Data collection

Data sources

1st-party data: Data you collect yourself

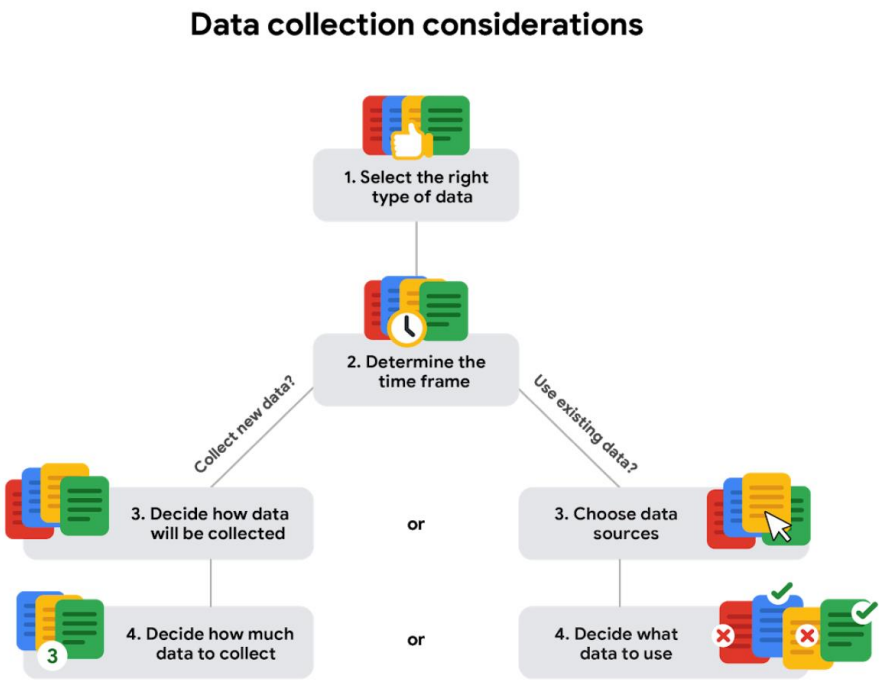
2nd-party data: is collected directly by another group and then sold.

3rd party data: from a number of different sources, but sold by a provider not collected it themselves

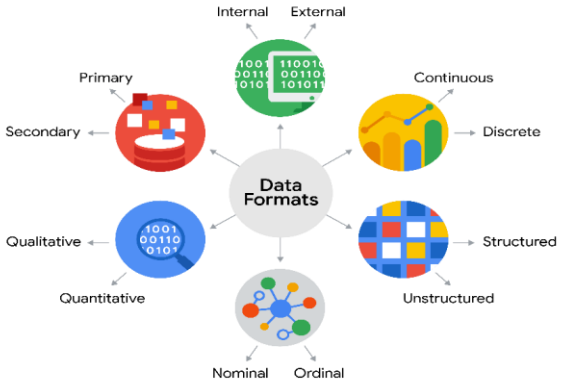
What to collect: Only data help you with your questions

How much data to collect: Each project has its own needs.

Time-frame: Use the flowchart below if data collection relies heavily on how much time you have (on the project)



Data formats in practice



Primary and 2nd data

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 and external data

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 and discrete data

Data Format Classification	Definition	Examples
Continuous data	Data that is measured and can take on any value within a certain range.	- Height of kids in third-grade classes (52.5 inches, 65.7 inches) - Runtime markers in a video - Temperature
Discrete data	Data consists of whole, concrete numbers with **specific and fixed data values** determined by counting.	- 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 and quantitative data

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 and ordinal data

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 - Yes/No questions
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 and Unstructured data

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

- 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

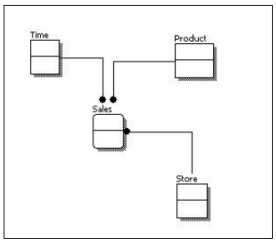
Data modelling level and techniques

Commented [2HL1]: As a junior data analyst, you won't be asked to design a data model. But you might come across existing data models your organization already has in place.

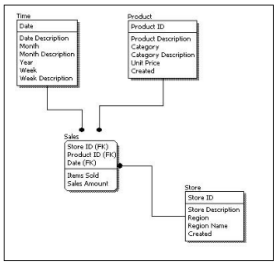
Data modeling is the process of creating diagrams that visually represent how data is organized and structured.

Levels of data modelling:

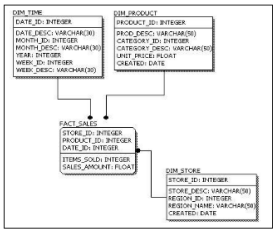
Conceptual Model Design



Logical Model Design



Physical Model Design

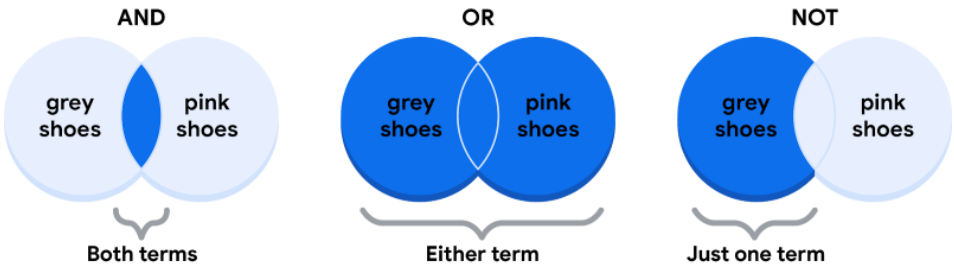


- **Conceptual data modeling**: gives a high-level view of the data structure, such as how data interacts across an organization. A conceptual data model doesn't contain technical details.
- **Logical data modeling** focuses on the technical details of a database such as relationships, attributes, and entities. For example, a logical data model defines how individual records are uniquely identified in a database. But it doesn't spell out actual names of database tables. That's the job of a physical data model.
- **Physical data modeling** depicts how a database operates. A physical data model defines all entities and attributes used; for example, it includes table names, column names, and data types for the database.

Data types and Boolean logic

Data types in spreadsheet: number, text, boolean (data just have 2 values (eg: True; False))

Boolean logic



The AND operator

Condition: “If the color of the shoe has any combination of grey and pink, you will buy them.”

⇨ “IF (Color=”Grey”) AND (Color=”Pink”) then Buy

The OR operator

“If the shoes are grey or pink, you will buy them.”

⇨ “IF (Color=”Grey”) OR (Color=”Pink”) then buy

The NOT operator

"You will buy any grey shoe except for those with any traces of pink in them."

⇨ “IF (Color="Grey") AND (Color=NOT “Pink”) then buy them.

Note: When discussing structured databases, data analysts refer to the data contained in a **row** as a **record**, and that in a **column** as a **field**.

Wide data and Long data

Wide data: Data in which every data subject has a single row with multiple columns. Each row contains multiple data points for the particular items identified in the columns.

Eg: Stock prices

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

Long data: Data in which every data subject has multiple rows with 01 column. Each row contains a single data point for a particular item.

Eg: Stock prices

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

Wide data is easier to read and understand.

Wide data is preferred when	Long data is preferred when
Creating tables and charts with a few variables about each subject	Storing a lot of variables about each subject. For example, 60 years worth of interest rates for each bank

Comparing straightforward line graphs	Performing advanced statistical analysis or graphing

Unbiased and objective data

Bias data can lead to skewed answers. To avoid bias, in the collect data phrase, we already need to ensure fairness by collecting an unbiased sample, which is representative of the whole population. (one way is to make sure the pie chart of the sample looks familiar with the population pie chart)

Types of data bias

- Sampling bias
- Observer bias:
- Interpretion bias: The tendency to always interpret ambiguos situations in a positive/ negative way
- Confirmation bias: The tendency to search for/ interpret information in a way that confirms pre-existing beliefs

Good data sources

Reliable
Original: Collect data from its original sources
Comprehensive: answer all your vital questions
Current:
Cited: Who created this data?

Bad data sources

NOT Reliable (bias, misleading)
NOT Original: comes from 2, or third party that you are not sure
NOT C
NOT Current
NOT Cited

- ⇨ Should avoid bad data sets
- ⇨ For good data, stick with vetted public data sets, academic papers, financial data and governmental agency data.

Data ethics

Well- founded standards for data. At the end of the day, we should ask: How we can improve the people who gives the data?

Aspects of data ethics

- Ownership: Who owes the data? Individual providing the data
- Transaction transparency: All data-processing activities and algorithms should be completely explainable and understood by individuals providing the data -> they can judge the fairness of the data and raise concern if needed
- Consent: The owner'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 the scale of these transactions.

- Privacy:
- Open:

Self-reflect and understand what it is that you're doing and the impact that it has.

Data anonymization

Data anonymization is the process of protecting people's private or sensitive data by eliminating *Personally identifiable information*

Types of data should be anonymized: Healthcare and financial data, or personal data (telephone numbers, name, etc.)

⇨ Need de-identification, which is a process used to wipe data clean of all personally identifying information.

Open data

Open data is part of data ethics, refers to free access, usage, and sharing of data. But for data to be considered open, it has to:

- 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

The third-party data can be publicly available. However, Personal identifiable information (PII) should be kept safe.

Sites and resources for open data

1. **U.S. government data site**: Data.gov is one of the most comprehensive data sources in the US. This resource gives users the data and tools that they need to do research and even helps them develop web and mobile applications and design data visualizations.
2. **U.S. Census Bureau**: This open data source offers demographic information from federal, state, and local governments, and commercial entities in the U.S. too.
3. **Open Data Network**: This data source has a really powerful search engine and advanced filters. Here, you can find data on topics like finance, public safety, infrastructure, and housing and development.
4. **Google Cloud Public Datasets**: There are a selection of public datasets available through the Google Cloud Public Dataset Program that you can find already loaded into BigQuery.
5. **Dataset Search:** The Dataset Search is a search engine designed specifically for data sets; you can use this to search for specific data sets.
6. Kaggle has an Open Data search function that can help you find datasets to practice with.
7. Finally, BigQuery hosts 150+ public datasets you can access and use.

Public health datasets

1. Global Health Observatory data: You can search for datasets from this page or explore featured data collections from the World Health Organization.
2. The Cancer Imaging Archive (TCIA) dataset: Just like the earlier dataset, this data is hosted by the Google Cloud Public Datasets and can be uploaded to BigQuery.

3. 1000 Genomes: This is another dataset from the Google Cloud Public resources that can be uploaded to BigQuery.

Public climate datasets

1. National Climatic Data Center: The NCDC Quick Links page has a selection of datasets you can explore.
2. NOAA Public Dataset Gallery: The NOAA Public Dataset Gallery contains a searchable collection of public datasets.

Public social-political datasets

1. UNICEF State of the World's Children: This dataset from UNICEF includes a collection of tables that can be downloaded.
2. CPS Labor Force Statistics: This page contains links to several available datasets that you can explore.
3. The Stanford Open Policing Project: This dataset can be downloaded as a .CSV file for your own use.

Managing Data

Databases

A **relational database** is a database that contains a series of tables that can be connected to show relationships.

- ⇨ Organizing data in a relational database: **Normalization**

The key to relational databases

A primary key: is a column of a table that is used to uniquely identify each record within that table

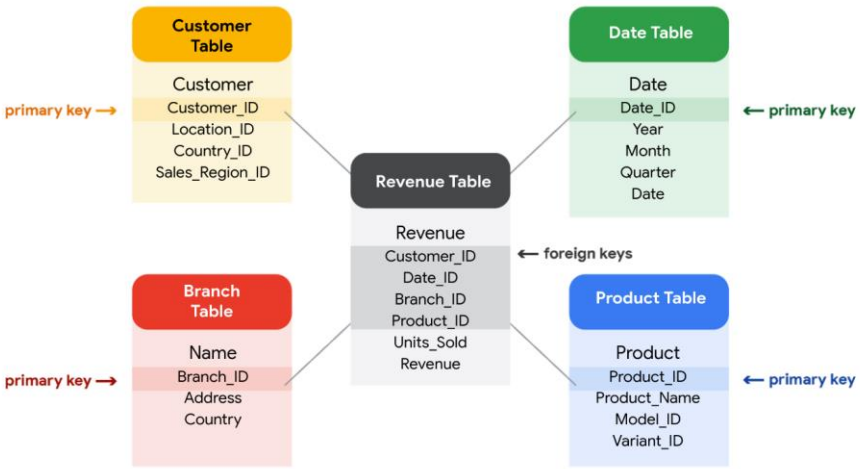
- ⇨ The value assigned to the primary key in a particular row must be unique within the entire table.
- ⇨ There is only one primary key in the table

Some tables don't require a primary key. A primary key may also be constructed using multiple columns of a table (**composite key**).

A foreign key is a field within a table that is a primary key in another table.

- ⇨ A table can have multiple foreign keys.

Commented [2HL2]: For example, if customer_id and location_id are two columns of a composite key for a customer table, the values assigned to those fields in any given row must be unique within the entire table.



Inspecting a dataset: A guided, hands-on tour

Before you begin an analysis, it's important to inspect your data to determine if it contains the specific information you need to answer your stakeholders' questions. In any given dataset, it may be the case that:

- The data is not there (you have sandwich data, but you need pizza data) => ask
- The data is insufficient (you have pizza data for June 1-7, but you need data for the entire month of June) => Ask for clarification, do not assume
- The data is incorrect (your pizza data lists the cost of a slice as $250, which makes you question the validity of the dataset)

⇨ Your next step: You may be able to recover this data from an external source or at least recommend to your stakeholders that another data source be used.

The scenario

⇨ Having the answer, remember to answer "so what?" questions

Metadata

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

Elements of metadata

- Title and description
  What is the name of the file or website you are examining? What type of content does it contain?
- Tags and categories

What is the general overview of the data that you have? Is the data indexed or described in a specific way?

- Who created it and when
  Where did the data come from, and when was it created? Is it recent, or has it existed for a long time?
- Who last modified it and when
  Were any changes made to the data? If yes, were the modifications recent?
- Who can access or update it
  Is this dataset public? Are special permissions needed to customize or modify the dataset?

Types of metadata

- **Descriptive metadata** describes a piece of data or can be used to identify it at any time.
- **Structural metadata** indicates exactly how many collections data live in. It provides information about how a piece of data is organized and whether it's part of one, or more than one, data collection.
- **Administrative metadata** indicates the technical source and details for a digital asset.

A metadata repository is a database specifically created to store metadata.

- Describe the state and location of metadata
- Describe the structures of the tables inside
- Describe how data flows through the repository
- Keep track of who accesses the metadata and when

⇨ A source to ensure trustfulness of the data: it is ROCCC?

⇨ Metadata is stored in a single, central location and it gives the company standardized information about all of its data, so the company can find the right data at the right time

Accessing different data sources

Import external sources to a spreadsheet

Google Sheets: IMPORTRANGE function

Excel: Data/Get data/ From File/ From Excel Workbook/ Import. Load: import all, transform: a part…

Importing data from CSV files

Google Sheets: File/ Import

Excel: Data/Get data/ From File/ From Text/CSV

Importing HTML tables from web pages

Google Sheets: IMPORTHTML function

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

Excel: Data/Get data/ From Web

Databases and spreadsheets for sorting and filtering

Question	Spreadsheet	Database
How do they store data	In cells	In tables
How are they use to interact with data	To analyze ?	To show relationships
How powerful is each		
Their pros and cons when sorting		
Their pros and cons when filtering		

BigQuery