Page No.
Date:

Course 3:
Prepare Data for Exploration

That will cover

- How data is generated.
- Different formats
- Analyze bias and credebility
- What is "clean data"
- Data bases
- Extract own data - Spreadsheets & SQL.
- Basics of data org.
- Protecting data

Week 1

~ Data is generated all time, every minute millions text and hundreds of millions of emails are sent. Millions of online searches are made and videos are viewed. Those numbers are only growing.

Page No.:
Date:

"Every piece of information is Data"

All data is usually generated as a result of our activity in the world.

- Interviews, Observations, Forms, Questionnaires, Surveys, Cookies.

Factors to consider while collecting data:

- How the data will be collected
- Choose data sources
- Decide what data to use
- How much to collect
- Select the right Data type
- Determine timeframe

FP, SP, TP → Choose data sources

Sample, population → Decide what data to use

Historical Data → Determine timeframe

- Buzz Words...

First Party Data:

Data collected by an individual or a group using their own resources.

- Prefered: As you collect only those which you know for exactly what.

Second Party Data:

Data collected by a group Directly from its audience and then sold.

Third Party Data:

Data collected from outside sources who did not collect it directly.

- Might not reliable.

Sample:

Sample is a part of population that is representative of the population

Page No.:
Date:

Discrete data:
Can be counted and have limited number of values.

Continuous data:
Can be measured and can have almost many numeric value.

Internal Data:
Data that lives inside a company's own systems

External Data:
Lives outside of company's database

Nominal & Ordinal data.

Structured & Unstructured.

Mostly the data being generated is unstructured. Audio, Video, Images, emails, social media text,

Page No.:
Date:

Data Model:

A Model that is used for organizing data elements and how they relate to one another.

Data elements = Attributes.

Structured	Unstructured
• Defined DTypes	• Varied DTypes
• Mostly quantitative	• Mostly qualitative
• Easy - Organize - Search - Analyze	• Difficult to search • Provides more freedom for analysis
• Stored - RDBMS - Warehouse	• Stored - Data Lakes NoSQL Warehouses
• Contained in Rows & Cols	• Can't be put Row & Cols.

"Data models help us keep data consistent and they give us map of how data is organized"

What is Data Modeling?

It is the process of creating diagrams, that visually represent how data is organized and structured.

'Visual representations are called Data Models.'

'A blueprint of a house'

'Each level of data modeling has different level of detail'.

CONCEPTUAL

LOGICAL

PHYSICAL

(1) **Conceptual:**

Gives a high-level view of your data structure - How you want data to interact across an organization.

(2) **Logical:**

Focuses on the technical details of the model such as relationships, attributes & entities.

Date:

(3) Physical:

Show in actual that how the database is built.

Here you lay-out how each database will be put in place and DB, APPs, Features will interact in specific detail

→ Data Modeling Techniques:

Two common: ERD (high level)

&

UML (detailed)

Page No.:
Date:

* Data Type:

A specific kind of data attribute that tells what kind of value the data is.

use - Record (Row) & Field (for Col).

WIDE DATA

Data in which every data subject has a single row with multiple columns to hold the values of various attributes of the subject.

LONG DATA

Data in which each row is one time point per subject, so each subject will have data in multiple rows.

(Only one column to add for new attribute)

Date:

NAME	YEARS	HEIGHT
Ayush	2011	120

Wide

Name	YR-2011	YR-2012	YR-2013
Ayush	120	130	170
Baej	120	135	140
Heena	110	115	116

Long

Name	Years	Height
Aayush	2011	120
Aayush	2012	130
Aayush	2013	170
Baej	2011	120
Baej	2012	135
Baej	2013	140
Heena	2011	110
Heena	2012	115
Heena	2013	116

Data Transformation:

A process of changing data's format, structure or values.

Doing this - you know before your analysis?

- Adding, Copying, replicating
- Deleting fields or records
- Standardizing names of variables
- Renaming, Moving, combining
- Joining.

Why to transform the data?