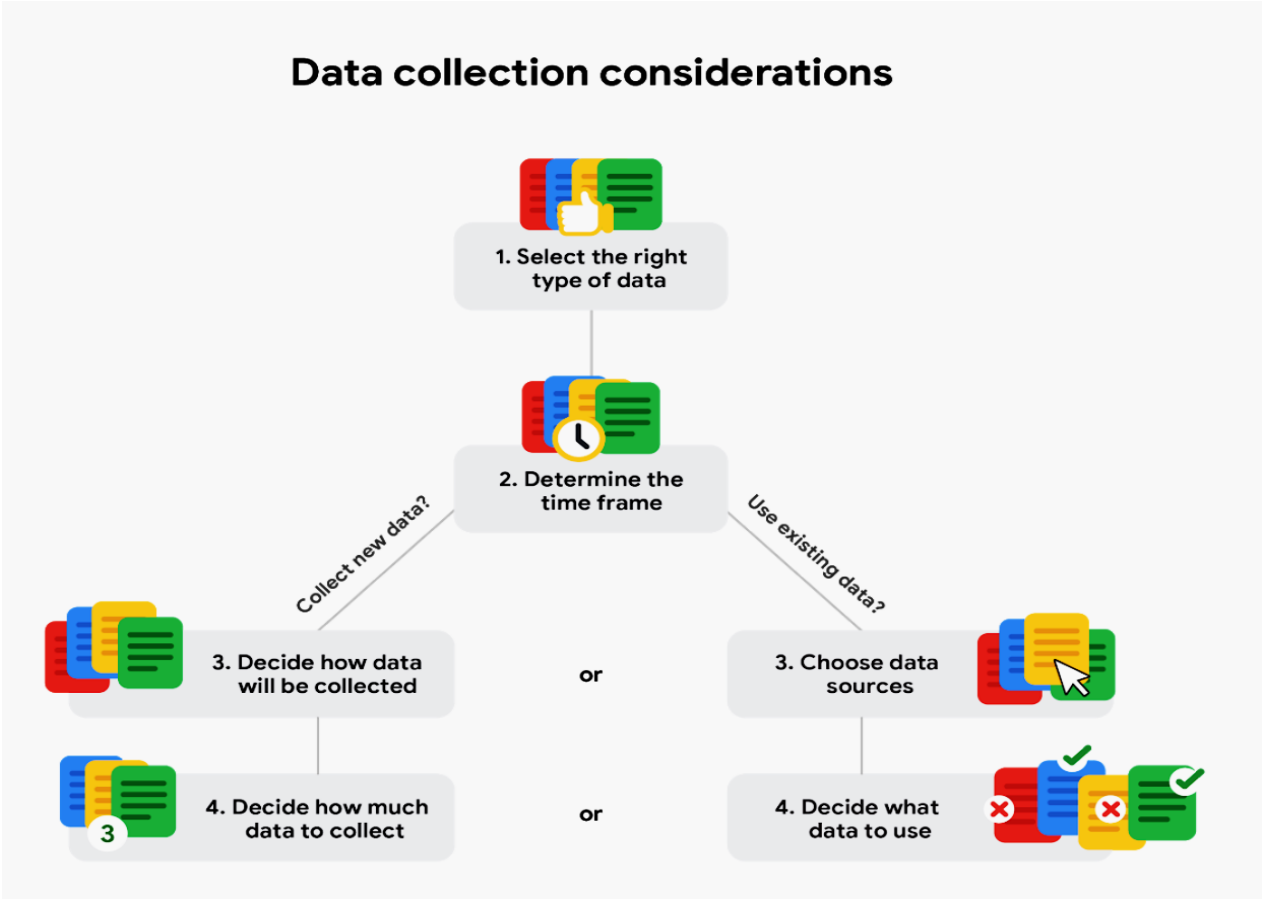
Google data analytics professional course

Week - 1

Collecting Data

Data collection considerations

- How the data will be collected
- Data sources
- Decide what data to use
- How much data to collect
- Select the right data type
- Determine the time frame for data collection



First-party data, *This is data collected by an individual or group using their own resources. Collecting first-party data is typically the preferred method because you know exactly where it came from.*

Second-party data*, which is data collected by a group directly from its audience and then sold. In our example, if you aren't able to collect your own data, you might buy it from an organization that's led traffic pattern studies in your city.*

Third-party data, *or data collected from outside sources who did not collect it directly. This data might have come from a number of different sources before you investigated it.*

Differentiate between data formats and data structures

Discover data formats

Two types of data

Quantitative

Qualitative

We can go even deeper into quantitative data and break it down into discrete or continuous data.

Quantitative

- ***Discrete data:***

This is data that's counted and has a limited number of values.

- ***Continuous data:***

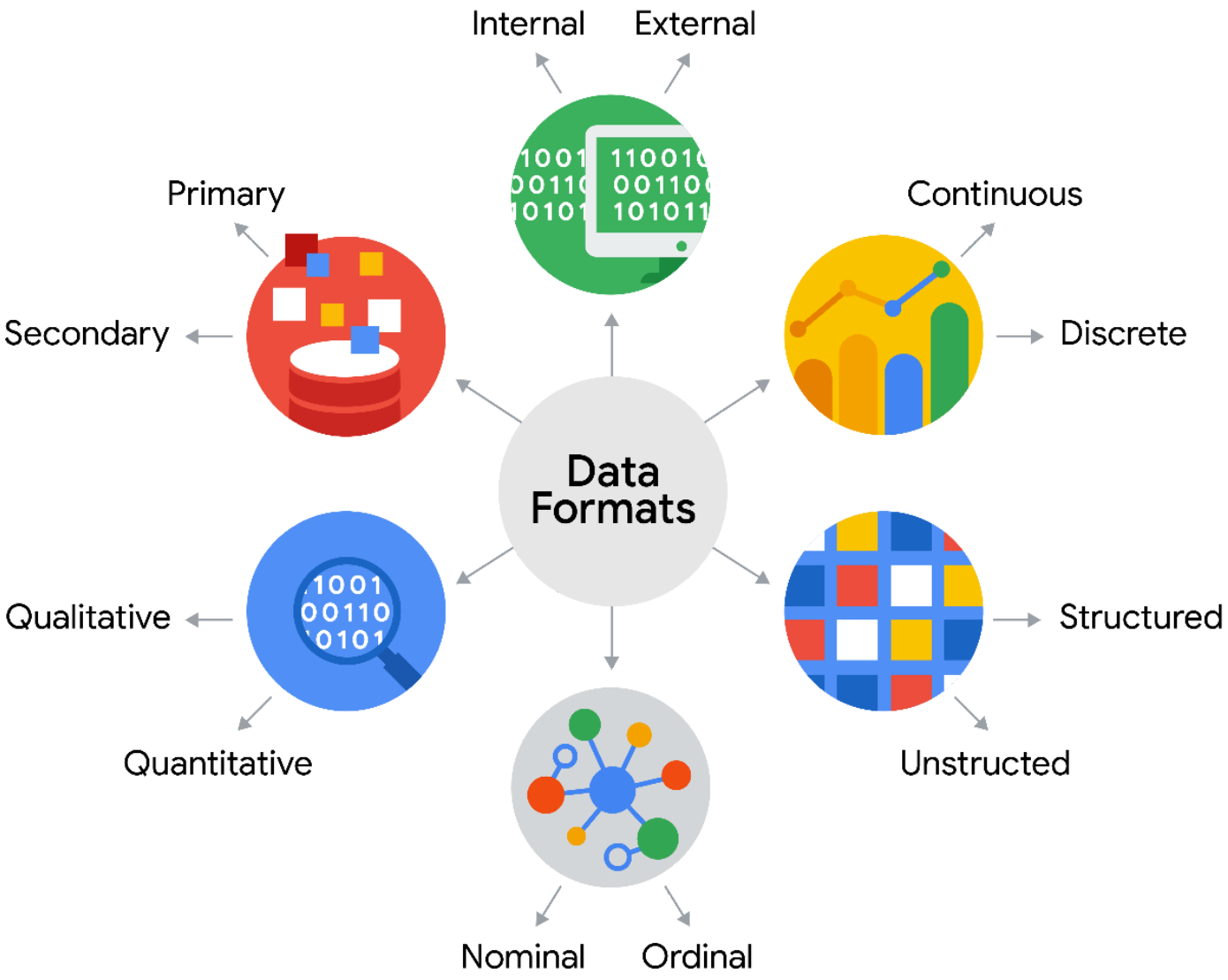
It can be measured using a timer, and its value can be shown as a decimal with several places.Example You could express that movie's run time as 110.0356 minutes. You could even add fractional data after the decimal point if you needed to.

Qualitative

- ***Nominal*** Eg: yes or no type data
- ***Ordinal data*** Eg: Rating between some values

Another types of data:

Internal data,
which is data that lives within a company's own systems.
External data, *is, you guessed it, data that lives and is generated outside of an organization.*
Structured data *is data that's organized in a certain format, such as rows and columns.*
This also helps make data visualization pretty easy because structured data can be applied directly to charts, graphs, heat maps, dashboards and most other visual representations of data.
Unstructured data,
This is data that is not organized in any easily identifiable manner. Audio and video files are examples of unstructured data because there's no clear way to identify or organize their content.


Internal
External
Primary
Secondary
Continuous
Discrete
Data Formats
Qualitative
Quantitative
Structured
Unstructed
Nominal
Ordinal

Primary vs. Secondary

Data Format Classification	Definition	Examples
Primary data	Collected by a researcher from first-hand sources	- Data from an interview you conducted - Data from a survey returned from 20 participants - Data from questionnaires you got back from a group of workers
Secondary data	Gathered by other people or from other research	- Data you bought from a local data analytics firm's customer profiles - Demographic data collected by a university - Census data gathered by the federal government

Internal vs. External

Data Format Classification	Definition	Examples
Internal data	Data that lives inside a company's own systems	- Wages of employees across different business units tracked by HR - Sales data by store location - Product inventory levels across distribution centers
External data	Data that lives outside of a company or organization	- National average wages for the various positions throughout your organization - Credit reports for customers of an auto dealership

Continuous vs Discrete

Data Format Classification	Definition	Examples
Continuous data	Data that is measured and can have almost any numeric value	- Height of kids in third grade classes (52.5 inches, 65.7 inches) - Runtime markers in a video - Temperature
Discrete data	Data that is counted and has a limited number of values	- Number of people who visit a hospital on a daily basis (10, 20, 200) - Room's maximum capacity allowed - Tickets sold in the current month

Qualitative vs. Quantitative

Data Format Classification	Definition	Examples
Qualitative	Subjective and explanatory measures of qualities and characteristics	- Exercise activity most enjoyed - Favorite brands of most loyal customers - Fashion preferences of young adults
Quantitative	Specific and objective measures of numerical facts	- Percentage of board certified doctors who are women - Population of elephants in Africa - Distance from Earth to Mars

Nominal vs. Ordinal

Data Format Classification	Definition	Examples
Nominal	A type of qualitative data that isn't categorized with a set order	- First time customer, returning customer, regular customer - New job applicant, existing applicant, internal applicant - New listing, reduced price listing, foreclosure
Ordinal	A type of qualitative data with a set order or scale	- Movie ratings (number of stars: 1 star, 2 stars, 3 stars) - Ranked-choice voting selections (1st, 2nd, 3rd) - Income level (low income, middle income, high income)

Structured vs. Unstructured

Data Format Classification	Definition	Examples
Structured data	Data organized in a certain format, like rows and columns	- Expense reports - Tax returns - Store inventory
Unstructured data	Data that isn't organized in any easily identifiable manner	- Social media posts - Emails - Videos

Understanding structured data

Data model

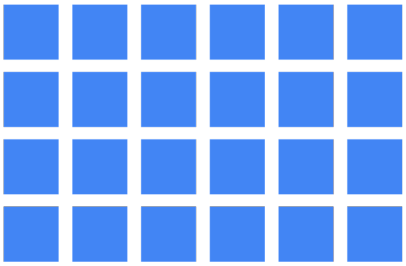
A model that is used for organizing data elements and how they relate to one another.

Data elements

They're pieces of information, such as people's names, account numbers, and addresses.

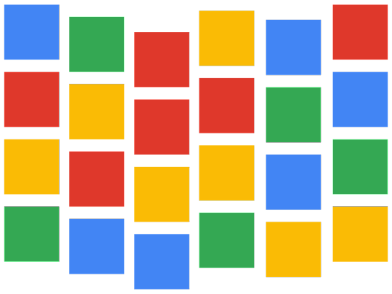
The structure of data

Structured data



- Defined data types
- Most often quantitative data
- Easy to organize
- Easy to search
- Easy to analyze
- Stored in relational databases & data warehouses
- Contained in rows and columns
- Examples: Excel, Google Sheets, SQL, customer data, phone records, transaction history

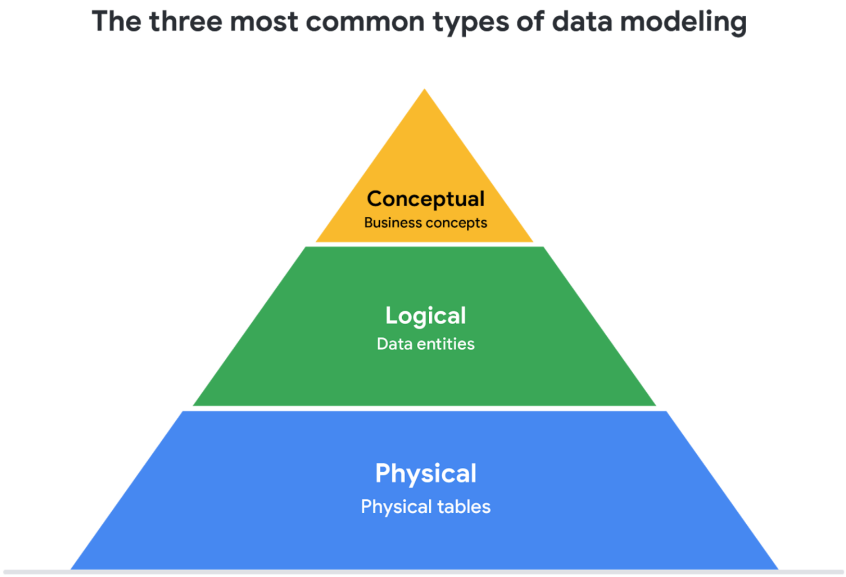
Unstructured data



- Varied data types
- Most often qualitative data
- Difficult to search
- Provides more freedom for analysis
- Stored in data lakes, data warehouses, and NoSQL databases
- Can't be put in rows and columns
- Examples: Text messages, social media comments, phone call transcriptions, various log files, images, audio, video

Data modeling levels and techniques

Types of data modeling



Data-modeling techniques

There are a lot of approaches when it comes to developing data models, but two common methods are the ***Entity Relationship Diagram (ERD)*** *and the* ***Unified Modeling Language (UML)*** *diagram.*

Explore data types, fields and values

Know the type of data you're working with

Data type

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

Basic data types:

- String
- Number
- Boolean

Understanding Boolean logic

- AND
- OR
- NOT

Additional Reading/Resources

- https://www.maa.org/press/periodicals/convergence/origins-of-boolean-algebra-in-the-logic-of-classes-george-boole-john-venn-and-c-s-peirce
- https://libguides.mit.edu/c.php?g=175963&p=1158594

Meet wide and long data

Transforming data

Data transformation is the process of changing the data's format, structure, or values. As a data analyst, there is a good chance you will need to transform data at some point to make it easier for you to analyze it.

Data transformation usually involves:

- 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 values (CSV) file.

Goals for data transformation might be:

- *Data **organization**: better organized data is 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**: apples-to-apples comparisons of the data can then be made*

Week - 2

Unbiased and objective data

Bias: From questions to conclusions

***Bias** has evolved to become a preference in favor of or against a person, group of people, or things.*
***Data bias** is a type of error that systematically skews results in a certain direction.*

Biased and unbiased data

Sampling bias is when a sample isn't representative of the population as a whole. You can avoid this by making sure the sample is chosen at random, so that all parts of the population.
Unbiased sampling results in a sample that's representative of the population being measured.

Understanding bias in data

Three more types of data bias, [Sampling bias +]

- observer bias (or Experimenter bias or research bias)
- interpretation bias
- confirmation bias

Observer bias

The tendency for different people to observe things differently.

Interpretation bias

The tendency to always interpret ambiguous situations in a positive, or negative way. Eg: Communicating with friend and manager same content but different understanding.

Confirmation bias

It is the tendency to search for, or interpret information in a way that confirms preexisting beliefs.

Explore data credibility

Identifying good data sources

ROCCC

- **R- Reliable**
- **O- Original**
- **C- Comprehensive**
- **C- Current**
- **C- Cited**

For good data, stick with vetted public data sets,academic papers, financial data and governmental agency data.

What is "bad" data?

Not ROCCC

Data ethics and privacy

Introduction to data ethics

Ethics *refers to well-founded standards of right and wrong that prescribe what humans ought to do, usually in terms of rights, obligations, benefits to society, fairness or specific virtues.*

Data ethics *refers to well- founded standards of right and wrong that dictate how data is collected, shared, and used.*

GDPR *- General Data Protection Regulation of the European Union*

There are lots of different aspects of data ethics but we'll cover six:

- *ownership,*
- *transaction transparency,*
- *consent,*
- *currency,*
- *privacy, and*
- *openness.*

Ownership

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

Transaction transparency

Which is the idea that ***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.*** *They should know answers to questions like why is the data being collected? How will it be used? How long will it be stored? The best way to give consent is probably a conversation between the person providing the data and the person requesting it.*

Currency

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

Privacy *means* ***preserving a data subject's information*** *and activity any time a data transaction occurs.*

- *Protection from unauthorized access to our private data,*
- *Freedom from inappropriate use of our data,*
- *The right to inspect, update, or correct our data,*
- *Ability to give consent to use our data,and*
- *Legal right to access our data.*

Data anonymization

Personally identifiable information - PII

Data anonymization

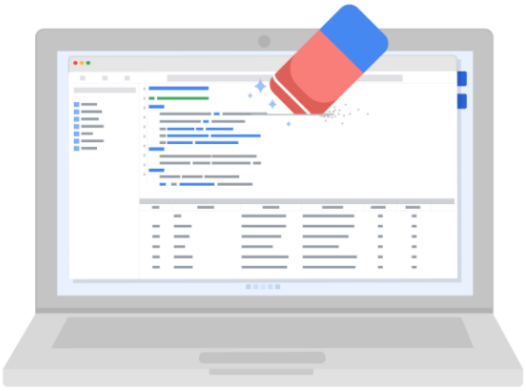
Data anonymization is the ***process of protecting people's private or sensitive data*** *by eliminating that kind of information. Typically, data anonymization involves* ***blanking, hashing, or masking personal information,*** *often by using fixed-length codes to represent data columns, or hiding data with altered values.*

What types of data should be anonymized?

Healthcare and financial data are two of the most sensitive types of data. These two industries usually goes through ***de-identification****, which is* ***a process used to wipe data clean of all personally identifying information****.*

Data anonymization is used in just about every industry.

- *Telephone numbers*
- *Names*
- *License plates and license numbers*
- *Social security numbers*
- *IP addresses*
- *Medical records*
- *Email addresses*
- *Photographs*
- *Account numbers*



Understanding open data

*Openness refers to **free access, usage and sharing of data**.preferably by downloading over the Internet in a convenient and modifiable form.*
Eg:data.gov

- *Be available and accessible to the public as a complete dataset*
- *Be provided under terms that allow it to be reused and redistributed*
- *Allow universal participation so that anyone can use, reuse, and redistribute the data*

***Data Interoperability** is the ability of data systems and services to openly connect and share data.*

Sites and resources for open data

- https://www.data.gov/
- https://www.census.gov/data.html
- https://www.opendatanetwork.com/
- https://cloud.google.com/solutions/datasets
- https://datasetsearch.research.google.com/

Week-3

Working with databases

All about databases

Database

A database is a collection of data stored in a computer system.

Metadata

Metadata is data about data.

Database features

Relational database

A relational database is a database that contains a series of related tables that can be connected via their relationships.

Primary key

A primary key is an identifier that references a column in which each value is unique.
A table can only have one primary key.

Foreign key

A foreign key is a field within a table that's a primary key in another table. A table in a relational database is allowed to have multiple foreign keys.

Managing data with metadata

Exploring metadata

Metadata *summarizes basic information about data.*

There are three common types of metadata:

- *descriptive,*
- *structural, and*
- *administrative*

Descriptive metadata:

It is metadata that describes a piece of data and can be used to identify it at a later point in time.

Eg: The descriptive metadata of a book in a library would include the code you see on its spine, known as a unique International Standard Book Number, also called the ***ISBN****.*

Structural metadata

Which is metadata that indicates how a piece of data is organized and whether it's part of one or more than one data collection.

Eg: Index of a book

Administrative metadata

It is metadata that indicates the technical source of a digital asset.

Eg: Details of the photo includes size, time, height, width etc..

Metadata is as important as the data itself

Metadata tells the who, what, when, where, which, how, and why of data.

Elements of metadata

- Title and description
- Tags and categories
- Who created it and when
- Who last modified it and when
- Who can access or update it

Using metadata as an analyst

A metadata repository is a database specifically created to store metadata.
Analysts use meta data because it tells what the data is.

Metadata management

Data governance is a process to ensure the formal management of a company's data assets.

Accessing different data sources

Working with more data sources

- Internal
- External

Internal data from internal source

Data from spreadsheet to spreadsheet

Eg:

=IMPORTRANGE("https://docs.google.com/spreadsheets/d/1utuuy9wrDP0g6TbkBzZgKIU6qVzPV7q2dLKm5urU_x4/edit#gid=0", "A1:O12")

From external source to a spreadsheet

Import from files like csv, xlsx etc..

or

Spreadsheet - IMPORTHTML(“ url ”, ” table ”, 1)

GUIDE

- *https://www.thedataschool.co.uk/anna-prosvetova/web-scraping-made-easy-import-html-tables-or-lists-using-google-sheets-and-excel/*

TABLE IN WEBSITE

- *https://en.wikipedia.org/wiki/Demographics_of_India*

Eg:

=IMPORTHTML("http://en.wikipedia.org/wiki/Demographics_of_India","table",1)

- *We can draw data only from table or list.*
- *The number is the index that refers to the order of the tables on a web page.*

Microsoft Excel

*You can import data from web pages using the **From Web** option:*

***Step 1 :** Open a new or existing spreadsheet.*

***Step 2:** Click Data in the main menu and select the **From Web** option.*

***Step 3:** Enter the URL and click OK.*

***Step 4:** In the Navigator, select which table to import.*

***Step 5:** Click **Load** to load the data from the table into your spreadsheet.*

Importing data from spreadsheets and databases

Data from the websites

https://www.who.int/data/gho/

Exploring public datasets

- https://cloud.google.com/solutions/datasets
- https://datasetsearch.research.google.com/
- https://www.kaggle.com/datasets
- https://cloud.google.com/bigquery/public-data

Public health datasets

- https://www.who.int/data/collections
- https://cloud.google.com/healthcare/docs/resources/public-datasets/tcia
- https://cloud.google.com/life-sciences/docs/resources/public-datasets/1000-genomes

Public climate datasets

- https://www.climate.gov/maps-data/all?listingMain=datasetgallery
- https://www.ncei.noaa.gov/weather-climate-links

Public social-political datasets

- https://data.unicef.org/resources/dataset/sowc-2019-statistical-tables/
- https://www.bls.gov/cps/tables.htm
- https://openpolicing.stanford.edu/

Sorting and filtering

Sort and filter in spreadsheet

Working with large datasets in SQL

Bigquery commands

```
SELECT
FROM
WHERE
```

```
SELECT
    count(*) as num_of_bikestrips  or   count(duration) as num_of_bikestrips
FROM
   `bigquery-public-data.london_bicycles.cycle_hire`
WHERE
    duration >= 1200;
SELECT
 name,
```

```
  count
FROM
  `babynames.names_2014`
WHERE
  gender = 'M'
ORDER BY
  count DESC     //count in descending order
LIMIT
  5
```

IN-Depth Bigquery

Dialects

Vendors of SQL databases may use slightly different variations of SQL. These variations are called SQL **dialects**.

- MySQL, PostgreSQL, and SQL Server, aren't case sensitive. This means if you searched for country_code = 'us', it will return all entries that have 'us', 'uS', 'Us', and 'US'.
- BigQuery is case sensitive, so that same search would only return entries where the country_code is exactly 'us'.
- We can use single or double quotations ' ' or " " .

-- command line

```
--This is an important query used later to join with the accounts table
SELECT
	rowkey, --key used to join with account_id
	Info.date, --date is in string format YYYY-MM-DD HH:MM:SS
	Info.code --e.g., 'pub-###'
FROM
	Publishers
```

Write neatly

```
SELECT
	CASE
		WHEN genre = 'horror' THEN 'Will not watch'
		WHEN genre = 'documentary' THEN 'Will watch alone'
		ELSE 'Watch with others'
		END AS watch_category, COUNT(movie_title) AS number_of_movies
FROM
	MovieTheater
GROUP BY
	1
```

Multi line command

```
/*
Date: September 15, 2020
Analyst: Jazmin Cisneros
Goal: Count the number of rows in the table
*/
SELECT
	COUNT(*) AS number_of_rows -- the * stands for all so count all
FROM
	table
```

Week-4

Effectively organize data

Benefits of organize data

- It makes it easier to find and use,
- Helps you avoid making mistakes during your analysis and
- Helps to protect it.

Best practices you can use when organizing data, including

- Naming conventions,
- Foldering, and
- Archiving older files.
- Align your naming and storage practices with your team to avoid any confusion
- Develop metadata practices

All about file naming

Naming conventions

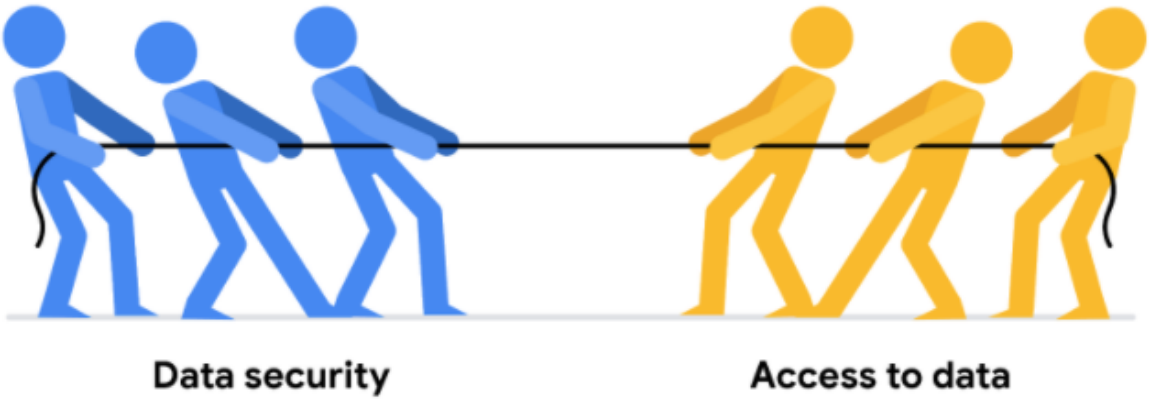
That describes the content, date, or version of a file and its name.

Securing data

Data security

Data Security means protecting data from unauthorized access or corruption by adopting safety measures.

Balancing security and analytics



Encrypt data after analysis

Week-5

Create or enhance your online presence

Why an online presence is important

A professional online presence

- It can help potential employers find you.
- lets you make connections with other data analysts in your field,
- learn and share data findings, and
- maybe even participate in community events.

Network building

- LinkedIn
- Instagram
- Github
- Kaggle
- Dataelixir https://dataelixir.com/newsletter-archives/
- Meetup https://www.meetup.com/topics/data-analytics/
- Tableau https://www.tableau.com/learn/series/how-we-do-data
- Kdnuggets https://www.kdnuggets.com/meetings/index.html
- Conference https://www.digitalanalyticsassociation.org/
- Data Science Assn https://www.datascienceassn.org/

Quick Review

Week-1

- ★ *Collect data*

Week-2

- ★ *Checking data (good or bad, biased or unbiased)*

Week-3

- ★ *Database*

Week-4

- ★ *Organize the data*

Week-5

- ★ *Network building*

Dhamodharan

09/10/2021