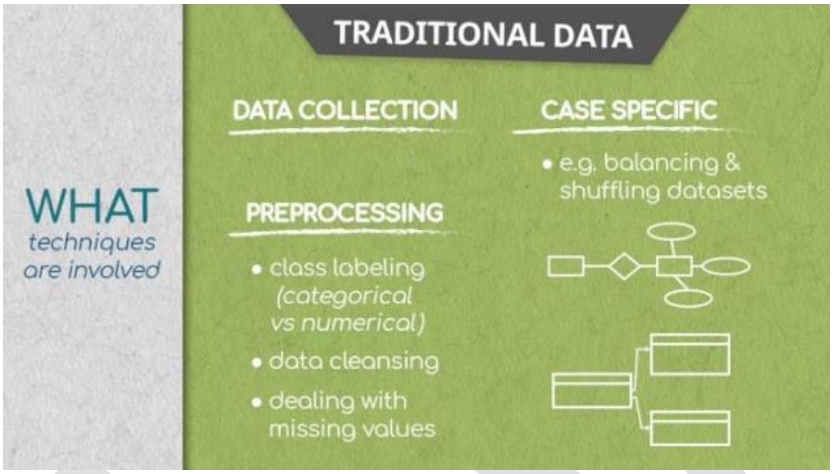
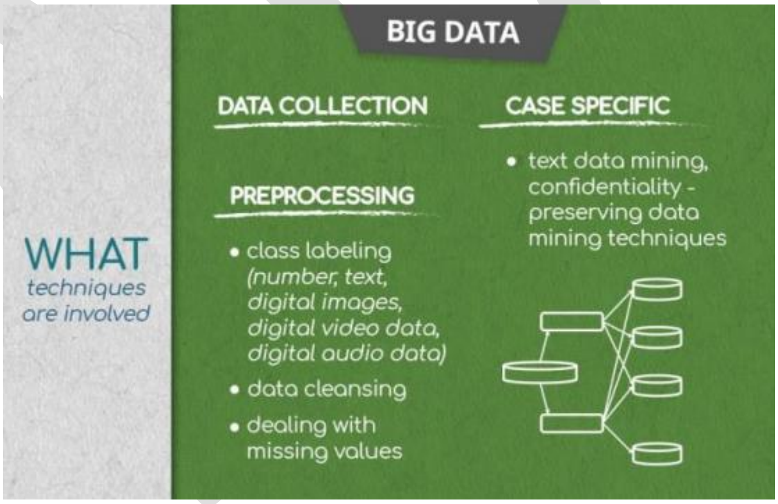
Unit – 2

1. Define traditional and Big Data?

Traditional data is the structured data which is being majorly maintained by all types of businesses starting from very small to big organizations. In traditional database system a centralized database architecture used to store and maintain the data in a fixed format or fields in a file



Big Data is a collection of data that is huge in volume, yet growing exponentially with time. It is a data with so large size and complexity that none of traditional data management tools can store it or process it efficiently. Big data is also a data but with huge size. Big data is a field that treats ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software.



~~~~~~

2. Techniques involved in traditional data?

Same as above qn

~~~~~~

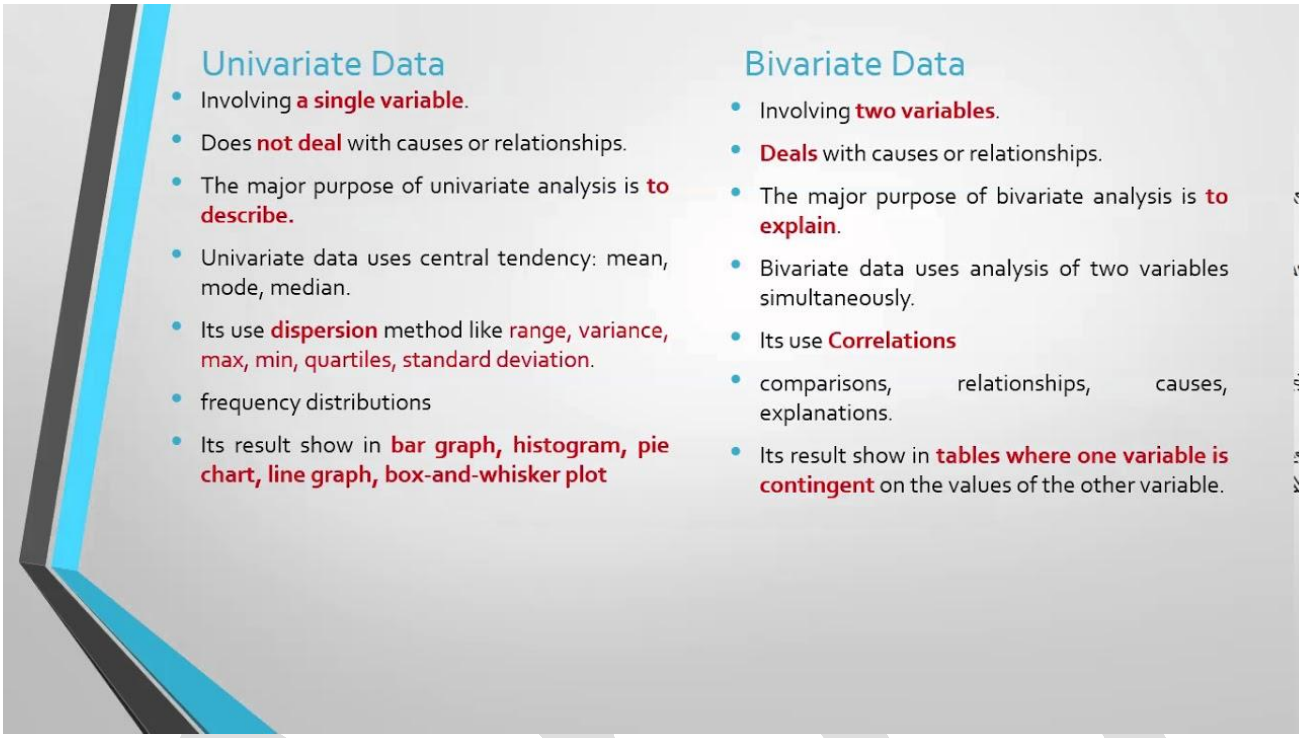
3. Define data shuffling

Data shuffling : Re-arranging data points to eliminate unwanted patterns and improve predictive performance further on. This is applied when, for example, if the first 100 observations in the data are from the first 100 people who have used a website; the data isn't randomized. Simply we put, shuffling techniques aim to mix up data and can

optionally retain logical relationships between columns. It randomly shuffles data from a dataset within an attribute (e.g. a column in a pure flat format) or a set of attributes (e.g. a set of columns)

---

4. What is Univariate and Bi-variate Analysis?



Example -

Sample question: How many of the students in the freshman class are female?	Sample question: Is there a relationship between the number of females in Computer Programming and their scores in Mathematics?

---

5. Mention few data repositories?

---

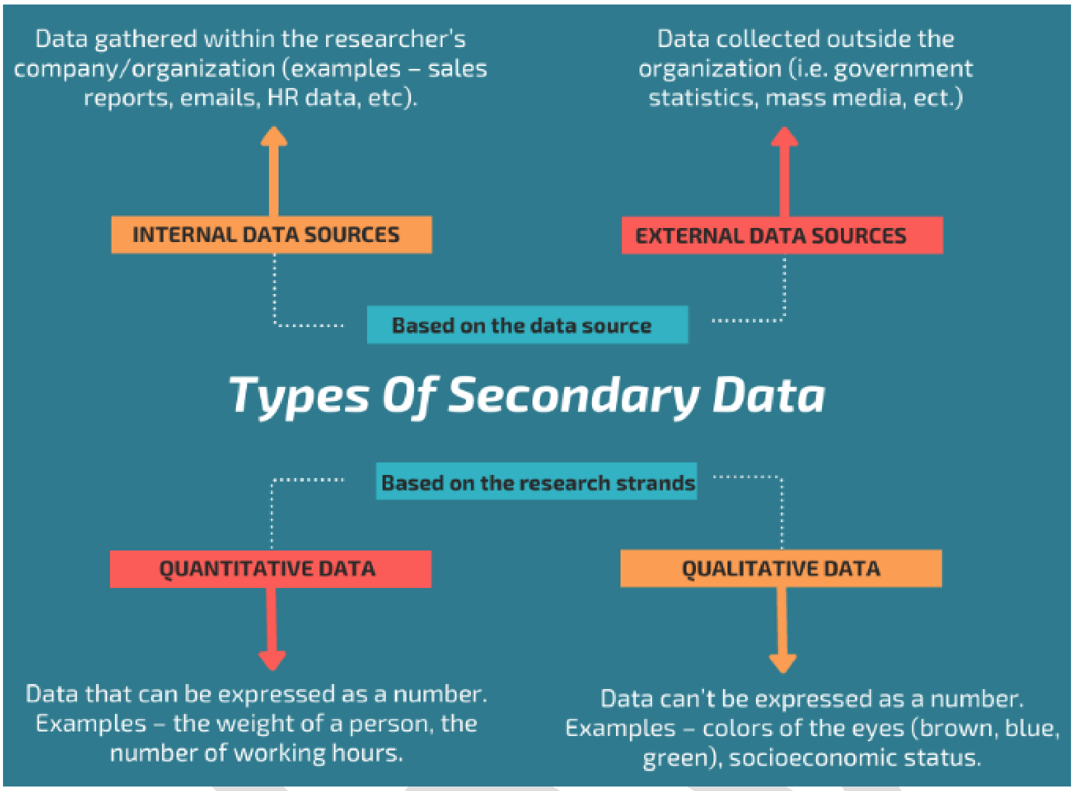
6. Explain any 2 Categories of API.

Web APIs: Web APIs are APIs that can be accessed using the HTTP protocol. The API defines endpoints, and valid request and response formats. Web APIs include the APIs used to communicate with the browser (see list). They may be services such as web notifications and web storage. Different web APIs feature varying levels of security and privacy, including open, internal and partner APIs. Multiple web APIs can be combined into a composite API - a collection of data or service APIs.

Open APIs: Open APIs, also known as external or public APIs, are available to developers and other users with minimal restrictions. They may require registration, and use of an API key, or may be completely open. They are intended for external users (developers at other companies, for example) to access data or services. As an example, take a look at the provided by the UK government. Any developer can access it, without even registering, allowing app builders to include governmental data on restaurant standards in their apps.

---

7. List and Explain 2 types of Secondary data briefly.



There are two types of secondary data, based on the data source:
Internal sources of data: information gathered within the researcher's company or organization (examples – a database with customer details, sales reports, marketing analysis, your emails, your social media profiles, etc).
External sources of data: the data collected outside the organization (i.e. government statistics, mass media channels, newspapers, etc.)

Also, secondary data can be 2 types depending on the research strands:
Quantitative data – data that can be expressed as a number or can be quantified. Examples – the weight and height of a person, the number of working hours, the volume of sales per month, etc. Quantitative data are easily amenable to statistical manipulation.
Qualitative data – the information that can't be expressed as a number and can't be measured. Qualitative data consist of words, pictures, observations, and symbols, not numbers. It is about qualities. Examples – colors of the eyes (brown, blue, green), your socioeconomic status, customer satisfaction, and etc.

~~~~~~

8. What is feature engineering?

Feature engineering is the process of using domain knowledge to extract features from raw data. These features can be used to improve the performance of machine learning algorithms. Feature engineering can be considered as applied machine learning itself.
For example, the decision tree based algorithms take into consideration only one feature at a time and divide the set into one part where the values of a considered feature are higher than an arbitrary threshold and the second part where values are lower.

~~~~~~

9. Define Outliers and causes Outliers?

Outliers are extreme values that deviate from other observations on data , they may indicate a variability in a measurement, experimental errors or a novelty. In other words, an outlier is an observation that diverges from an overall pattern on a sample.

Most common causes of outliers on a data set:

- Data entry errors (human errors)
- Measurement errors (instrument errors)
- Experimental errors (data extraction or experiment planning/executing errors)
- Intentional (dummy outliers made to test detection methods)
- Data processing errors (data manipulation or data set unintended mutations)
- Sampling errors (extracting or mixing data from wrong or various sources)
- Natural (not an error, novelties in data)

In the process of producing, collecting, processing and analyzing data, outliers can come from many sources and hide in many dimensions. Those that are not a product of an error are called novelties.Detecting outliers is of major importance for almost any quantitative discipline (ie: Physics, Economy, Finance, Machine Learning, Cyber Security). In machine learning and in any quantitative discipline the quality of data is as important as the quality of a prediction or classification model

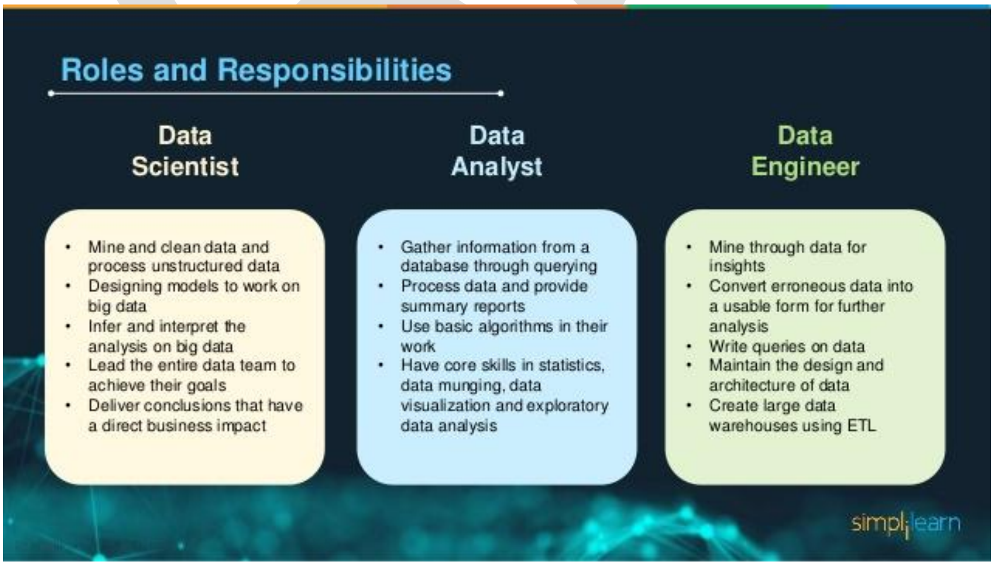
~~~~~~

10. What is pre-processing?

Data preprocessing is an important step in the data mining process. The phrase "garbage in, garbage out" is particularly applicable to data mining and machine learning projects. Data-gathering methods are often loosely controlled, resulting in out-of-range values, impossible data combinations, and missing values, etc. In any Machine Learning process, Data Preprocessing is that step in which the data gets transformed, or Encoded, to bring it to such a state that now the machine can easily parse it. In other words, the features of the data can now be easily interpreted by the algorithm.

~~~~~~

11. What is the difference between Data Analyst and Data Engineer?



~~~~~~

12. Explain Data pre-processing?

Data preprocessing is an important step in the data mining process. The phrase "garbage in, garbage out" is particularly applicable to data mining and machine learning projects. Data-gathering methods are often loosely
~~~~~~

controlled, resulting in out-of-range values, impossible data combinations, and missing values, etc. In any Machine Learning process, Data Preprocessing is that step in which the data gets transformed, or Encoded, to bring it to such a state that now the machine can easily parse it. In other words, the features of the data can now be easily interpreted by the algorithm.

~~~~~~

13. Explain about Data repositories?

A data repository can be defined as a place that holds data, makes data available to use, and organizes data in a logical manner. A data repository may also be defined as an appropriate, subject-specific location where researchers can submit their data.

Some common types of data repositories include:

- Data Warehouse. A data warehouse is a large data repository that brings together data from several sources or business segments. ...
- Data Lake. ...
- Data Mart. ...
- Metadata Repositories. ...
- Data Cubes. ...

~~~~~~

14. Explain missing value treatments?

Some techniques to treat the missing values:
<u>I. Deletion:</u> Unless the nature of missing data is 'Missing completely at random', the best avoidable method in many cases is deletion.
a. Listwise: In this case, rows containing missing variables are deleted.

User	Device	OS	Transactions
~~A~~	~~Mobile~~	~~NA~~	~~5~~
B	Mobile	Android	3
~~C~~	~~NA~~	~~iOS~~	~~2~~
D	Tablet	Android	1
E	Mobile	iOS	4

In the above case, the entire observation for User A and User C will be ignored for listwise deletion
b. Pairwise: In this case, only the missing observations are ignored and analysis is done on variables present.

User	Device	OS	Transactions
A	Mobile	~~NA~~	5
B	Mobile	Android	3
C	~~NA~~	iOS	2
D	Tablet	Android	1
E	Mobile	iOS	4

In the above case, 2 separate sample data will be analyzed, one with the combination of User, Device and Transaction and the other with the combination of User, OS and Transaction. In such a case, one won't be deleting

any observation. Each of the samples will ignore the variable which has the missing value in it. Both the above methods suffer from loss of information. Listwise deletion suffers the maximum information loss compared to Pairwise deletion. But, the problem with pairwise deletion is that even though it takes the available cases, one can't compare analyses because the sample is different every time.

II. Imputation

a. **Popular Averaging Techniques**: Mean, median and mode are the most popular averaging techniques, which are used to infer missing values. Approaches ranging from global average for the variable to averages based on groups are usually considered.

For example: if you are inferring missing value for Revenue, you might assign the average defined by mean, median or mode to such missing value. You could also consider taking into account some other variables such as Gender of the User and/or the Device OS to calculate such an average to be assigned to the missing values. Though you can get a quick estimate of the missing values, you are artificially reducing the variation in the dataset as the missing observations could have the same value. This may impact the statistical analysis of the dataset since depending on the percentage of missing observations imputed, metrics such as mean, median, correlation, etc may get affected.

OS	Revenue
Android	1,804
iOS	3,027
iOS	8,788
Android	NA
Android	3,735
Android	1,056
iOS	9,319
Android	6,199
Android	2,235
iOS	NA
Android	1,146

OS	Global Mean	Group Mean
Android	1,804	1,804
iOS	3,027	3,027
iOS	8,788	8,788
Android	4,145	2,696
Android	3,735	3,735
Android	1,056	1,056
iOS	9,319	9,319
Android	6,199	6,199
Android	2,235	2,235
iOS	4,145	7,045
Android	1,146	1,146

The above table shows the difference in imputed missing values of Revenue arrived by taking its global mean and mean based on which OS platform it belongs to.

b. **Predictive Techniques:** Imputation of missing values from predictive techniques assumes that the nature of such missing observations are not observed completely at random and the variables chosen to impute such missing observations have some relationship with it, else it could yield imprecise estimates.

In the examples discussed earlier, a predictive model could be used to impute the missing values for Device, OS, Revenues. There are various statistical methods like regression techniques, machine learning methods like SVM and/or data mining methods to impute such missing values.

~~~~~~

15. Explain any 2 methods of collecting Primary Data in detail.

Few methods of collecting Primary Data

1. **Interview method:** • The data collected during this process is through interviewing the target audience by a person called interviewer and the person who answers the interview is known as the interviewee. • Some basic business or product related questions are asked and noted down in the form of notes, audio or video and this data is stored for processing. • These can be both structured and unstructured like personal interviews or formal interviews through telephone, face to face, email etc.

2. **Survey Method:** • The survey method is the process of research, where a list of relevant questions are asked and answers are noted down in the form of text, audio, or video. • The survey method can be obtained in both online
~~~~~~

and offline mode like through website forms and email. • Then that survey answers data are stored for analyzing. Examples are online surveys or surveys through social media polls.

3. **Observation method**: The observation method is a method of data collection in which the researcher keenly observes the behavior and practices of the target audience using some data collecting tool and stores the observed data in the form of text, audio, video, or any raw formats. • In this method, the data is collected directly by posting a few questions on the participants. • For example, observing a group of customers and their behavior towards the products. • The data obtained will be sent for processing.

~~~~~~

16. What is an API? Explain 3 Basic elements of an API.
Same as qn 21 below

~~~~~~

17. What are the different methods to explore data?

1. Unique value count: One of the first things which can be useful during data exploration is to see how many unique values are there in categorical columns. This gives an idea of what is the data about. A unique value count of categorical columns in the cars dataset is shown here.
2. Frequency Count: Frequency count is finding how frequent individual values occur in column. For example, here is the frequency count for column "make".
3. Variance: When it comes to analysing numeric values, some basic information such as minimum, maximum and variance are very useful. Variance gives a good indication how the values are spread.
4. Pareto Analysis: Pareto analysis is a creative way of focusing on what is important. Pareto 80–20 rule can be effectively used in data exploration.
5. Histogram: Histogram are one of the data scientists favourite data exploration techniques. It gives information on the range of values in which most of the values fall. It also gives information on whether there is any skew in data.
6. Correlation Heat-map between all numeric columns: The term correlation refers to a mutual relationship or association between two things. In almost any business or for personal reasons, it is useful to express something in terms of its relationship with others. Finding correlation is very useful in data exploration, as it gives an idea on how the columns are related to each other.
7. Pearson Correlation and Trend between two numeric columns: Once you have visualised correlation heat-map , the next step is to see the correlation trend between two specific numeric columns.
8. Cramer-V correlation between all Categorical columns: Cramer-V is a very useful data exploration technique to find the correlation between categorical variables. And the result of Cramer-V can also be visualised using heat-map.
9. Correlation between two specific categorical columns: Once you have checked correlation between categorical columns using Cramer-V correlation matrix, you can further explore correlation between any two categorical columns.
10. Cluster size Analysis: We live in a world with immense amount of data. It is very easy to get bogged down by data overload. In order to survive in this ever-increasing data world, we need to look things from a high-level perspective.
11. Clustering or Segmentation: Once you have determined the number of clusters, the next step is to divide all data into specific number of clusters or segments.

~~~~~~

18. What are the methods to treat missing values?
Same as qn 14 above

~~~~~~

19. Explain any 4 Steps of Data Exploration and Preparation in detail.

Data exploration refers to the initial step in data analysis in which data analysts use data visualization and statistical techniques to describe dataset characterizations, – such as size, quantity, and accuracy, in order to better understand the nature of the data.

Steps of Data Exploration and Preparation

• Remember the quality of your inputs decide the quality of your output.

• So, once you have got your business hypothesis ready, it makes sense to spend lot of time and efforts here.

• Data exploration, cleaning and preparation can take up to 60-70% of your total project time.

• Below are the steps involved to understand, clean and prepare your data for building your predictive model:

1. Variable Identification : In this step, you have to first identify the input and output variables. Then, identify the data type and category of the variables.
2. Univariate Analysis : In univariate analysis, variables are explored one-by-one. This method depends on whether a variable type is categorical or continuous
3. Bi-variate Analysis (and Multi-variate) : Bivariate analysis is the analysis of bivariate data. It used to find out if there is a relationship between two sets of values. It usually involves the variables X and Y.
4. Missing values treatment : It is necessary to understand the concept of missing values because if it not handled properly, then inaccurate interference occurs. It can lead to incorrect prediction and classification.
5. Outlier treatment: The outlier is a data point that is distant from another different point. These outliers should remove from datasets.
6. Variable transformation: Data does not always come in a form that is immediately suitable for analysis. We often have to change variables before analysis. A transformation is a recursion of data using a function or some mathematical operation on each observation.
7. Variable creation: Creating variable if needed to fill the missing value.

• Finally, we will need to iterate over steps 4 – 7 multiple times before we come up with our refined model.

~~~~~~

20. What are the different sources of data collection?

Few methods of collecting Primary Data

1. **Interview method:** • The data collected during this process is through interviewing the target audience by a person called interviewer and the person who answers the interview is known as the interviewee. • Some basic business or product related questions are asked and noted down in the form of notes, audio or video and this data is stored for processing. • These can be both structured and unstructured like personal interviews or formal interviews through telephone, face to face, email etc.

2. **Survey Method:** • The survey method is the process of research, where a list of relevant questions are asked and answers are noted down in the form of text, audio, or video. • The survey method can be obtained in both online and offline mode like through website forms and email. • Then that survey answers data are stored for analyzing. Examples are online surveys or surveys through social media polls.

3. **Observation method**: The observation method is a method of data collection in which the researcher keenly observes the behavior and practices of the target audience using some data collecting tool and stores the observed data in the form of text, audio, video, or any raw formats. • In this method, the data is collected directly by posting a few questions on the participants. • For example, observing a group of customers and their behavior towards the products. • The data obtained will be sent for processing.

4. **Experimental method:** The experimental method is the process of collecting data through performing experiments, research and investigation.

Few methods of collecting Secondary Data [Secondary data is the data which has already been collected and used for some valid purpose.] :
~~~~~~

a) Internal source: • These types of data can easily be found within the organization such as market record, a sales record, transactions, customer data, accounting resources, etc. • The cost and time consumption is less in obtaining internal sources. • Most companies have a program for maintaining key-data, so much of the cleaning work may already be done. • This data can be stored in official data repositories such as databases, data marts, data warehouses and data lakes maintained by a team of IT professionals.

Data Repositories : - • The primary goal of a Database is to store data. • A Data warehouse is designed for reading and analyzing that data. • A Data mart is a subset of the data warehouse and used for serving a specific business unit. • Data lakes contains data in its natural or raw format. • But the possibility exists, that your data still resides in Excel files on the desktop of a domain expert.

b) External source: • The data which can't be found at internal organizations and can be gained through external third party resources is external source data. • The cost and time consumption is more because this contains a huge amount of data. • Examples of external sources are: Government publications, news publications, Registrar General of India, planning commission, international labor bureau, syndicate services and other non governmental publications. • Some companies provide data so that you, in turn they can enrich their services and ecosystem. – Such asTwitter, LinkedIn and Facebook. • Although data is considered an asset more valuable than oil by certain companies, more and more governments and organizations share their data for free with the world. – This data is helpful and convenient to train your data science skills at home.

3. Other Sources: a) Sensors data: With the advancement of IoT devices, the sensors of these devices collect data which can be used for sensor data analytics to track the performance and usage of products. b) Satellites data: Satellites collect a lot of images and data in terabytes on daily basis through surveillance cameras which can be used to collect useful information. c) Web traffic: Due to fast and cheap internet facilities many formats of data which is uploaded by users on different platforms can be predicted and collected with their permission for data analysis. – The search engines also provide their data through keywords and queries searched mostly.

~~~~~~

21. What is an API? Explain about any social media API for data collection in detail.

• Application programmable interface (API) allows one piece of code to interact with other code.
• One of the most common use case for APIs is on the web, – Sharing things on social media, making payments over the web, displaying list of tweets through a social handle, all of these services use API at the back.

<u>Basic elements of an API</u>

An API has 3 primary elements:
1. Access: is it the user or who is allowed to ask for data or services?
2. Request: is the actual data or service being asked for (e.g., if I give you current location from my game (Pokemon Go), tell me the map around that place). A Request has two main parts:
– Methods: i.e. the questions you can ask.
– Parameters: additional details you can include in the question or response.
3. Response: the data or service as a result of your request.

<u>Categories of API</u>

1. Web-based system:
• A web API is an interface to either a web server or a web browser.
• These APIs are used extensively for the development of web applications.
• These APIs work at either the server end or the client end.
• Companies like Google, Amazon, eBay all provide web- based API.
• Some popular examples of web based API are Twitter REST API, Facebook Graph API, Amazon S3 REST API, etc.
2. Operating system:
~~~~~~

• There are multiple OS based API that offers the functionality of various OS features that can be incorporated in creating windows or mac applications.
• Some of the examples of OS based API are Cocoa, Carbon, WinAPI, etc.
3. Database system:
• Interaction with most of the database is done using the API calls to the database.
• These APIs are defined in a manner to pass out the requested data in a predefined format that is understandable by the requesting client.
• This makes the process of interaction with databases generalised and thereby enhancing the compatibility of applications with the various database.
• Some popular examples are Drupal 7 Database API, Drupal 8 Database API, Django API.
4. Hardware System
• These APIs allows access to the various hardware components of a system.
• They are extremely crucial for establishing communication to the hardware.
• Due to which it makes possible for a range of functions from the collection of sensor data to display on your screens.
• For example, the Google PowerMeter API will allow device manufacturers to build home energy monitoring devices that work with Google PowerMeter.
• Some other examples of Hardware APIs are: QUANT Electronic, WareNet CheckWare,OpenVX Hardware Acceleration, CubeSensore, etc.

Example:
Facebook API
• Facebook API provides an interface to a large amount of data generated everyday. The innumerable post, omments and shares in various groups & pages produces massive data.
• This massive public data provides a large number of opportunities for analyzing the crowd.
• It is also incredibly convenient to use Facebook Graph API with both R and python to extract data.
• The Graph API is the primary way to get data into and out of the Facebook platform. It's an HTTP-based API that apps can use to programmatically query data, post new stories, manage ads, upload photos, and perform a wide variety of other tasks.
• The Graph API is named after the idea of a "social graph" — a representation of the information on Facebook. It's composed of:

- nodes — basically individual objects, such as a User, a Photo, a Page, or a Comment
- edges — connections between a collection of objects and a single object, such as Photos on a Page or Comments on a Photo
- fields — data about an object, such as a User's birthday, or a Page's name

Typically you use nodes to get data about a specific object, use edges to get collections of objects on a single object, and use fields to get data about a single object or each object in a collection.