**Got Data? Use Spark in Azure Databricks**

Challenge 2 – Running a Spark Job

What You’ll Need

To complete the labs, you will need the following:

• A web browser

• A Microsoft account

• A Microsoft Azure subscription

• A Windows, Linux, or Mac OS X computer

• Azure Storage Explorer

• The lab files for this course <only if you get stuck – proctors have access to this content>

**Note**: To set up the required environment for the lab, follow the instructions in the [**Setup**](https://microsoft-my.sharepoint.com/personal/laedell_microsoft_com/Documents/Hackathons/ML/Databricks/databricks_intro/databricks-introFinal/Setup%20Guide.docx)document for

this course. Specifically, you must have signed up for an Azure subscription.

Challenge Background: Spark jobs enable you to run data processing code on-demand or at scheduled intervals. This enables you to build data processing solutions for unattended execution.

Business Challenge: your business has asked you to analyze voluminous web server logs

Questions:

1. What types of analyses would be good when working with website log data ? Remember, you want to provide value to the business quickly.
2. If you look at the ProcessLog.py files, what Spark modules are being used and why do you think they were chosen? Would you have chosen alternate modules and why?
3. What else would you add to the script to enhance what it is already doing using what you learned in Challenge 1?
4. Can you explain the purpose of the different files that were created by the Spark job process in your blob storage?
5. What is the file name(s) created that contains the Processed Results? Does it contain the # of webpage views for each product that you were expecting?

To get you started, the 1st thing you will need to do is to Import your ProcessLog.py or .scala script into your Databricks Workspace. How? Ask your table mates.

\*\* remember: you can use scala or python; you can use Databricks notebooks or native scripts (.py, for example) to solve the Challenge \*\*

\*\*\*Lastly. If you get stuck, proctors can help you without providing the answer.

**Challenge 2:**

2.1) Access Databricks Workspace from Challenge 1 or Provision a New Databricks Workspace if you didn’t complete Challenge 1

2.2) Access existing Storage Account & Create a new Container for Challenge 2 (if you didn’t complete Challenge 1, create a new Storage Account / Container)

**helpful tips:**

\*\*Upload the log source files - These logs have been made available as part of the Teams site under Files <IISLog.txt>

2.3) Create Spark Job in Azure Databricks

**helpful tips:**

\*\* you will need to import the ProcessLog.py or ProcessLog.scala file from Teams to your Databricks Workspace (file found in Teams files)

\*\*you will need to edit the scripts with your storage account details – also, add other transformers or use different modules if you want to test out your knowledge or have a different idea for how to approach the challenge – points rewarded for teamwork and for creative thinking!!

**Challenge considerations:**

1. Access your Databricks workspace to create Spark job – Spark jobs can be used to automate execution of your scripts. Spark jobs that are created can be checked into a DevOps CICD pipeline along with artifacts created in your Databricks workspace (eg. Notebooks and other script files).

2. For the **Task** option, click **Select Notebook** and choose the **ProcessLog** code file you

uploaded previously (either .py or .scala)

3. For the **Parameters** option, enter the following key-value pair **<note: values are case sensitive>**

logfile

IISlog.txt

4. **Cluster** settings - , click **Create New (this is your jobs cluster):**

• Cluster Type: New Cluster

• Databricks Runtime Version: *Choose the latest available version*

• Python Version: 3

• Driver Type: Same as worker (leave default)

• Workers: 1

* **Important setting to append:**

• **Spark Config**: Add a key-value pair for your storage account and key like this:

fs.azure.account.key.***your\_storage\_account***.blob.core.windows.net ***your\_access\_key1\_value***

5. For the **Schedule** option, choose a schedule if you want OR click **Run Now** to start the job.

6. View job status as it moves from **Pending to Running to Succeeded.**  Notice how Active Runs changes as status changes

END OF CHALLENGE 2