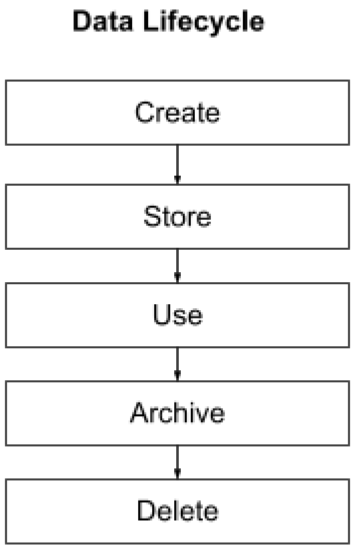
Data Lifecycle Stages

It’s good to think about data in the context of the various “lifecycles” that it exists in. From the Data Analyst’s point of view, there are of two lifecycles:

- The Data Lifecycle: the journey of the data itself.
- The Data Analysis Lifecycle: the journey of the data through analysis activities.

The Data Lifecycle

This is the journey of the data from ‘birth to death’. Consider the following five stages:



These are explained in more detail below.

Create

You can create data in different ways, including:

- Typing it into an online form (e.g. filling out an application).
- Capturing it from sensors (e.g. a temperature sensors in a greenhouse).
- Capturing it from an interaction a person has with a machine (e.g. scanning a travel pass).
- Acquiring existing data (e.g. provided by another company).

Store

Generally, you should store data created for a purpose where it can fulfil that purpose. For example, you can store data:

- On a hard disk on a server.
- On a laptop.
- On a USB drive.

You also need to think of the format in which you store data, for example:

- In a database (e.g a database behind a website front end).
- In a file (e.g. a spreadsheet).

Use

There is a vast range of ways that you can use data, depending on its purpose. For example:

- To report on a company's financial performance.
- To control some industrial machinery.
- To assess someone's medical condition.

Archive

Current methods of data storage may require fast or expensive storage media. So, rather than just deleting it, you may move it to cheaper (and usually slower) storage media when it has served its purpose.   That way, you can retrieve the data if, for some reason, you need to reference it. For example:

- Data about ex-customers is removed from a website's database but stored in an offline archive (for regulatory reasons).
- Data for an ad-hoc analytics process is archived to optical storage once the analysis is complete.
- A company archives data containing company financials older than five years as they no longer need to report on it.

Delete

At some point, you may decide that you can permanently delete data that has served its purpose.   For example:

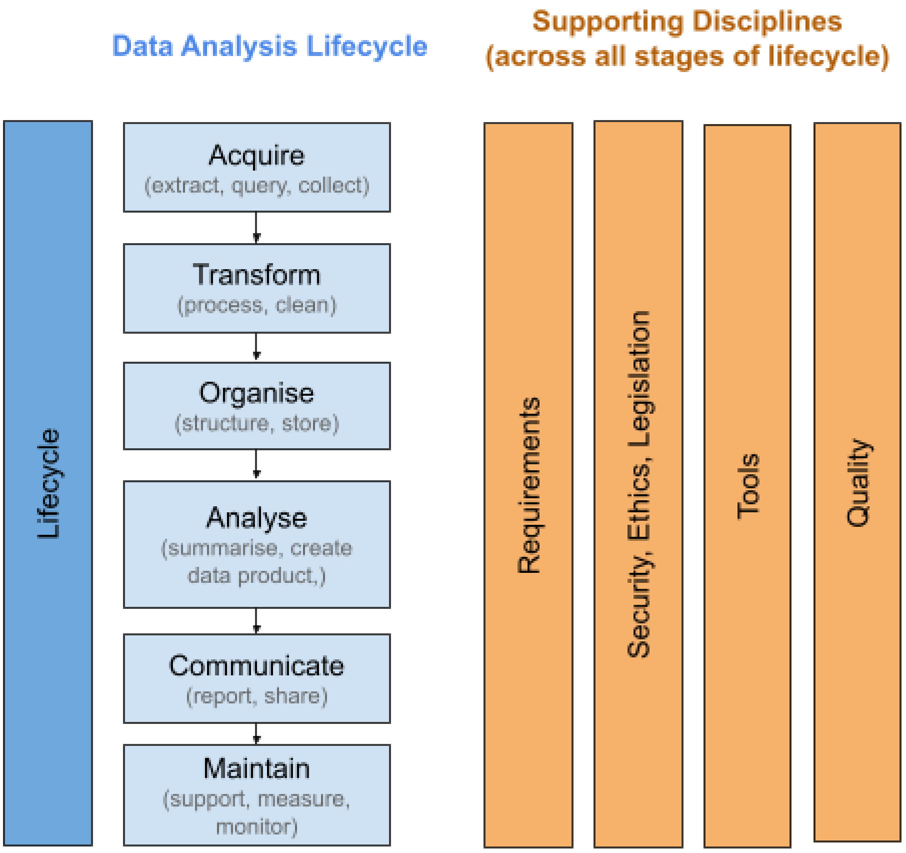
- Delete data from sensors on a decommissioned machine.
- Destroy backups of a company's servers older than 10 years.
- A customer requests that you destroy their data.

The Data Analysis Lifecycle

When thinking specifically about data analysis activities, consider the lifecycle of the data through those activities. This is referred to as the data analysis lifecycle. Because it is about processing the data, it's also often called the data analysis pipeline (think of data flowing through a pipe from one process to another, each doing something to that data).

These activities would fall under the use block in the data lifecycle above, or possibly across the create, store and use blocks if the data acquisition and storage are part of the data analysis project.

Take a look at the following diagram:



The activities in the data analysis lifecycle are in blue. Think of the orange blocks as supporting disciplines that don't apply to one activity but need to be considered in all.

Let's examine each one.

Acquire

You acquire the data for analysis in some way. (E.g. from a web API, queried from a database or collected from an online form.)

Transform

When you acquire data, it is rarely in a form that you can immediately use for analysis. (E.g. you may need to perform some processing or cleaning.).

Organise

Data should be organised in a way that facilitates the analysis. It may require some thought about how it is structured and stored.

Analyse

You can perform a wide variety of analyses on the data. Some may be simple summaries, but sometimes you need a more complex analysis to create some "data product". (E.g. the data product could be a predictive model, a dashboard, etc.)

Communicate

An analysis is of no value unless you communicate the results. This may be a formal or informal sharing of the results, a detailed report, etc.

Maintain

Once you create a data product, it is not the end of the story. The data product may need ongoing support, including measuring its effectiveness, monitoring its usage, etc.

Supporting Disciplines

For each of the above stages of the data analysis lifecycle, you must consider several supporting disciplines. Here are some examples of the questions you need to ask for each stage of the lifecycle against the supporting disciplines:

	Requirements	Security, Ethics, Legislation	Tools	Quality
Acquire	What data do you need to satisfy the requirements? *(E.g. data about page hits when analysing web traffic.)*	Is there a legal, ethical or security reason why we should not use the data? *(E.g. using patient data acquired from a hospital to send marketing emails.)*	What tools do we need to use to get hold of the data? *(E.g. data stored in a database would need to be queried and extracted.)*	Is the data good enough for our purpose? *(E.g. is there any missing data?)*
Transform	How should we process the data to make it suitable for our needs? *(E.g. transforming temperatures from fahrenheit to celsius.)*	Are we allowed to process and derive new information from this data? *E.g. restricting survey data from commercial exploitation.)*	What tools do we need to transform the data? _*E.g. data in Excel needs to be loaded and processed by a program that can read them.)_	How can we make sure we don’t lose quality as we transform the data? *(E.g. losing significant data when rounding some numbers.)*
Organise	Where should we store data? *(E.g. storing company website data where it can be accessed.)*_	Are there restrictions about where data can be stored? *(E.g. can we legally store it in a cloud server in a different country?)*	What storage tools are needed? *(E.g. a database.)*	How do we ensure data integrity? *(E.g. ensuring that bank account transactions are linked and stored against the correct account.)*
Analyse	What are the analysis requirements? *(E.g. an airline*	Are there restrictions about the sorts of analysis?	What are the appropriate tools for carrying out the analysis?	How can we ensure the quality of our analysis? *(E.g. you*

	engine manufacturer requires a sensor data analysis to assess fuel efficiency.)	*(E.g. it would be unethical to profile customers based on race and gender.)*	*(E.g. an end-user may prefer a visual tool, whereas a programmer prefers to write code.)*	*analyse effectiveness of a drug, but is the result statistically significant?)*
Communicate	How should we communicate the result to the intended audience? *(E.g. a daily sales report emailed to a CEO.)*	Are there any security implications of how the information is communicated? *(E.g. a list of MPs' mobile phone numbers posted on an unencrypted USB drive.)*	What tools should we use to communicate? *(E.g. a dashboard)*	How can we ensure the analysis is well-communicated? *(E.g. a poorly presented chart may fail to get an important message across.)*
Maintain	How can we repeat the analysis? *(E.g. a company-produced list of customers for a targeted marketing campaign wants to regenerate a new list every 6 months using the same techniques.)*	How can we support the system in a secure manner? *(E.g. a third-party provides ongoing HR database management. \ow does the company protect their data?)*	What tools should we use to provide ongoing monitoring? *(E.g. how to send alerts of potential fraudulent credit card transactions.)*	How can we ensure that quality does not degrade over time? *(E.g. an analysis of disposable income becomes inaccurate because of a financial downturn.)*