOR MANAGERS

(N2MBA07)

Syllabus

Unit-1: Big Data, Database

- Overview of Big Data
- Data, Information, Database

Unit-2: SQL

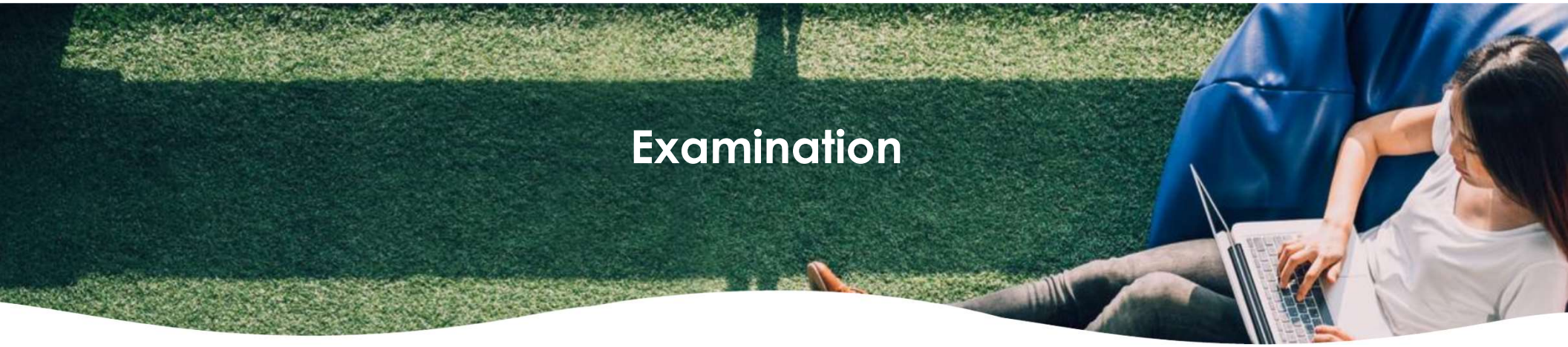
- Introduction to SQL, MySQL
- Data retrieval using MySQL

Unit-3: R Programming

- Introduction to R Language
- Data Manipulation, Graph, Regression

Unit-4, 5: Python

- Introduction to Python Programming Concept
- Data Manipulation, Time Series & Text Analytics using Python



Examination

Internal assessment

- 50 Marks Question Paper
- Practical Exam
- Write SQL/R/Python code

Semester end assessment

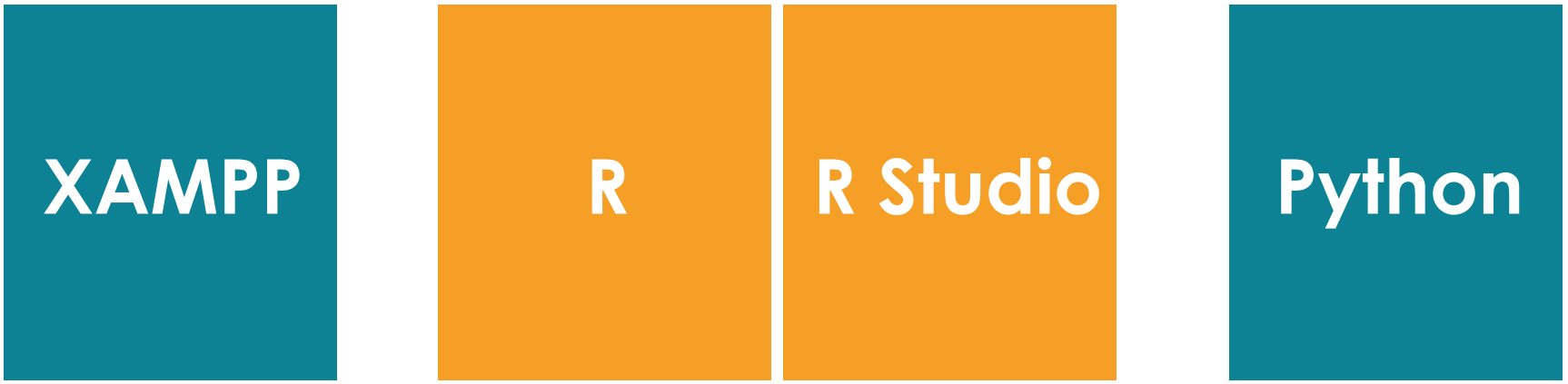
- 50 Marks MCQ paper
- 50 questions with multiple choice option
- 1.5 hrs

Class timings

Weekly once (on Saturday)

09:00AM – 10:10AM	Theory Sessions	~3 hrs	Theory Sessions	01:00PM – 02:10PM
10:10AM – 10:30AM	Break		Break	02:10PM – 02:30PM
10:30AM – 12:00PM	Practical Sessions		Practical Session	02:30PM – 04:00PM

Software Tools for BDTM



Unit 2 Unit 3 Unit 4 & 5

What is a Data ?

- ✓Data is a collection of information gathered by observations, measurements, research or analysis
- ✓It consists of facts, numbers, names, figures or even description of things.
- ✓Data can be organized in the form of free text, images, graphs, tables
- ✓Example :

City	Min. Temp. (in Degrees)	Max. Temp. (in Degrees)	Rain
Mumbai	25	40	22%
Delhi	32	45	16%
Bangalore	23	35	28%
Chennai	33	48	21

What is a Digital Data ?

✓Digital data is data that represents other forms of data using specific machine language systems that can be interpreted by various technologies.

✓The most fundamental of these systems is a binary system, which simply stores complex audio, video or text information in a series of binary characters, traditionally ones and zeros, or "on" and "off" values.

Data Sources

Internal data sources

Information that comes directly from the company's systems and are specific to the company in question

Example:

- Sales, Cash Flow, Production
- Customer Relationship Management(CRM)
- Enterprise Resource Planning(ERP) system
- OLTP and operation data

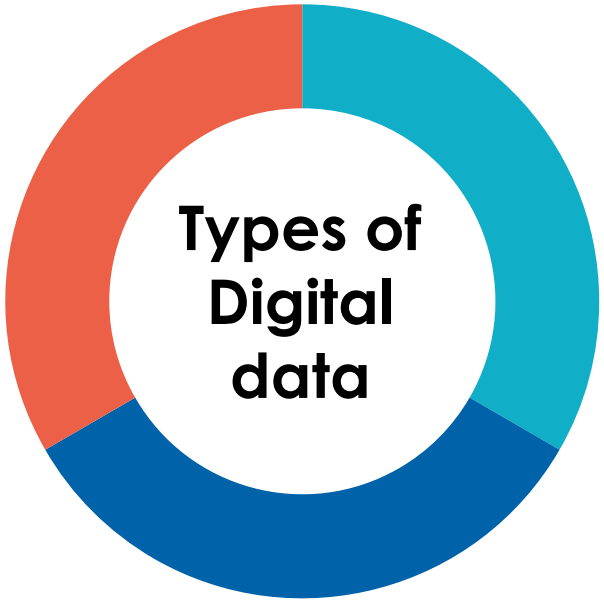
External data sources

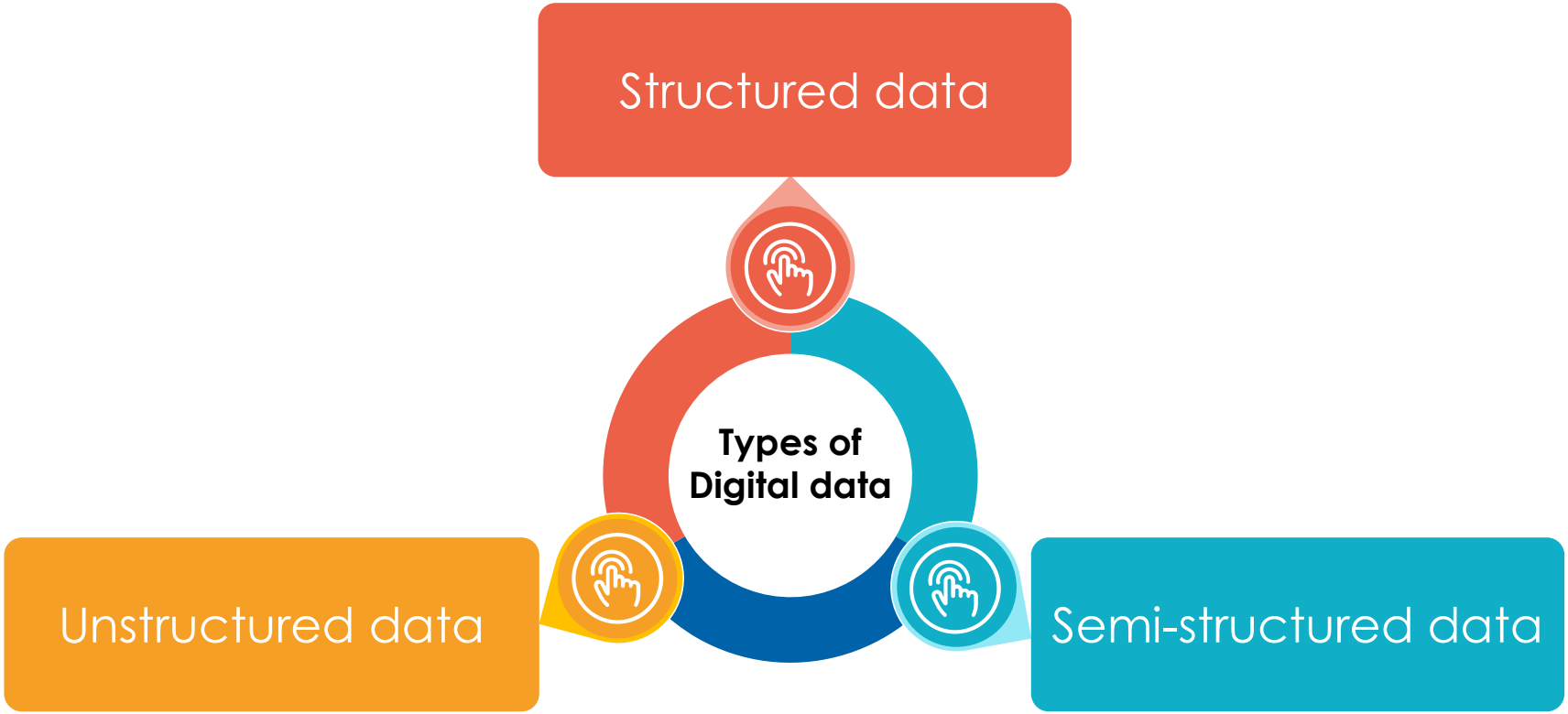
Information that comes from outside of company's or provided 3rd party vendor.

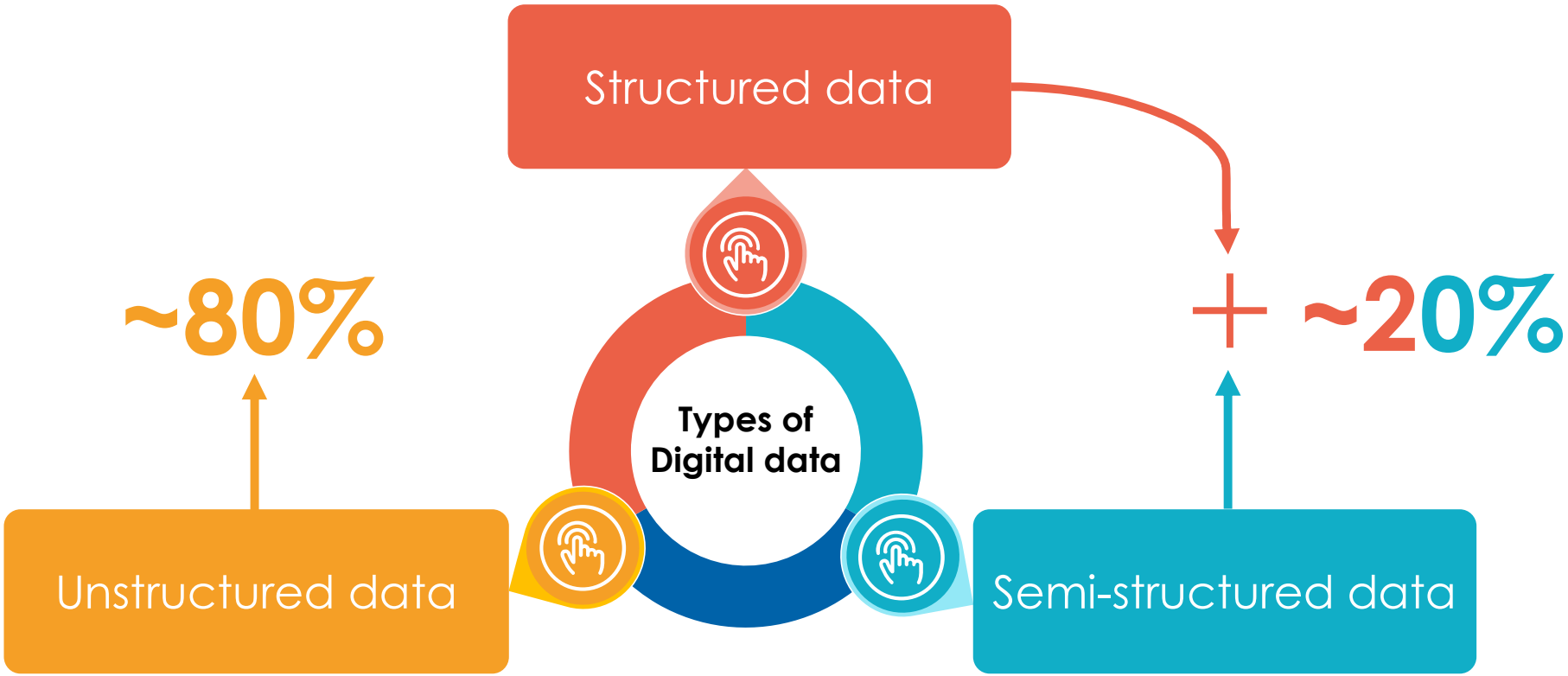
Example:

- Internet, Social Media data
- Government
- Market Research Organization
- Business Partners**

Types of Digital Data




Structured data
Types of Digital data
Unstructured data
Semi-structured data


Structured data
~80%
+ ~20%
Types of Digital data
Unstructured data
Semi-structured data

Structured Data

- Structured data can be defined as the data that has a defined repeating pattern.
- This pattern makes it easier for any program to sort, read, and process the data.
- Processing structured data is much easier and faster than processing data without any specific repeating patterns.
- Example:

Customer ID	Customer Name	Product ID	City	State
12345	Smith	214	Mumbai	Maharashtra
23456	John	365	Bangalore	Karnataka
34567	Nick	222	Pune	Maharashtra
45678	Sagar	456	Chennai	Tamil Nadu

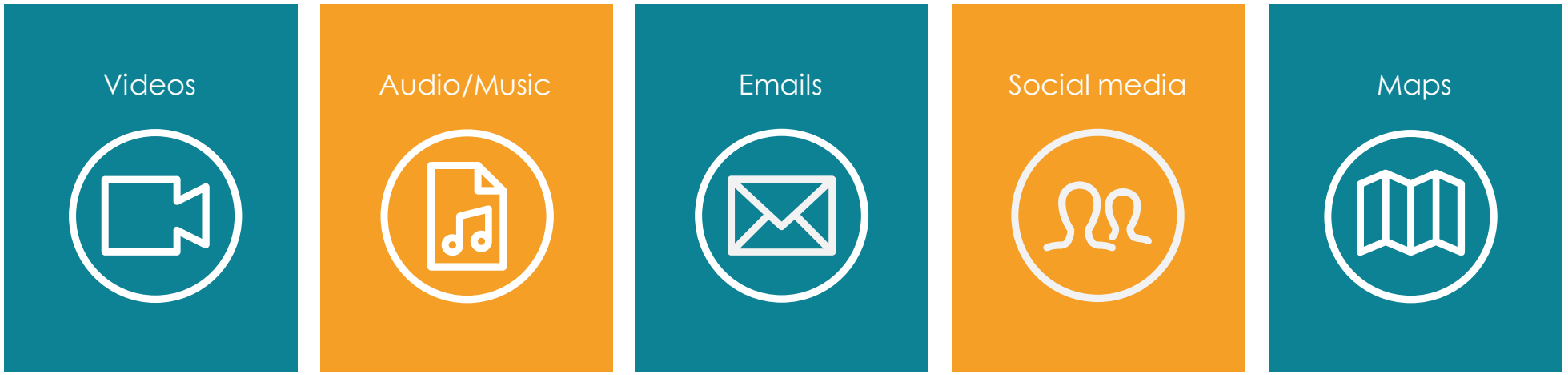
• Semi-Structured Data •

- Semi-structured data, also known as having a schema-less or self describing structure, refers to a form of structured data that contains tags or markup elements in order to separate out the elements and generate hierarchies of records and fields in the given data.
- Example:

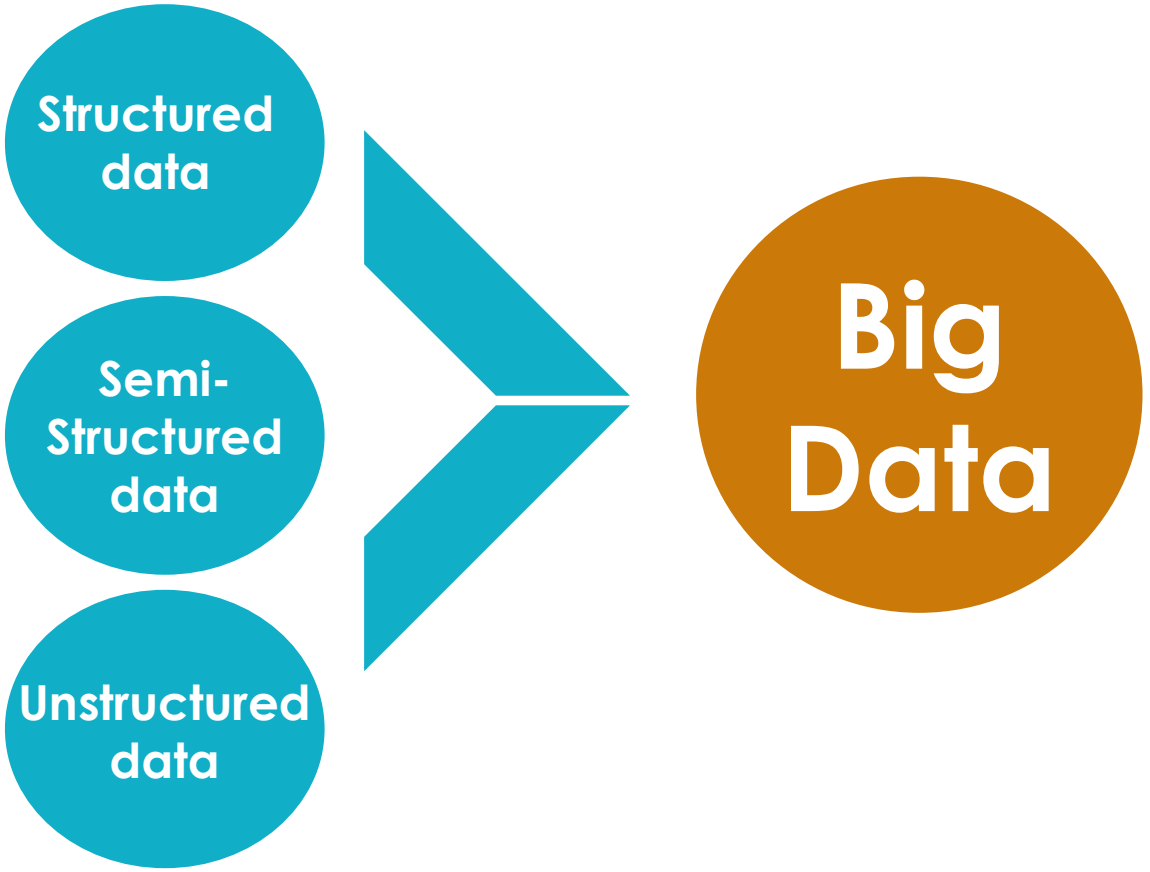
#No	Name	Email
1	Sam Jocabs	smj@xyz.com
2	First Name : David Last Name : Brown	davidb@xyz.com
3	Nick Sagar	Email-1: nick.sager@xyz.com Email-2: nicksager@gmail.com
4	First Name : John Middle Name : P Last Name : Todd	Personal Email: johntodd@gmail.com Business Email: john@xycompany.com

Unstructured Data

- Unstructured data is a set of data that might or might not have any logical or repeating patterns. About 80% of enterprise data consist of unstructured content.
- What are the unstructured data ?



What if it gets combine?



What is a Big Data?

- ✓**Big data refers to the datasets that are too large or complex to be manage by traditional data-processing software.**

- ✓**Big data is structured, semi-structured and unstructured or heterogeneous in nature. It becomes difficult for computing system to manage "Big Data" because of the extreme speed and volume at which it is generated**

Characteristics of Big Data

Big Data

Challenge with existing system	Classified into 4Vs	Complex data ?
Is a new data challenge that requires leveraging existing system differently	Volume Velocity Variety Veracity	Is usually unstructured and qualitative in nature

BIG DATA

Tweets

Every second ~6000 tweets

UPI

Every day around

~6.5 billion transactions

Facebook

Every minute
~5 lacs comments,
~3 lacs status update,
~1.5 lacs photos upload

E-Commerce

Every day approximate
1.6 million shipment

Big Data - 4Vs

According to Gartner, data is growing at the rate of 59% every year. This growth can be depicted in terms of the following four Vs.

Volume

Volume is the amount of data generated by organizations or individuals.

Variety

Variety describe the different formats for data such as images, text, video, audio, GPS.

Velocity

Velocity describes the rate at which data is generated, captured, and shared

Veracity

Veracity generally refers to the uncertainty of data. Whether the obtained data is correct or consistent.

THE 4 V'S OF BIG DATA

Volume

SCALE OF DATA

40 ZETTABYTES
of data will be created by 2020, an increase of 300 times from 2005

6 BILLION PEOPLE
have cell phones
WORLD POPULATION: 7 BILLION

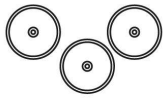
2.5 QUINTILLION BYTES
of data are created each day

Most companies in the U.S. have at least
100 TERABYTES
of data stored

Variety

DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES



30 BILLION PIECES OF CONTENT
are shared on facebook every month

4 BILLION + HOURS OF VIDEO
are watched on You Tube each month

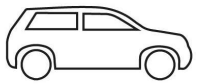
4 MILLION TWEETS
are sent per day by about 200 million monthly active users

Velocity

ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures
1TB OF TRADE INFORMATION
during each trading session

Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure



Veracity

UNCERTAINITY OF DATA

1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions

27% OF RESPONDENTS
in one survey were unsure of how much of data was inaccurate

Reference : http://www.ibmbigdatahub.com/infographic/four-vs-big-data

Future of Big Data

- **Most organizations today consider data and information to be their most valuable assets, that big data can unlock significant value by making information transparent.**

- **Sophisticated analytics can improve decision-making, minimize risks, and actionable insights that would otherwise remain hidden.**

- **At the same time, the volume and variety of data is also increasing at the immense rate every day, so big data can be used to develop the next generation of products and services.**

Thank You