**1. The data pipeline**

Alright, we've mentioned the term pipeline several times by now, so let's focus on it for this lesson.

**2. If data is the new oil...**

You may have heard that "Data is the new oil", as first coined by The Economist, so let's follow this idea.

1. 1 The Economist, 2017-05-06, by David Parkins

**3. Oil well**

We extract crude oil from an oil field.

**4. Oil well pipe**

We move the crude oil to

**5. Distilling**

a distillation unit, where we

**6. Residue**

separate

**7. Heavy oil**

the

**8. Diesel**

oil

**9. Kerosene**

into

**10. Napththa**

several

**11. Gasoline**

products. Some products are sent directly to their final users.

**12. Airport**

For example, some pipes go straight to airports to deliver kerosene. Other products, like gasoline,

**13. Gas storage facility**

are sent to gas storage facilities and stored in big tanks,

**14. Gas stations**

before being distributed to gas stations.

**15. Naphtha is transformed**

Other products, like naphtha, go through several chemical transformations,

**16. Factory receives plastic**

Manufacturers use synthetic polymers to create products, like CDs. As you can see, we have many pipelines tying it all together.

**17. Back to data engineering**

CDs? So last century, Vivian thinks. However, to manage data for Spotflix, she follows a procedure similar to oil processing. Companies ingest data from many different sources, which needs to be processed and stored in various ways. To handle that, we need data pipelines that efficiently automate the flow from one station to the next, so that data scientists can use up-to-date, accurate, relevant data. This isn't a simple task and that's why data engineers are so important.

**18. Mobile**

At Spotflix, we have sources from which we extract data. For example, the users' actions and listening history on the mobile Spotflix app

**19. Computer**

and the desktop Spotflix app,

**20. Website**

and the Spotflix website itself. We also have websites Spotflix uses internally, like their HR management system for payroll and benefits.

**21. Ingesting artists data**

The data is ingested

**22. Ingesting albums data**

into

**23. Ingesting artists data**

Spotflix's system,

**24. Data lake**

moving from their respective sources to our data lake (no fear, we will talk about data lakes in the next chapter).

**25. First pipelines**

These are our first three pipelines.

**26. Artists database**

We then organize the data, moving it into databases (we will talk more about databases in Chapter 2 as well). It could be artist data, like name, number of followers, and associated acts,

**27. Albums database**

albums data, like label, producer, year of release,

**28. Tracks database**

tracks data, like name, length, featured artists, and number of listens,

**29. Playlists database**

playlists data, like name, song it contains,and date of creation,

**30. Customers database**

customers data, like username, account opening date, subscription tier,

**31. Employees database**

or employees data, like name, salary, reporting manager, updated by human resources.

**32. Six more pipelines**

These are six new pipelines.

**33. Album covers database**

Some albums data can be extracted and stored directly. For example, album cover pictures all have the same format, so we can store them directly without having to crop them.

**34. One more pipeline**

One more pipeline!

**35. Sales employees table**

Employees could be split in different tables by department, for example sales,

**36. Engineering employees table**

engineering,

**37. Support employees table**

support, etc. We will talk about tables in Chapter 2 as well.

**38. Three more pipelines**

For now, three more pipelines!

**39. Sales USA employees table**

These tables could be further split by office, for example the US,

**40. Sales Belgium employees table**

Belgium,

**41. Sales UK employees table**

and the UK. If data scientists had to analyze employee data(to investigate employee turnover for example), this is the data they would use.

**42. Three more pipelines**

Three more pipelines!

**43. Checking for corrupted tracks**

Tracks would need to be processed, first to check if the track is readable, then to check if the corresponding artist is in the database, to make sure the file is in the correct size and format, etc.

**44. One more pipeline**

That's one more pipeline, that we will unpack into Chapter 3 when we will talk about data processing.

**45. Clean tracks database**

The data can then be stored in a new, clean tracks database. This is one of the databases data scientists could use to build a recommendation engine by analyzing songs for similarity, for example.

**46. One last pipeline**

And that's our last pipeline!

**47. You get a pipeline!**

Alright! That's

**48. You get a pipeline!**

a lot

**49. Everybody gets a pipeline!**

of pipelines!

**50. Data pipelines ensure an efficient flow of the data**

In a nutshell, data pipelines ensure the data flows efficiently through the organization. They automate extracting, transforming, combining, validating, and loading data, to reduce human intervention and errors, and decrease the time it takes for data to flow through the organization. Don't worry, we'll cover this in detail in the last chapter.

**51. ETL and data pipelines**

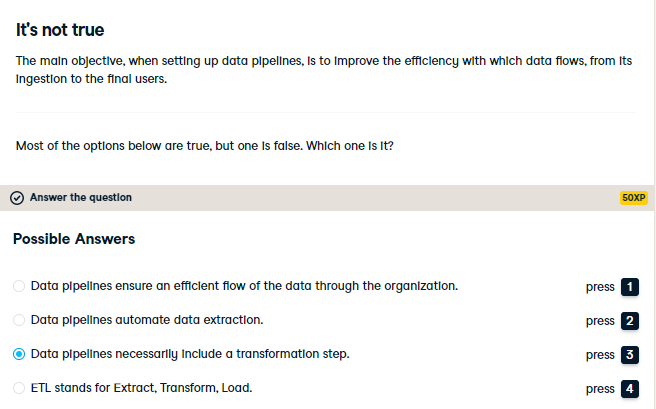
One term you will hear a lot is "ETL". It's a popular framework for designing data pipelines. It breaks up the flow of data into three sequential steps: first E for extracting the data, then T for transforming the data, and finally, L for loading this transformed data to a new database. The key here is that data is processed before it's stored. In general, data pipelines move data from one system to another. They may follow ETL, but not all the time. For instance, the data may not be transformed, and routed directly to applications like visualization tools or Salesforce.

**52. Summary**

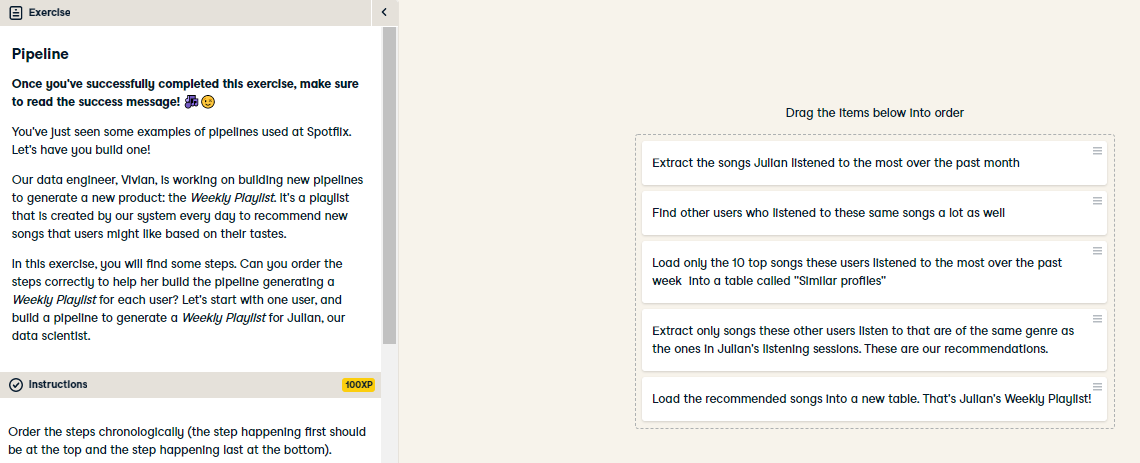
OK, now you understand what a data pipeline is, what it's used for, why it's important, how we use them at Spotflix, and where ETL fits in.

**53. Let's practice!**

Let's solidify your understanding of data pipelines with a couple exercises, and then onwards to Chapter 2 to dive into the details data storage.



ETL stands for Extract, Transform, Load and is a popular framework for designing data pipelines. However, data pipelines do not necessarily include a transformation step.



Excellent job! And just like that, you reached the end of Chapter 1. Congratulations! Jump into Chapter 2 to learn more about the different ways to data and optimize this storage, before diving into data processing in Chapter 3!

By the way, you may have noticed the exercise titles are actually song titles. That remains true throughout the course. So here's the deal: if you make it to the end of the course, I will tell you where to find the playlist!