**1. Processing data**

Welcome to this third and final chapter! Now we have a clear idea of what data engineering is, and how data can be stored in different ways. This chapter will cover the last step on our roadmap: moving and processing data.

**2. Our data pipeline**

Let's have a quick look at the pipeline again to build some intuition.

**3. Our data pipeline**

When we move data to the data lake,

**4. Our data pipeline**

when we split it into different tables,

**5. Our data pipeline**

or when we remove corrupted tracks, we are processing data.

**6. A general definition**

So what does it mean to "process" data? In a nutshell, data processing consists in converting raw data into meaningful information.

**7. Data processing value**

Precisely, why do we need to process data? Well, there may be some data that we don't need at all. When rolling out a new feature, we may be watching a lot of indicators to ensure it's working as expected. But once we're sure it's stable and well integrated, we don't need this data anymore. Storing and processing data is not free, so we want to optimize our memory, process and network costs. Uncompressed data can be ten times larger than compressed one: imagine if we had to process that! Our whole business model would collapse. Some data may come in a type, but would be easier to use in another. For example, there is a tradeoff between file size and sound quality of the music tracks.

**8. At Spotflix**

At Spotflix,

**9. At Spotflix**

artists may upload data in wav or flac format, which are high quality master files. Letting users stream these big files would incur big network costs. The data is processed by converting the master files to the .ogg format, a lighter format with slightly lower sound quality.

**10. At Spotflix**

It's these files that we will stream to our users.

**11. Data processing value**

We want to move and organize data so it is easier for analysts to find what they need, like you saw on the data pipeline graph. Music files also contain metadata, like the name of the artist and the genre. The data is again processed to extract the metadata and store it in a database, for easy access by data analysts and data scientists. You may want your data to fit a certain schema or structure, to reap the benefits covered in the previous chapter. We gather employee data and fit it to the specific table schema you saw with the employee table, separating name and last name, using logic instead of text to distinguish between part-time and full-time employees, etc. Data processing also increases productivity. At Spotflix, we automate all the data preparation steps we can, so that when it arrives to data scientists, they can analyze it almost immediately. The value they add to the company originates from the insights derived from their analyses, so we need to help them focus on and deliver exactly that.

**12. How data engineers process data**

In terms of data processing, data engineers have different responsibilities. They perform data manipulation, cleaning, and tidying tasks that can be automated, and that will always need to be done, regardless of the analysis anyone wants to do with them. For example, rejecting corrupt song files, or deciding what happens with missing metadata. What should we do when the genre is missing? Do we reject the file, do we leave the genre blank, or do we provide one by default? They also ensure that the data is stored in a sanely structured database, and create views on top of the database tables for easy access by analysts. Views are the output of a stored query on the data. For example, artist data and album data should be stored in separate tables in the database, but people will often want to work on these things together. That means data engineers need to create a view in the database combining both tables. A<qqqqq.

**13. Tools**

There are a bazillion data processing tools, but they are out of the scope of this course.

1. 1 The difference between batch and stream will be explained in the next lesson!

**14. Apache Spark**

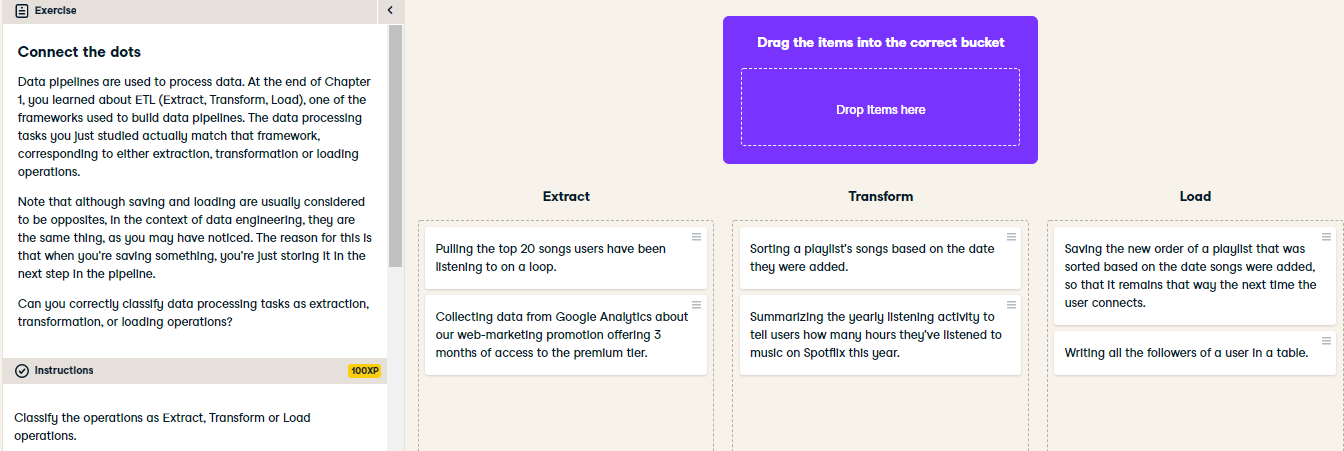
One such tool is Apache Spark,for which you can find courses on DataCamp if you're interested.

**15. Summary**

Alright, now you can tell what data processing is, why it's necessary, what it consists in and how we process data at Spotflix.

**16. Let's practice!**

Let's hammer the nail with some exercises!



Great job connecting the dots! As you can see, most of the data processing steps are transformation steps including selecting, encoding, calculating new values, sorting and ordering, joining data from multiple sources, aggregating, pivoting tables, splitting one table into several ones, removing duplicate values.

Depending on the context, validating data can be considered part of the extract step (you don't extract if it's not necessary), or the transform step (you check that your results make sense before saving them).