**1. Data engineering and big data**

Welcome to Data Engineering for Everyone! My name is Hadrien, and I will be your instructor for this course.

**2. About the course**

This is a conceptual course; there is no coding involved. If you are not a developer, the objective is to provide you with a solid enough understanding of the topic so that you can exchange with a data engineering team. If you are interested in actually developing data engineering projects, the objective is to equip you with conceptual knowledge allowing you to get the most out of our data engineering curriculum.

**3. Chapter 1**

This first chapter will clarify what Data engineering is; specifically, how it relates to big data and how a data engineer differs from a data scientist. Data engineers build data pipelines, so we will end this chapter by looking into these as well.

**4. Chapter 2**

Building on these foundations, in the second chapter, we will then take things in order. We will study data storage: the different types of data structures, the central role that the SQL language plays in data engineering, and some storage solutions.

**5. Chapter 3**

Once data is stored, it is ready to be processed. This will be the topic of the third chapter, where we will dive deeper into processing methods and tools, scheduling, parallel computing and cloud computing.

**6. Spotflix**

Throughout the course, we will look at how all these data engineering concepts are implemented at a fictional music streaming company named Spotflix.

**7. Data workflow**

Let's take things from the start then. There are four general steps through which data flows within an organization. First, we collect and ingest data, from web traffic, surveys, or media consumption for example.

**8. Data workflow**

Data is stored in raw format. The next step is to prepare it, which includes "cleaning data", for instance finding missing or duplicate values, and converting data into a more organized format.

**9. Data workflow**

Once the data is clean and organized, it can be exploited. We explore it, visualize it, build dashboards to track changes or compare two sets of data.

**10. Data workflow**

Finally, once we have a good grasp of our data, we're ready to run experiments, like evaluate which article title gets the most hits, or to build predictive models, for example to forecast stock prices.

**11. Data engineers**

Data engineers are responsible for the first step of the process: ingesting collected data and storing it. They have a great responsibility as they lay the ground work for data analysts, data scientists and machine learning engineers. If the data is scattered around, corrupted, and difficult to access, there's not much to prepare, explore, or experiment with.

**12. Data engineers**

And that's exactly why you need a Data engineer: their job is to deliver the correct data, in the right form, to the right people, as efficiently as possible.

**13. A data engineer's responsibilities**

They ingest data from different sources, optimize the databases for analysis, and manage data corruption. Data engineers develop, construct, test, and maintain architectures such as databases and large-scale processing systems to process and handle massive amounts of data. If you're not sure what this all means, that's okay! The course will unpack all this jargon and explain the what, why, and how.

**14. Data engineers and big data**

With the advent of big data,

**15. Data engineers and big data**

the demand for data engineers has increased. Big data can be defined as data so large you have to think about how to deal with its size, because it's difficult to process using traditional data management methods.

**16. Big data growth**

This graph helps make sense of the growth of big data. In order of volume, big data is mainly composed of sensors and devices data, social media data, enterprise data and VoIP data.

1. 1 Data Age 2025, Seagate, November 2018

**17. The five Vs**

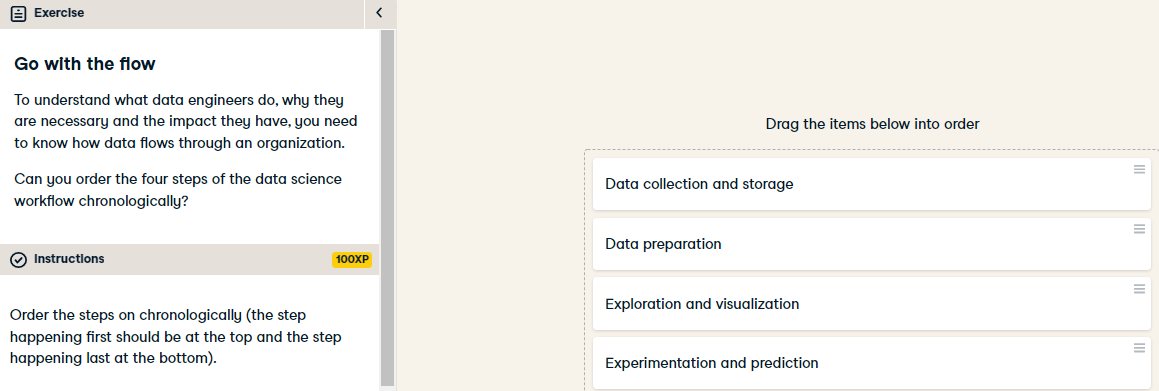
Big data is commonly characterized by five Vs: volume (the quantity of data points), variety (type and nature of the data: text, image, video, audio), velocity (how fast the data is generated and processed), veracity (how trustworthy the sources are), and value (how actionable the data is). Data engineers need to take all of this into consideration.

**18. Summary**

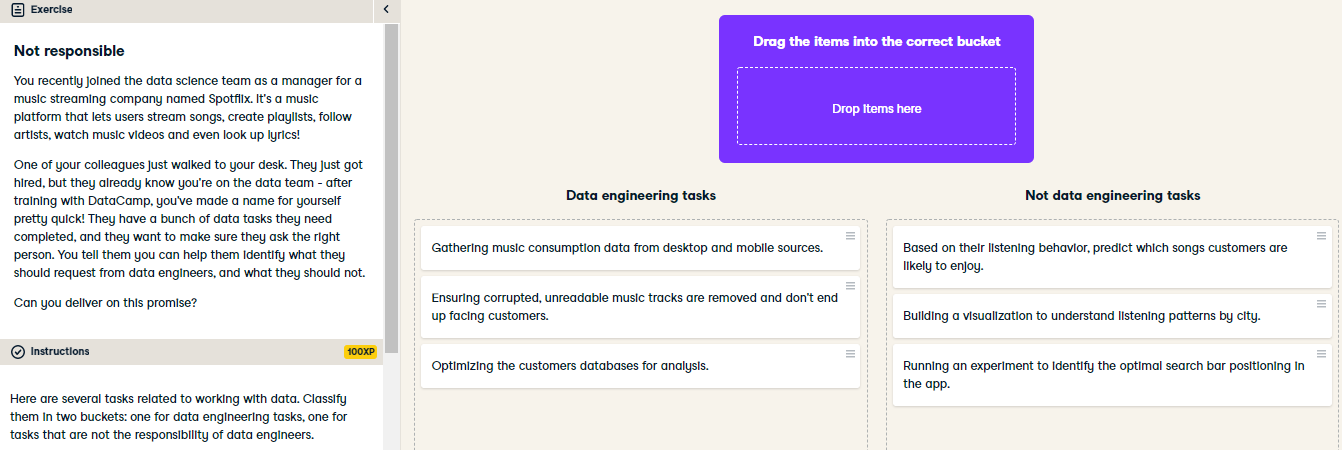
Great! Now you not only know what's waiting for you in this course, but also how data flows, when a Data engineer intervenes, what their responsibilities are, and how they relate to big data.

**19. Let's practice!**

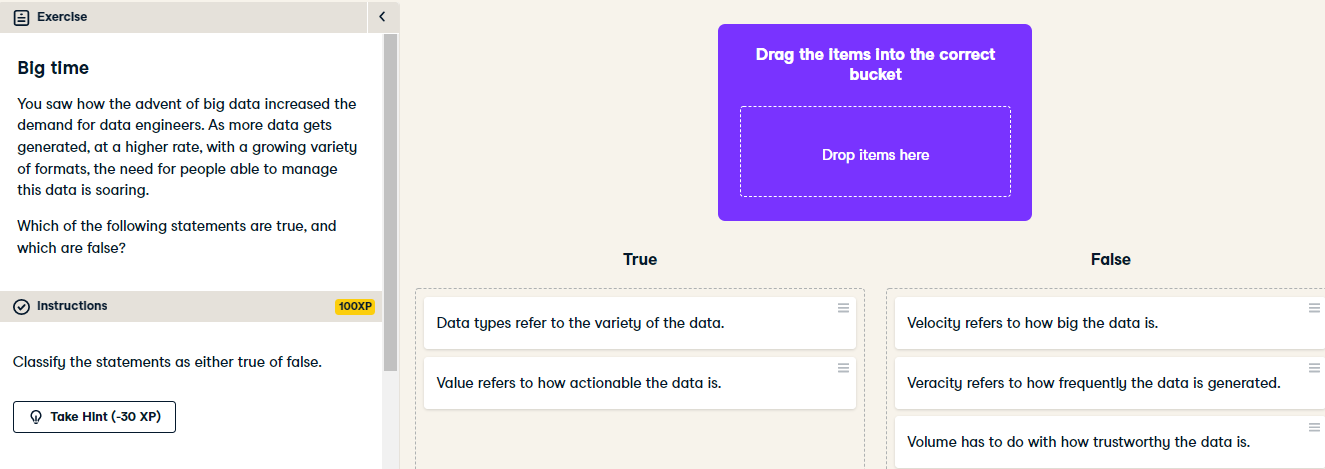
Let's check your understanding!



Great job! You know how data flows through an organization. Let's jump in the next exercise to check how well you understand a data engineer's responsibilities!



Fantastic! You definitely know what a data engineer is responsible for, and it's also much clearer for your colleague now.



Awesome job! Now you have a clear understanding of what qualifies big data, beyond the buzzword.