16.3.4

Spark API

Now that you and Jennifer have an overview of how to use Spark, it's time to dig in and plug in the necessary APIs.

Language APIs

Spark is very accessible through different programming languages. Spark was written in Scala, a tough programming language. However, there is an API that works with many languages to translate the code into Spark code and execute.

Spark has an API that supports the following languages:

- Java
- Python
- SQL
- R

This course will focus on using the Python flavor of Spark, or PySpark.

Data APIs

Spark supports two different sets of data APIs:

- **Low-level unstructured API** is mostly outdated but deals with **resilient distributed datasets (RDDs)**, which are an immutable collection of records that can be operated in parallel. These were part of early versions of Spark.
- **High-level API** consists of structured data such as comma-separated values (CSV).

The high-level API consists of three core forms of data that you'll work with in Spark. Notice that all three contain familiar structures:

- Datasets

- DataFrames
- SQL tables and views

We'll focus on Spark's high level, but it's important to understand that there is some lower level data that Spark can work with. You won't use the low level APIs very often, and generally almost all situations will be better served using structured APIs. However, there are a couple of scenarios where low-level APIs might apply:

- You need finely tuned control over the data in your clusters that high-level APIs can't provide.
- Your role involves maintaining legacy code that uses low-level APIs.

Now that you have some background in Spark, let's set it up on your machine.

Retake