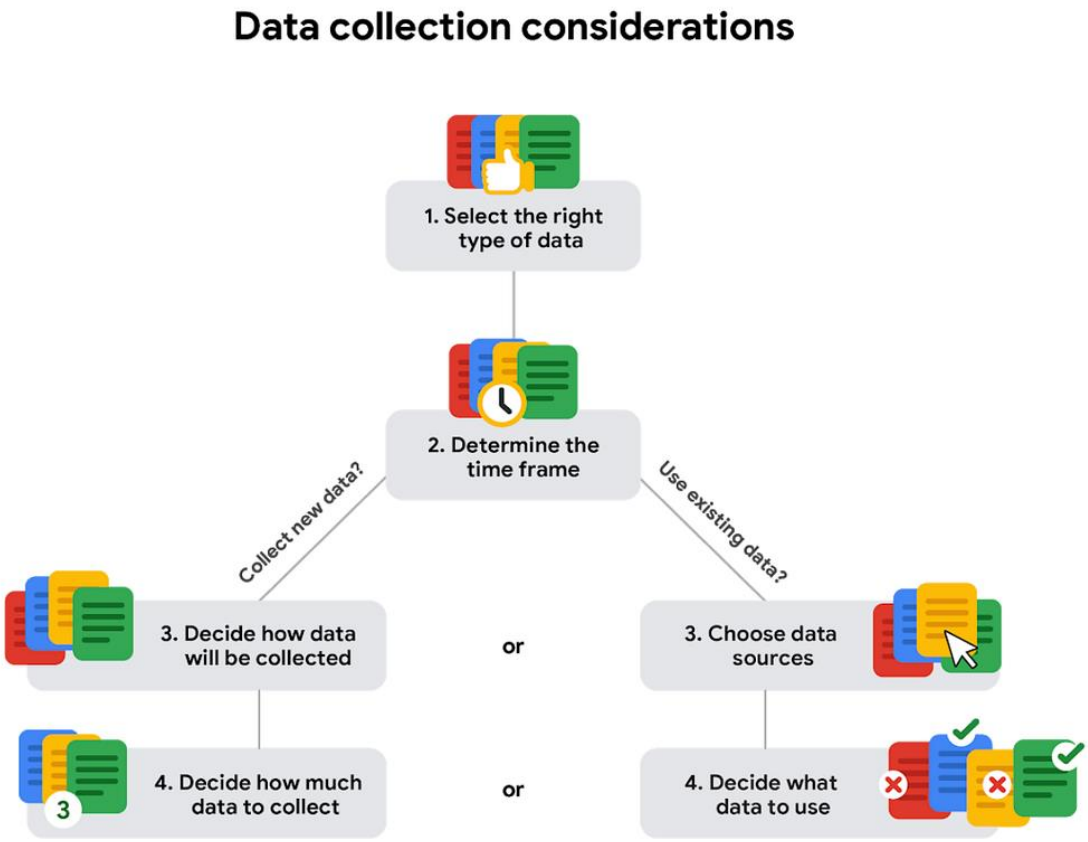
Course 3 - Prepare Data for Exploration

In Data Analysis, the third phase is **"The Prepare"** phase where data is gathered and organized for data analysis and this was what was treated in the third part of the Google Data Analytics Professional Course.

Data Preparation is where understanding the different types of data and data structures take place. It is the phase where you give answers to the question, "What type of data is right for the analysis I intend doing?"

In the process of collecting data, you need to consider some factors which are:



1. How will the data be collected?

You need to have a structure in place as to how the data needed for your analysis will be collected and ensure the process is efficient.

2. What data source?

Having answered the question of how your data will be collected, you need to know the data source and there are 3 data sources, which are:

- **First-party Data:** This is the form of data collected by an individual or group using their own resources. Collecting this type of data is the preferred method because you will know where

the data came from and you're certain of the steps that were put in place before the data was collected.

- **Second-party Data:** In this section, data is collected by a group directly from its audience and then sold to people who need such data for their analysis.
- **Third-party Data:** This is the type of data sold by a provider who didn't collect the data themselves.

Note that no matter the type of data you get, you need to inspect it for accuracy and trustworthiness.

3. How much data to collect?

In this section, you need to consider two things: "Population and Sample".

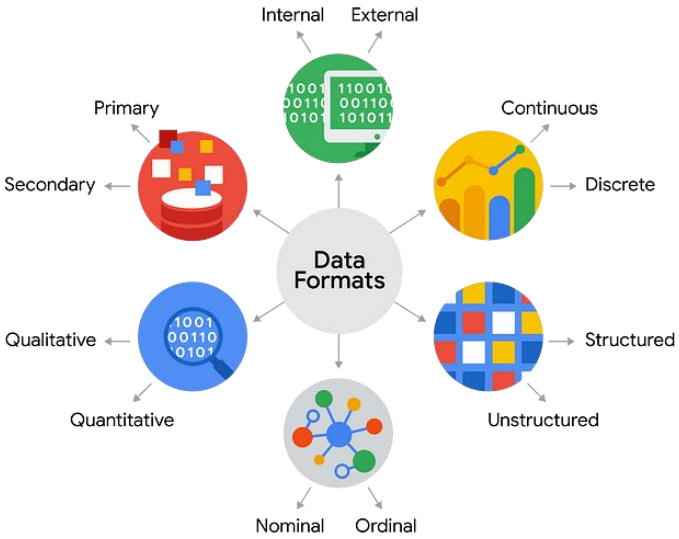
Population: This is all the data values in a certain dataset which can be stressful while generating data.

Sample: This is a part of a population that is representative of the population.

4. As you select data, you want to make sure you select the right data type.

5. Determine the time frame for data collection.

In the process of data preparation, there are different types of data formats such as:



Qualitative data, which is subdivided into **Nominal data** (this type of data is categorized without a set order, it doesn't have a sequence) and **Ordinal data** (this type of data has a set order or scale).

Quantitative data which is subdivided into **Discrete data** (data that is counted and has a limited number of values) and **Continuous data** (data that is measured and can have almost any numeric value).

Additionally, we have external, structured and unstructured data.

External Data: This type of data lives and is generated outside of an organization.

Structured Data: This type of data is organized in certain formats such as rows and columns. Spreadsheets and relational databases are softwares.

Unstructured Data: This data type is not organized in any easily identifiable manner (audio files and video files).

Structured data works nicely with a data model. Meanwhile, a model is used for organizing data elements and how they relate to each other.

Data models help to keep data consistent and provide a map of how data is organized. There are 3 types of data modeling:

1. **Conceptual data modeling:** This gives a high-level view of the data structure, such as how data interacts across an organization. For example, a conceptual data model may be used to define the business requirements for a new database.
2. **Logical data modeling:** This focuses on the technical details of a database such as relationships, tributes and entities. For example, a logical data model defines how individual records are uniquely identified in a database.
3. **Physical data modeling:** This depicts how a database operates. A physical data model defines all entities and attributes used.

A lot of approaches when it comes to data modeling techniques have two common methods: Entity Relationship Diagram (ERD) and Unified Modeling Language (UML) diagram.

Entity Relationship Diagrams (ERDs) are visual ways to understand the relationship between entities in the data model.

Unified Modeling Language (UML) diagrams are very detailed diagrams that describe the structure of a system by showing the system's entities, attributes, operations and their relationships.

Another way to describe data is through the **Data type**. Data type is the specific kind of data attribute that tells what kind of value a form data is and this can vary based on the query language that is being used. This brings us to the term **Data Transformation**.

Data transformation is the process of changing the format, structure or values of any form of data. Data transformation usually involves some steps such as:

- Adding, copying or replicating data
- Deleting fields or records
- Standardizing the names of variables
- Renaming, moving or combining columns in a database
- Joining one set of data with another
- Saving a file in a different format. For example: Saving a spreadsheet as a comma separated value (CSV) file.

Why transform data?

- Data Organization: It helps data to be better organized and easier to use.
- Data Compatibility: Different applications or systems can then use the same data.

- Data Migration: Data with matching formats can be moved from one system to another.
- Data Merging: Data with the same organization can be merged together.
- Data Enhancement: Data can be displayed with more detailed fields.
- Data Comparison: Comparison of the data can be made.

From there, data bias was discussed and was defined to be a type of error that systematically skews results in a certain direction. We have different types of data bias such as:

- **Sampling Bias:** This is when a sample isn't representative of the population as a whole.
- **Unbiased Sampling:** This is when a sample is representative of the population being measured.
- **Observer Bias:** It's the tendency for different people to observe things differently.
- **Interpretation Bias:** The tendency to always interpret ambiguous situations in a positive or negative way.
- **Confirmation Bias:** Search for or interpret information in a way that confirms pre-existing beliefs.

Data bias needs to be avoided in preparation for analysis because it makes our analysis skewed to one direction or in favor of a particular sample which makes our analysis not credible.

Now, when choosing a data source, there are best practices to ensure that we make the right choice and those practices are acronymed ROCCC.

Reliable: Accurate, complete and unbiased information that's been vetted and proven fit for use.

Original: Validate with the original source of the data.

Comprehensive: Contains critical information to solve the problem at hand.

Current: The usefulness of data decreases as time passes so you need to be sure you're using a current piece of data.

Cities: Citing the information source makes it easy for credibility.

After getting our data from the right data source, there are some things we need to consider before using the data and they are called **"Data Ethics"**.

Data ethics are well founded standards of right and wrong that dictate how data is collected, shared and used. Some aspects of data ethics are:

Ownership: Individuals own the raw data they provide and they have the primary control over its usage, how it's been processed and how it's shared.

Transaction transparency: All data-processing activities and algorithms should be completely explainable and understood by the individual who provides their data.

Consent: This is an individual's right to know explicit details about how and why their data will be used before agreeing to provide it.

Currency: Individuals should be aware of financial transactions resulting from the use of their personal data and scale of these transactions.

Privacy: Preserving a piece of data subjects information and activity anytime a data transaction occurs.

In the process of ensuring data privacy, the need to consider **"Data anonymization"** is also important and this is the process of protecting people's private or sensitive data by eliminating their personally identifiable information (PII). PII are pieces of information that can be used to track down a person's identity. Some examples of data that need to be anonymized are: Telephone numbers, names, license plates, license numbers, etc.

Openness: Free access, usage, and sharing of data.

Other subjects that were discussed were databases, metadata and data sorting & organisation.

Databases are collections of data stored in a computer system. An example is the relational database.

Relational database is the type that contains a series of related tables that can be connected via their relationships. For two or more tables to be connected, they must all have one or more similar fields to make them related. Hence, the name.

The Branch ID is the key to connecting these fields together and they are of two types:

Primary Key: An identifier that references a column in which each value is unique.

Foreign Key: A field within a table that is a primary key in another table.

Note: A table can only have one primary key but multiple foreign keys.

Metadata is a form of data about data. This is used in database management to help data analysts interpret the contents of the data within the database. The 3 common types are:

Descriptive metadata: Metadata that describes a piece of data and can be used to identify it at a later point.

Structural metadata: Metadata that indicates how a piece of data is organized and whether it is part of one or more than one data collection.

Administrative metadata: Metadata that indicates the technical source of a digital asset.

Data sorting as it were, is the arranging of data into a meaningful order to make it easier to understand, analyze and visualize.

Best practices when organising data:

Naming convention: Consistent guidelines that describe the content, date or version of a file in its name.

Foldering:

Organizing files into folders.

Archiving older folders.

Aligning naming and storage practices with your team.

Developing metadata practices.