SCHOOL OF DATA
by Springboard

Spark Optimization Mini Project
Optimize Existing Code

Estimated Time: 1-2 hours

Spark can give you a tremendous advantage when it comes quickly processing massive datasets. However, the tool is only as powerful as the one who wields it. Spark performance can become sluggish if poor decisions are made in the layout of the code and the functions that are chosen.

This exercise gives you hands-on experience optimizing PySpark code. You'll look at the physical plan for the query execution and then modify the query to improve performance.

1. Download the code and the data from here
2. Extract the archive into a folder on your local computer with PySpark setup
3. The file named 'optimize.py' contains the code that you need to optimize

For example, we want to compose a query in which we get for each question also the number of answers to this question for each month. A version of this query exists in the .py file, but does so in a suboptimal way. Your task is to try to rewrite it and achieve more optimal performance.

As you might remember from our spark subunit, there are several ways one can improve performance of a Spark job:

1. By picking the right operators
2. Reduce the number of shuffles and the amount of data shuffled
3. Tuning Resource Allocation
4. Tuning the Number of Partitions
5. Reducing the Size of Data Structures
6. Choosing Data Formats

Deliverables:

1. Push code to GitHub with the Readme MarkDown file explaining what was the issue with the code and how you achieved a better performance.