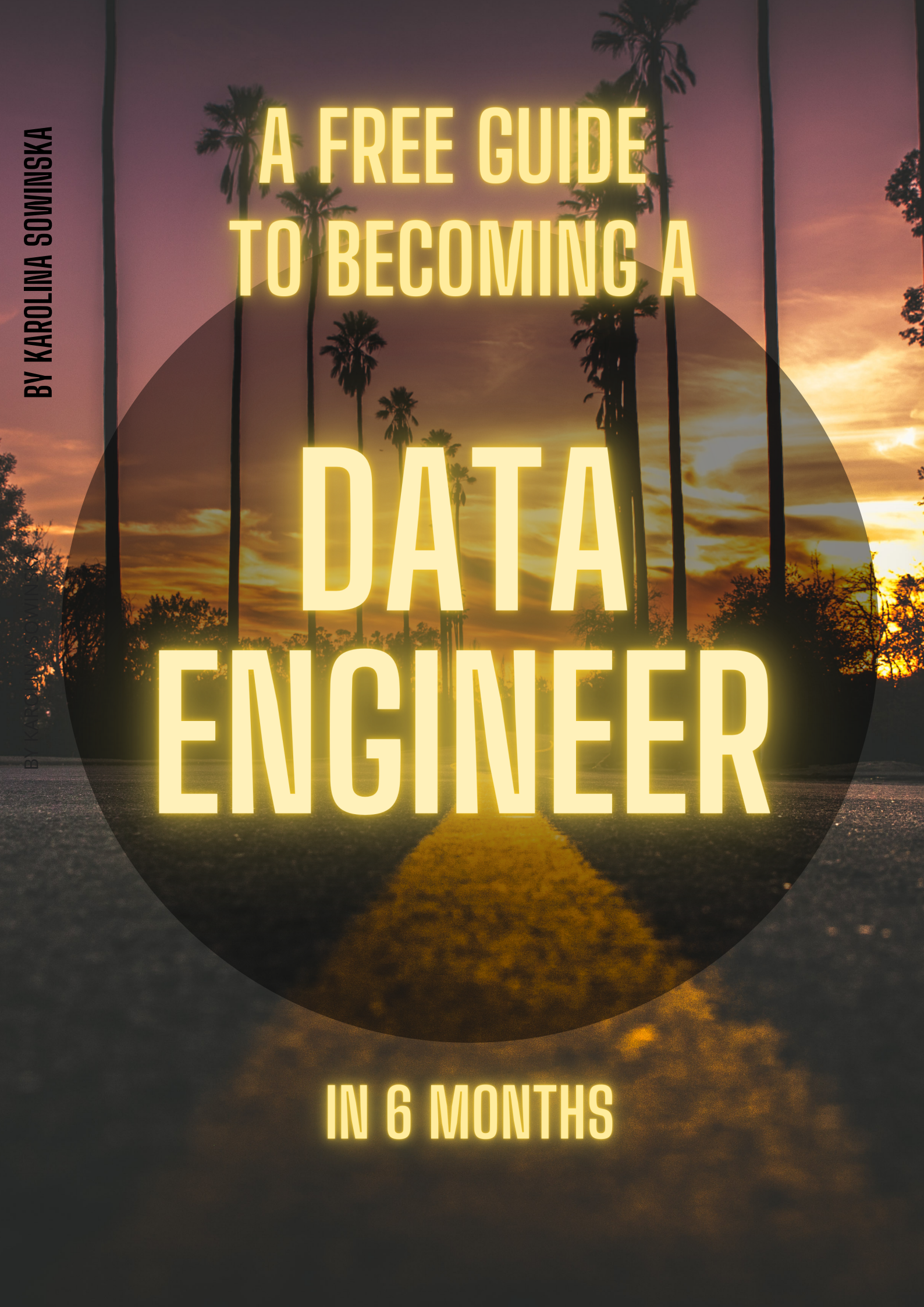

BY KAROLINA SOWINSKA
A FREE GUIDE
TO BECOMING A
DATA
ENGINEER
IN 6 MONTHS

WHY BECOME A DATA ENGINEER?

Data Engineering has overtaken Data Science as the sexiest job in the tech industry.

It's a fact. According to the 2021 Data Science Interview Report the number of data engineering interviews have grown by **40%** in the past year, far exceeding the **10%** growth in Data Science interviews. The increasing demand for Data Engineers is also reflected in their very high salaries, as they range between **$110,000-$155,000** according to Glassdoor.

But that's not all. Most importantly, there's currently a shortage of data engineers.

For every open position, there are only **2.53** applicants on average. To put that into perspective, the competition is almost twice as tough for Data Science positions, with **4.76** applicants per position. That makes data engineering the safe bet - if you learn the required skills, you're almost guaranteed to find a job in this field.

WHAT DO DATA ENGINEERS DO?

Some people say that data is the new oil, as it is every organisation's most valuable resource.

However, before a company can tap into it, layers of foundational work need to be done first. Bluntly speaking - raw data is a mess. **Data engineers' role is to clean it up and put it in order.**

In practice, that might involve:
-**designing and maintaining data warehouses** (e.g. dealing with decisions such as which cloud platform is the best based on the business goals, identifying security risks, choosing the right type of a database)
-**building data pipelines** for (Extract, Transform, Load processes which onboard data to your organisation from a data vendor e.g. via an API or ftp server)
-**monitoring stability of the pipelines** (as they might break on any given day due to a number of reasons, e.g. a vendor changing data format or perhaps limited disk space on your server)
-**setting up data-management tools** (e.g. Tableau, Amazon Web Services) for business users
-**contributing to data governance** and ensuring data quality

DATA ENGINEERING CASE STUDY

Stakeholder:
"I'd like to have access to daily weather data in the US and UK. I found this data provider called SunnyData, and I think it's got what we need. We're on a trial with them, but we're potentially happy to buy the full license. Ideally, I need that data since 2015"

Data engineer:
Your first task as a data engineer is to figure out to access the data. Is it provided via an **API** or perhaps via an **ftp server?** You're most likely going to find the answer to that in the data vendor's documentation.
Next, you download a sample of the data to inspect it. You might want to put it into a database, and confirm with the stakeholder what fields are needed (define acceptance criteria). This is also called a **static load.** Make sure that the historical data is available since 2015, and that it doesn't contain plenty of missing values or duplicates. Does it have a good coverage for the specified regions?
If yes, then great - now it's time to build a **data pipeline.** Where do we need to store the data? In an on-prem database or in a cloud storage? What's the size of the data, and how much is it going to grow if we're going to download it every day?

The choice between an **on-prem vs cloud storage** will be dependent on a number of factors. Typically, if the data size is gigantic, it might work out to be cheaper to store it on internal servers in the long run. On the other hand, if we need flexibility and little upfront costs, cloud should be our way to go. In this case we're deciding to go with **Amazon AWS.** However, choosing a cloud storage option also isn't straightforward - do we pick Amazon EBS, EFS or S3? Each options has its unique benefits (which our further explored in the linked article), but let's pick S3 for our SunnyData case.

Having picked a storage for our data, we now need to define the tables where we're going to load the data, in other words we're defining the **table schema**. In the case of Amazon AWS, this can be done in Amazon Athena. We're specifying column names and column datatypes.

After we've defined the tables, we are ready to write Python scripts that will download the data from the data vendor, save them in a raw form in our **data lake,** validate them (e.g. by checking if we received the exact same column names and data types that we defined in our database), make any necessary transformations, and finally load the data to our defined tables.

Fantastic! However, the stakeholder wanted to have access to data since 2015. That's why the next step is **backfilling** - defining a function that will download historical data and load it to our databases.

Lastly, we schedule the data pipeline in **Airflow.** We might also use **Jenkins,** or some other job scheduling platform.

We're done! Hopefully the pipeline won't break much... We're going to have to **keep an eye on it!**

MY GUIDANCE

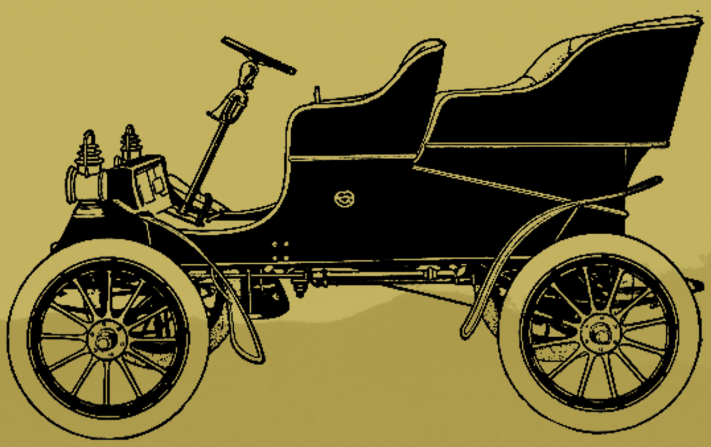
A lot of people from my YouTube community have asked me for guidance in finding the first Data Engineering job.

I know how overwhelming it can be to start learning a new skill. If self-doubt creeps in, you might end up procrastinating, and giving up on your dreams. That's why I'd like to share a few of my favourite quotes with you first:

HENRY FORD

WHETHER YOU THINK YOU CAN, OR YOU THINK YOU CAN'T...

YOU'RE RIGHT.



SENECA

LUCK HAPPENS WHEN PREPARATION MEETS OPPORTUNITY

ABRAHAM LINCOLN

DISCIPLINE IS CHOOSING BETWEEN WHAT YOU WANT NOW AND WHAT YOU WANT MOST

STEP-BY-STEP GUIDE

On that note, let's set on a journey to pivot your career the way your heart desires!

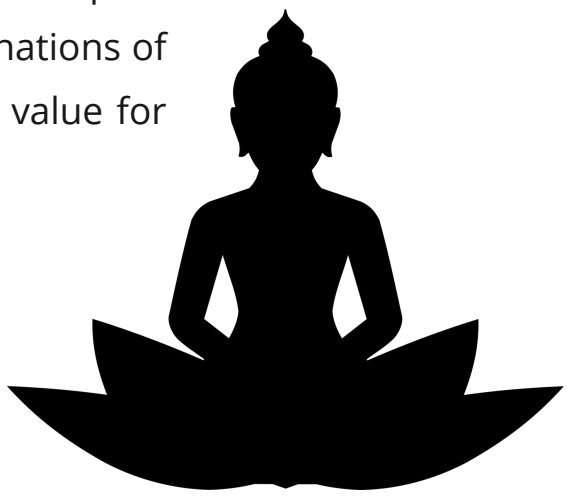
I've designed a step-by-step guide which will take you from zero to being ready for data engineering interviews.

I've made a couple of **assumptions:**

- You need 6 months, assuming that you also have other commitments such as a full-time job or full-time education.
- You will dedicate 10-15h each week

I've researched multiple books, articles and courses online to make sure that what I'm recommending below takes you on a path to learning the **least amount of most important knowledge** that you should acquire before you start applying to jobs. It is a combinations of free and paid resources to represent the best value for money.

Ready? Follow the steps below.



Step 1: Learn Python (2 months)

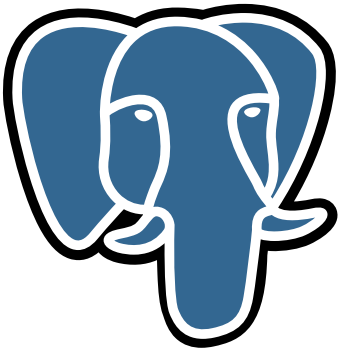
You will need at least one programming language. However, the overwhelming majority of job descriptions mention Python skills. It should be the first thing you learn. I recommend Python For Everybody Specialisation course. It does not only teach you Python syntax, but it shows you how to work with APIs and databases, which are crucial skills for a data engineer.

Step 2: Learn SQL & Databases (2 months)

A data engineer cannot get away without knowing SQL - this query language is standing strong as the lingua franca of data. The course I'm recommending doen't just cover SQL syntax, but it puts it into perspective of data analysis, with a good focus on handling timestamps (you'll find yourself doing that a lot as a data engineer!). Moreover, it teaches more advanced topics around distributed computing using Spark (another sought-after data engineering skill) - Learn SQL Basics for Data Science Specialisation.

Also check out this great introduction to modern databases.

On top of that, it's also important to understand basic principles of data warehouse design. For that, I recommend this well-written introductory article.

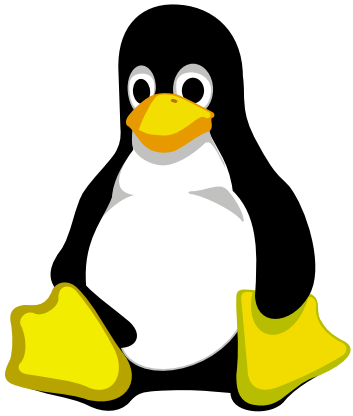


Step 3: Learn Linux and Bash scripting and CRON (1 month)
As a data engineer, you will often need to work with the command line. You should have at least basic familiarity with Linux and Bash scripting, for which I recommend Command Line in Linux course.

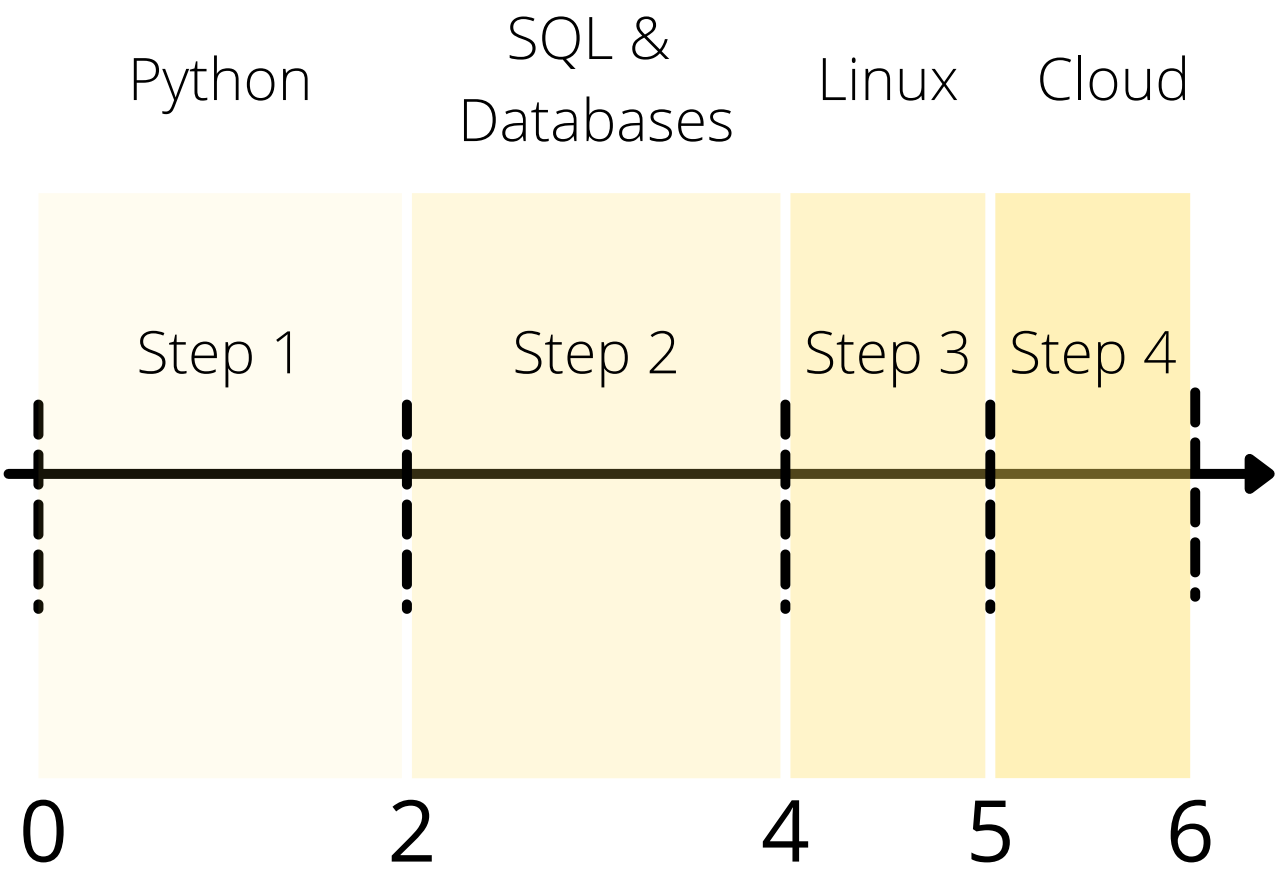
Moreover, familiarise yourself with CRON. It is a time-based job scheduler which you will most likely use to schedule data pipelines. A good article for learning about CRON.

Step 4: Familiarise yourself one cloud computing provider - AWS, GCP, Azure (1 month)
This step is often not crucial for Junior Data Engineer roles, however it will certainly set you apart from other candidates. Cloud is the primary choice for data storage for an increasing number of companies. There are a few significant cloud providers on the market: AWS, GCP, Azure. Based on my own experiences, I would recommend learning AWS from AWS Fundamentals Specialisation.

Disclosure: The above may contain affiliate links, meaning when you click the links and make a purchase, I may receive a commission (at no cost to you)



THE SIX-MONTH TIMELINE*



*Of course, feel free to set your own deadlines according to your individual needs. Perhaps you can learn faster or you're already faimiliar with some parts of the material - shorten the timeline accordingly!

Enjoy the process and good luck!

KAROLINA SOWINSKA

Tech and Lifestyle YouTuber