SCHOOL OF DATA
by Springboard

Data Engineering Career Track
Open-ended Capstone step 6: Scale Your Prototype

Estimated Time: 6-9 Hours

In this step of the project, you can work with the entire dataset that you acquired at the beginning of the project and use cloud resources to scale up the data pipeline prototype that you created in the last step of the project.

Suggested steps for scaling the prototype pipeline:

1. Move from Python code for the data pipeline to Apache Spark
    a. As you will be using the entire data set in this step, you would want to leverage the speed that Apache Spark provides for processing massive datasets. In order to achieve this, you would have to replicate the logic that you have developed so far in PySpark.
    b. Please make sure that OOP concepts and logging is implemented while developing PySpark code (same as step 5 of your capstone).
2. Create an Azure Blob Storage for the dataset to be stored (using techniques learned in 18.3.1)
3. Create a Spark cluster in Azure (as shown here) in order to process the data at scale using the Apache Spark code that you created above in in #1
4. Make sure that the Spark code points to the right blob storage in order to read and store the data using code below:

```
spark. jsc.hadoopConfiguration().set("fs.azure.account.keyprovider.{blobname}.blob.core.windows.net","org.apache.hadoop.fs.azure.SimpleKeyProvider")
spark. jsc.hadoopConfiguration().set("fs.azure.account.key.{blobname}.blob.core.windows.net","{storagekey==}")
spark.read.csv("wasb://{containername}@{blobname}.blob.core.windows.net/students/students.csv").show()
```

Replace `{blobname}` with your blob name where the container with data is stored. Replace `{containername}` with the name of blob container where your data is stored.

5. Submit you spark code for execution to the Spark Cluster and see the performance gain!

Deliverables:

- Update the project's GitHub repo with the work you complete for this step of the project and update the readme markdown file
- A slide deck with explanations of the setup you have for scaling up the prototype.

Slide deck updates:

From this step, you should include:

- Details about how you migrated your code to PySpark, and the performance gains you see as a result
- Snapshots of you cluster creation