W1 Data types and structures

Data exploration

- Understanding the different types of data and data structures
- What type of data is right for the question you're answering
- Practical skills about how to extract, use, organize, and protect your data
- Prepare step
- How data is generated
- Different formats, types, and structures of data
- Analyze data for credibility
- What "clean data" means
- Databases
- Extract your own data from a database using spreadsheets and SQL
- Process of data organization and protecting your data
- When data is extracted it may be biased instead of credible, dirty instead of clean

1. Understanding Data Types and Structures: how data analysts decide which data to collect, structured vs. non structured, data types, data formats
2. Understanding Bias, Credibility, Privacy, Ethics, and Access
3. Databases - where data lives: how to access them, extract, filter, sort data, and metadata
4. Organizing and Protecting your Data: file naming conventions to keep work organized
5. Engaging in the Data Community

Collecting Data

Data collection in our world

- The ways data can be generated and how industries collect data themselves
- How data is collected? Interviews, observations, forms, questionnaires, surveys, cookies
- Knowing how data is generated can help add context to the data and know how to collect it can help the data analyst process more efficiently

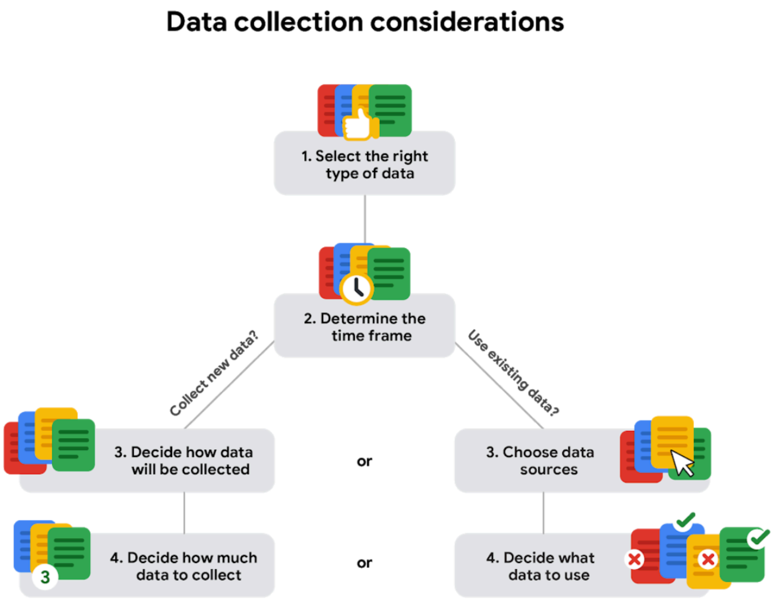
Determining what data to collect

- Deciding what kind of data to collect is a dilemma considering the nearly endless amount of data
- What factors to consider when collecting data?
- Usually the data will be given to you or your business talk will narrow down your choices

Data collection considerations

- How the data will be collected? --> your own resources? Or purchase?
- Choose data sources
    - First party data --> preferred method, you know where it came from
    - Second party data --> collected by a group from its audience and sold
    - Third party data --> outside sources collected, who didn't collect it directly, can be reliable as long as you check it for bias, accuracy, credibility
- Decide what data to use and not get distracted by other data: stay focused on your business problem
- How much data --> each project has its own needs
    - Population --> all possible data values in a certain database
    - Sample --> part of a population that is representative of the population
- Select right data type
- Determine time frame for data collection
    - Use historical data if answer is needed immediately
    - Might not have time to collect new data

Selecting the right data


Data collection considerations
1. Select the right type of data
2. Determine the time frame
Collect new data?
Use existing data?
3. Decide how data will be collected
or
3. Choose data sources
4. Decide how much data to collect
3
or
4. Decide what data to use

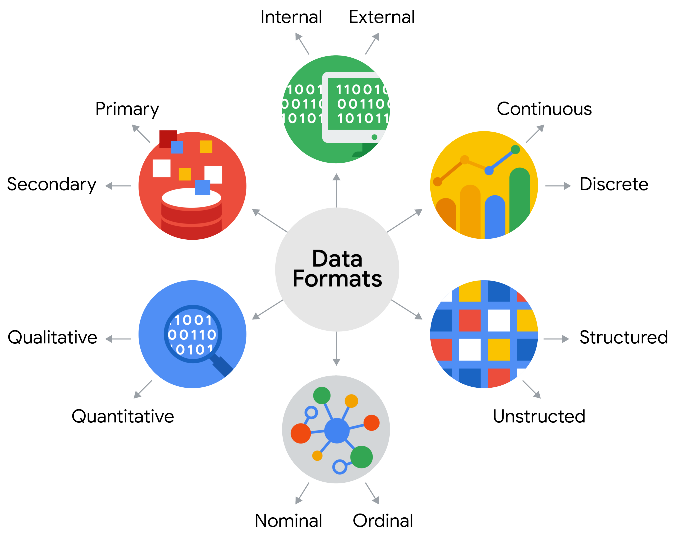
Differentiating between data formats and structures

- Discrete data --> data that is counted and has a limited number of values (could be money, points; when partial measurements aren't allowed its discrete)
- Continuous data --> data that is measured and can have almost any numeric value

- Nominal data --> a type of data that is categorized without a set order, no sequence
- Ordinal data --> a type of data with a set order or scale

- Internal data --> data that lives within a company's own system (usually more reliable and easier to collect)
- External data --> data that is generated outside of an organization (becomes valuable when your analysis)

- Structured data --> data organized in a certain format such as rows and columns (spreadsheets, relational databases); having a framework for the data makes the data easily searchable and more analysis ready
- Unstructured data --> data that is not organized in any easily identifiable manner (audio, video files); might have internal structure but doesn't fit neatly in rows/columns
  - You'll be working with structured data most of the time, it'll be turned to structured before you get to it
  - Structed data is in a **data model** --> a model used for organizing data elements and how they relate to one another, keeps data consistent
  - Data elements --> pieces of information such as people's names, account numbers, addresses to keep data consistent and well organized
  - Structured data is useful for databases and its easy to entry, query, analyze (if your data is exported the structure goes with the data)



the following table highlights the differences between primary and secondary data and examples of each

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

the following table highlights the differences between internal and external data and examples of each

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

the following table highlights the differences between continuous and discrete data and examples of each

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

the following table highlights the differences between qualitative and quantitative data and examples of each

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

the following table highlights the differences between nominal and ordinal data and examples of each

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)

the following table highlights the differences between structured and unstructured data and examples of each

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

Structured data	Unstructured data
• Defined data types • Most often quantitative • Easy to organize, search, analyze, store • Stored in relational databases and data warehouses • Contained in rows and columns • Ex. Excel, Sheets, SQL, customer data, phone records, transaction histories	• Varied data types • Most often qualitative data • Difficult to search • Provides more freedom for analysis • Stored in data lakes, data warehouses, NoSQL • Can't be put into rows, columns • Ex. Text messages, social media comments, phone call transcripts, image/video/audio

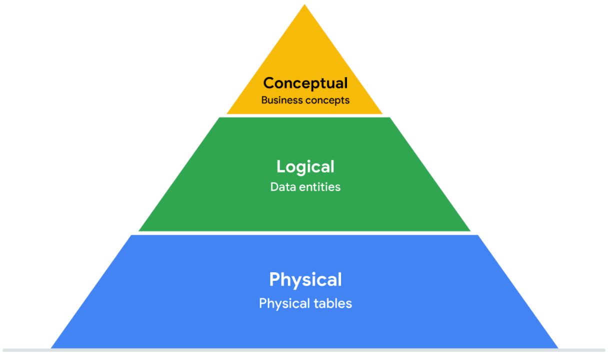
- Recent advancements in machine learning algorithms and AI are beginning to make it easier to search, manage, analyze unstructured data

Data Modeling Levels and Techniques

- Data models keep data consistent and enables people to map out how data is organized, a basic understanding helps data analysts and stakeholders make sense of their data and use it right
- Data modeling is the process of creating diagrams that visually represent how data is organized and structured, visual representations = data modeling
- It's like a blueprint, all the trades use the blueprint and have different relationship to it but they all understand the structure of the building, similarly different users may have different data needs and the data models gives them understanding of the structure as a whole

- Conceptual --> gives high level view of data structure like how it interacts across an organization, conceptual data doesn't include technical detail
- Technical --> details of a database like relationships, attributes, entities, logical data models define how individual records are uniquely identified in a database
- Physical --> how a database operates, physical data models define all entities and attributes used (ex table name, column name, data type)

The three most common types of data modeling



Data modeling techniques

- **Entity Relationship Diagram (ERD)** --> visual ways to understand the relationship between entities in the model
- **Unified Modeling Language (UMF)** --> very detailed diagrams that show the structure of a system by showing the systems entities, attributes, operations and relationships

- Data modeling can help you explore the high level details of your data and how it is related across the organization's information systems, understand how the data is put together so you know how to map the data models make it easier for everyone in your organization to understand and collaborate with you on your data

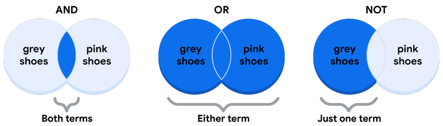
Explore data types, fields, and values

Know the type of data you're working with

- Data type: a specific kind of data attribute that tells what kind of value the data is, tells you what kind of data you're working with
- Data types can be different depending on the query language you're using
- Data types in spreadsheets --> numbers, texts/strings, Boolean
- Text/string data could include numbers (like street address or phone number) but they wouldn't be calculated so they're treated like text
- Boolean data --> two possible values like true/false
- Common issues in spreadsheets = mistaking data types with cell values (ex applying a formula from a set of number data to a set of text/string data and getting an error)

Understanding Boolean logic

- "Conditions" in Boolean statement are created with operators like AND, OR, NOT, Boolean similar to mathematical operators and create logical statements that filter results
- Boolean statements used for creating queries for searches and checking conditions for writing programming code



- AND statement --> "IF (Color = "Grey") AND (Color = "Pink") --- the AND operator lets you stack multiple conditions, Boolean logic will denote TRUE if both conditions are met otherwise it's a FALSE
- OR statement --> "IF (Color = "Grey") OR (Color = "Pink") --- if either condition is met
- NOT statement --> "IF (Color = "Grey") AND (Color = NOT "Pink")
- Power of Boolean logic is ability to combine multiple conditions in a single statement in order to filter your results

Data Table Components

- A data table (tabular data) has a very simple structure, arranged in rows/columns
- Rows --> records; Columns --> fields
- "Records/fields" can be used for any data table; "Rows/columns" usually reserved for spreadsheets
- "Field" is sometimes used to refer to a single piece of data (like value in a cell)
- Applying a function --> without "=" the spreadsheet will interpret the input as a string, for a formula to work it needs numeric data, when a given cell contains a string the program considers the numerical value of the cell to be zero

Meet Wide and Long Data Formats

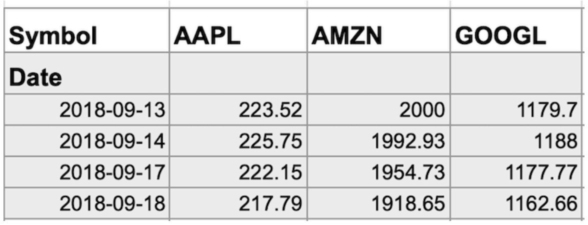
- Wide data --> data in which every data subject has a single row with multiple columns to hold the values of various attributes of a subject
- Long data --> data in which every row is one time point per subject, so each subject will have data in multiple rows
- Long data format is great for storing and organizing data when there's multiple variables for each subject at each time point we want to observe, you can store and analyze all the data using fewer columns; it keeps everything nice and compact

Data Transformation

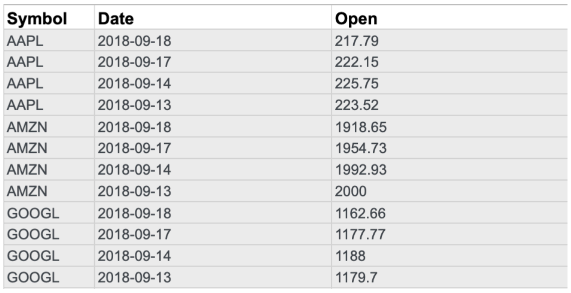
- Data transformation --> changing the data's format, structure, values
  - Adding, copying, replicating data
  - Deleting fields or records
  - Standardizing the names of variables
  - Renaming, moving, combining columns in a database
  - Joining one set of data with another
  - Saving a file as different format (CSV = comma separated values)

- Why transform?
  - Organization: easier to use
  - Compatibility: different applications or systems can use same data
  - Migration: matching formats can move from one system to another
  - Merging: data with same organization can be merged together
  - Enhancement: data can be displayed with more detailed fields
  - Comparison: apples-to-apples comparison can be made

- Long Data --> each row contains a single data point for a particular item
- Wide Data --> each row contains multiple data points for a particular item



Symbol	AAPL	AMZN	GOOGL
Date			
2018-09-13	223.52	2000	1179.7
2018-09-14	225.75	1992.93	1188
2018-09-17	222.15	1954.73	1177.77
2018-09-18	217.79	1918.65	1162.66



Symbol	Date	Open
AAPL	2018-09-18	217.79
AAPL	2018-09-17	222.15
AAPL	2018-09-14	225.75
AAPL	2018-09-13	223.52
AMZN	2018-09-18	1918.65
AMZN	2018-09-17	1954.73
AMZN	2018-09-14	1992.93
AMZN	2018-09-13	2000
GOOGL	2018-09-18	1162.66
GOOGL	2018-09-17	1177.77
GOOGL	2018-09-14	1188
GOOGL	2018-09-13	1179.7

- Long preferred when:
  - Storing lots of variables about each subject (ex 60 years worth of interest rates for each bank)
  - Performing advanced statistical analysis or graphing
- Wide preferred when:
  - Creating tables and charts with few variables about each subject
  - Comparing straightforward line graphs
- "Wide data subjects can have data in multiple columns, long subjects can have multiple rows that hold the values of subject attributes"

W2 Bias, credibility, access, ethics, privacy

Unbiased and Objective Data

- Avoid the conflict, overcome the challenges, and answer the questions
- Even the most sound data can be skewed or misinterpreted
- Good data vs. Bad data
- As data becomes more available we need to ask:
  - Who owns all this data?
  - How much control do we have over the privacy of data?
  - Can we reuse data however we want to?
- Data analysts make a lot of judgement calls on the privacy and ethics of data

Bias: From Questions to Conclusions

- Bias --> a preference in favor of or against a person, group of people, or things (conscious or unconscious)
- Once we know and accept that we have biases, we can start to recognize our own patterns of thinking and learn how to manage it
- Data bias --> a type of error that systematically skews results in a certain direction
- Sample bias, ways questions are worded in a survey
- People with disabilities tend to be under-identified, under-represented, excluded from mainstream health research
- The way you collect data can bias a data set
- Data analysts need to think about bias and fairness from the moment you start collecting data to the time you present your conclusions
- Sampling bias --> when a sample isn't representative of the population as a whole
- Use visualizations to uncover if you're working with biased data, easier to identify misalignment of your sample

Understanding bias in data

- You need all sides of the story to avoid sampling bias
- Observer bias (experimenter bias/researcher bias) --> tendency for different people to observe things differently, two people can observe something and draw different conclusions about it
- Interpretation bias --> tendency to always interpret ambiguous situations in a positive or negative way, two people hear/see the same thing and interpret it in a variety of different ways
- Confirmation bias --> tendency to search for or interpret information in a way that confirms pre-existing beliefs
- No matter what data you use it needs to be inspected for accuracy and trustworthiness

Exploring Data Credibility

Identifying good data sources

- "good" is subjective
- How to find and identify good data sources
- R: reliable
- O: original (validate 2nd and 3rd party data with original sources)
- C: comprehensive (all critical information needed to answer questions)
- C: current (usefulness of data decreases over time)
- C: citing (helps credibility)

- Best sources --> vetted public data sets, academic papers, financial data, government agency data

What is bad data?

- Not reliable --> inaccurate, incomplete, biased (sample selection bias/misleading visualizations)
- Not original --> if you can't locate original data or just relying on 2nd/3rd party data
- No comprehensive --> missing important information to answer question/solution
- Not current --> trusted sources refresh their data regularly so not out of data, irrelevant
- Not citied --> no go

- "Every good solution is found by avoiding bad data"

Data Ethics and Privacy

Introduction to data ethics

- Personal ethics evolve and become more rational overtime
- When analyzing data, we need to rely on more than just our code of ethics to address questions, challenges, opportunities
- Ethics --> 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 --> well founded standards of right and wrong that dictate how data is collected, shared, and used
- Aspects of data ethics: ownership, transaction, transparency, consent, privacy, openness
- Ownership --> who owns data? Individuals who own the raw data they provide and they have primary control over its usage, how its processed and how its shared
- Transactional transparency --> all data processing activities and algorithms should be completely explainable and understood by the individuals who provide their data
- Consent --> an individual's right to know explicit details about how and why their data will be used before agreeing to provide it (why data is collected, use for what, how long will it be stored), so much activity happening online, consent usually just looks like a terms and conditions checkbox, consent is very important to prevent all populations from being unfairly targeted
- Currency --> individuals should be aware of financial transactions resulting from the use of their personal data and the scale of these transactions

- "How do we actually improve the lives of people by using data"
- "Are the people represented in this data going to be benefitted by this"
- "At the end of the day, data are people"
- Thinking about how to keep aspects of their data protected and private
- We need to ensure that there's actionable ways in which people can consent to giving those data and ways they can ask for it to be revoked or removed
- As data is growing, people need ways to have power over their own data

Data Privacy

- Preserving a data subject's information and activity anytime a data transaction occurs
- Sometimes called information privacy or privacy protection
- Protection from unauthorized access to our private data
- Freedom from inappropriate use of our data
- The right to inspect, update, and correct our data
- Ability to give consent to use our data
- Legal right to access the data
- Being able to trust companies with your data is important, what makes a person want to use a company's product and share their information
- Openness --> free access, usage, sharing of data

Data Anonymization

- Personal identifiable information (PII) --> information that can be used by itself or with other data to track down a person's identity
- Data anonymization --> the process of protecting people's private or sensitive data by eliminating that kind of data (blanking, hashing, masking personal information)
- Healthcare and financial data are the two most sensitive types of data
- De-identification --> wipe data clean of PII
- Data that is often anonymized --> telephone numbers, names, license pates, SSN, IP addresses, medical records, email addresses, photographs, account numbers

Understanding Open Data

Features of Open Data

- Openness --> free access, usage, and sharing of data
- Availability and access --> open data must be available as a whole preferably by downloading in convenient, modifiable form ex data.gov
- Reuse and redistribution --> open data must be provided under terms that allow reuse and redistribution
- Universal participation --> no discrimination against fields, persons or groups; no one can mask restrictions on the data like only making it available for specific industries

Benefits of open data

- Credible databases can be used more widely
- All that good data can be leveraged, shared, or combined with other data
- A whole lot of resources are needed to make the technological shift to big data
- Interoperability --> the ability of data systems and services to openly connect and share data, compatible databases that allow information to be shared, this depends on openness

The Open-data debate

- 3rd party data --> 3rd parties might collect information about visitors from a website and leverage that info
- Because 3rd party data is readably available, its important to balance the openness of data with the privacy of individuals

Steps for ethical data: what data analysis can do to make sure they're looking at data from an ethical lens

- Self reflection and understanding impact of what you're doing
- Question who we are, this team working on a data set
- Continue to question the integrity, quality, representation that is present in the dataset
- What is potential harm or risk of holding onto the dataset if you continue to use it?
- What's the communication channel like? Are you informing those who you collected data from?

"As a data analyst, you stand in the intersection between the very people that will stand to benefit from the technology that's being developed and those in your organization that are trying to make more informed decision."

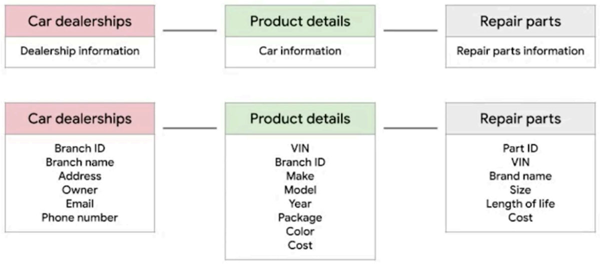
W3 All about databases

Databases: where data lives

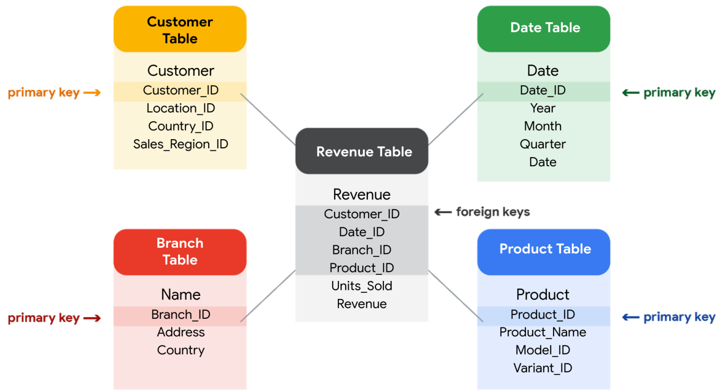
- Database: collection of data stored in a computer system
- How databases make it possible to find the exact piece of information for your analysis
- "Meta" -->if a character in a movie knows she's in a movie, or analyzing *how* you analyze data (quality check)
- Sometimes we get a little too close to our data, so it's important to step back and ask ourselves if our data processes make sense
- Metadata --> data about data, think of it like a reference guide, without it you have a bunch of data with no context explaining it, tells you where it comes from when it was created and what it's all about

Database features

- Just about all the data you access is stored within databases, make it easy for analysts to store and access information, faster insights data driven decisions and solve problems
- Relational Database --> a database that contains a series of related tables that can be connected via their relationships; for two tables to have a relationship, two or more fields must exist in both tables, that's what links the tables together (ex the VIN connecting Product details with Repair parts), allows analysts to organize data based on what's in common, tables in a relational database are connected by the fields they have in common



- Primary key --> an identifier that references a column in which each value is unique
- Unique identifier for each row in a table (Branch ID in Car Dealerships)
- If you include a primary key it should be unique meaning no two rows should have the same primary key, cannot be null (blank)
- If VIN is the primary key in Product details no two customers can have the same VIN value
- *Used to ensure data in a specific column is unique*
- *Uniquely identifies a record in a relational database table*
- *Only one primary key is allowed in a table*
- *Cannot contain null or blank values*
- Foreign key --> a field within a table that is a primary key in another table (how one table can be connected with another)
- In Repair parts the part ID is primary and all rows under it are foreign
- *A column or group of columns in a relational database table that provides a link between the data in two tables*
- *Refers to the field in a table that's the primary key of another table*
- *More than one foreign key is allowed to exist in a table*
- *These keys are what creates the relationship between tables in a relational database, which helps organize and connect data across multiple tables in a database*
- Some tables don’t require a primary key (Revenue Table has no primary key but multiple foreign keys
- Composite key --> a primary key may be constructed using multiple columns in a table (customer id and location id are two columns of a composite key, the values assigned to that row are unique in the table)
- Databases use SQL language to communicate, let's analysts communicate with the databases, in relational databases analysts can write queries to get data from the related tables



Managing data with metadata

- Metadata is the information that's used to describe the data that's contained in something, metadata is not the data itself, its data *about* the data
- Metadata is used in database management to help data analysts interpret the contents of the data within a database
- 3 common types of metadata: descriptive, structural, administrative
- Descriptive --> metadata that describes a piece of data and can be used to identify it at a later point in time (ISPN of a book)
- Structural -->metadata that indicates how a piece of data is organized, and if it's part of one, or more than one, data collection (how pages of a book are organized to create chapters), keeps track of the relationship between two things
- Administrative -> metadata that indicates the technical source of the digital asset (details of a digital photo showing type of file, when it was taken)

Metadata is as important as the data itself

- Metadata tells us the who, what, when, where, how, which, why
- Elements of metadata --> title/description, tags/categories, who created it/when, who last modified it/when, who can access it/update it
- Examples --> photos, emails, spreadsheets, websites, digital files, books
- Metadata is important to understanding full picture of your data, not just your data you are viewing but how the data comes together
- Ensures you are able to find, use, reuse, preserve for the future
- Your responsibility to manage data in its entirety

Why data analysts use metadata

- Putting data into context is the most valuable thing metadata does
- Additionally, metadata creates a single source of truth by keeping things consistent and uniform, uniform and consistency = great
- Uniform data can be organized, stored, classified, used effectively
- When data is uniform it's so much easy to discover relationships between the data inside it and data elsewhere, standardizing our process of retrieving data
- More reliable by making sure data is accurate, precise, timely, relevant (for finding root causes of problems that pop up)
- Metadata repository --> database specifically created to store metadata (physical or virtual location), describes where metadata came from, comes in an accessible form and common structure; meta repositories make it easier and faster to bring together multiple sources of data together for analysis
- Metadata repositories --> describe the state and location of the metadata, describe the structures of the tables inside, describe how the data flows through the repository, keep track of who accesses metadata and when, used to create a single source of truth
- When working with data generated outside your organization, important to understand the metadata of the external database, tell you how good the data is
- Important step when using external data is confirming we are allowed to use it, access or purchase

Metadata Management

- Components of metadata and how metadata analysts work to keep things organized
- Companies can have an overwhelming amount of data spanning across numerous processes and systems, pulling data together from so many places can be a big challenge, if the data they collect expands with cloud storage and 2nd/3rd party data with different rules regulations and organizes data in a different way
- Metadata is stored in a single, central location, and gives the company standardized information about all of its data
  - ○ (1) metadata included information about where each system is located and where data sets are located within those systems
  - ○ (2) metadata describes how the data is connected through the various systems
- Data governance --> a process to ensure the formal management of a company's data assets; gives company better control of their data, manage issues related to privacy or security, internal external data flows
- Metadata specialists --> organize and maintain company data, making sure its of the highest quality; they create basic metadata identification and discovery information, describe the way different data sets work together, explain different types of data resources; create standards and models used to organize data
- Metadata is the key to your larger data set, describes what in the rows/columns of the data, kind of a cliff notes version of a more complicated data set
- Important for understanding the resources you have to address a problem or what's missing

Accessing different data sources

- Internal (primary) vs. external (secondary) data
- Collecting internal data can be complicated, collecting from different sources or departments within a company; however internal data is worth it because it's relevant and free to access, sometimes internal data doesn't give the full picture
- Openness movement of data, sharing industry level perspectives

Importing data from spreadsheets and databases

- CSV file --> comma separated values, saves file in table format, spreadsheet will automatically detect the format but sometime you need to indicate that the separator is a different character; a delineator indicates a boundary or separation between two things

Sorting and Filtering Data

- Only focusing on the data relevant to the problem you are trying to solve, useful when dealing with complex spreadsheet
- Sorting and filtering data helps us customize the way data is presented, organize data so analysists can zoom in on the data that matters
- Sorting data --> arranging data into a meaningful order to make it easier to understand, analyze and visualize
- Sorting by multiple variables
- Multiple criteria sorting, make sure "data has header row" is highlighted
- Filtering --> only showing the data that meets a specific criteria while hiding the rest, simplifies a spreadsheet by showing us only the data we need to see
- Using BigQuery
    - BigQuery is a data warehouse on Google Cloud that allows data analysts to query, filter data results, aggregate results, and perform complex operations

W4 Effectively organize your data

Organizing and protecting your data

Feel confident about your data

- Certain procedures to follow to make sure your data is organized and easy to use
- Best practices for organizing data --> naming conventions, foldering, archiving older files, align your naming and storage practices with your team , develop metadata practices
- Naming conventions --> consistent guidelines that describe the content, date, or version of a file in its name; use logical and descriptive names for your files to make them easier to find and use
- Foldering --> organizing all files into folders, and break folders down into subfolders
- Archiving --> move old projects to a separate location to create an archive and cut down on clutter
- Think about how often you're making copies of data and storing it in different places, if data is stored in lots of different spreadsheets and databases it can contradict itself and create problems
- Storing multiple copies of data can take up a lot of space, relational databases help avoid data duplication and stores data for efficiently
- It's a good idea to take time early on in a project to consider what the best organizational methods will be for you and your team to stick to

Best practices for file naming conventions

Review the following file naming recommendations:

- Work out and agree on file naming conventions early on in a project to avoid renaming files again and again.
- Align your file naming with your team's or company's existing file-naming conventions.
- Ensure that your file names are meaningful; consider including information like project name and anything else that will help you quickly identify (and use) the file for the right purpose.
- Include the date and version number in file names; common formats are YYYYMMDD for dates and v## for versions (or revisions).
- Create a text file as a sample file with content that describes (breaks down) the file naming convention and a file name that applies it.
- Avoid spaces and special characters in file names. Instead, use dashes, underscores, or capital letters. Spaces and special characters can cause errors in some applications.

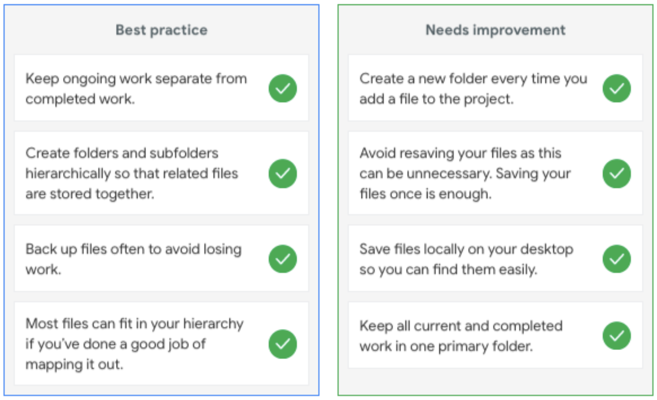
Best practices for keeping files organized

Remember these tips for staying organized as you work with files:

- Create folders and subfolders in a logical hierarchy so related files are stored together.
- Separate ongoing from completed work so your current project files are easier to find. Archive older files in a separate folder, or in an external storage location.
- If your files aren't automatically backed up, manually back them up often to avoid losing important work

File naming conventions

- File naming conventions --> consistent guidelines that describe the content, data, or version of a file in its name
- File naming DO's
  - Work out your conventions early
  - Align file naming with your team
  - Make sure file names are meaningful
  - Keep file name short and sweet
  - Format dates yyymmdd: SalesReport20201125 (follows international standard)
  - Lead revision numbers with 0: SalesReport20201125v02
  - Use hyphens, underscores, or capitalized letters: SalesReport_2020_11_25_v02
  - Create a text file that lays out all your naming conventions on a project



Best practice		Needs improvement	
Keep ongoing work separate from completed work.	✓	Create a new folder every time you add a file to the project.	✓
Create folders and subfolders hierarchically so that related files are stored together.	✓	Avoid resaving your files as this can be unnecessary. Saving your files once is enough.	✓
Back up files often to avoid losing work.	✓	Save files locally on your desktop so you can find them easily.	✓
Most files can fit in your hierarchy if you've done a good job of mapping it out.	✓	Keep all current and completed work in one primary folder.	✓

Securing Data

- Security features came help unauthorized people from viewing certain files, could also be how you lock your worksheets so you don't accidentally break your formulas
- Data security --> protecting data from unauthorized access or corruption by adopting safety measures
- Features to protect spreadsheets from being edited (from whole workbook to single cells)
- For Excel you can encrypt files with passwords

The battle between security and data analytics

- Data security means protecting data from unauthorized access or corruption by putting safety measures in place.
- Usually the purpose of data security is to keep unauthorized users from accessing or viewing sensitive data.
- Data analysts have to find a way to balance data security with their actual analysis needs.
- This can be tricky-- we want to keep our data safe and secure, but we also want to use it as soon as possible so that we can make meaningful and timely observations.
- In order to do this, companies need to find ways to balance their data security measures with their data access needs
- Luckily, there are a few security measures that can help companies do just that.
- The two we will talk about here are encryption and tokenization.
- Encryption uses a unique algorithm to alter data and make it unusable by users and applications that don't know the algorithm.
- This algorithm is saved as a "key" which can be used to reverse the encryption; so if you have the key, you can still use the data in its original form.
- Tokenization replaces the data elements you want to protect with randomly generated data referred to as a "token."
- The original data is stored in a separate location and mapped to the tokens.
- To access the complete original data, the user or application needs to have permission to use the tokenized data and the token mapping.
- This means that even if the tokenized data is hacked, the original data is still safe and secure in a separate location.
- Encryption and tokenization are just some of the data security options out there.
- There are a lot of others, like using authentication devices for AI technology.
- As a junior data analyst, you probably won't be responsible for building out these systems.
- A lot of companies have entire teams dedicated to data security or hire third party companies that specialize in data security to create these systems.
- But it is important to know that all companies have a responsibility to keep their data secure, and to understand some of the potential systems your future employer might use