Ask Bigger Questions

CAP - Developing with Spark and Hadoop:

Homework Assignment Guide for Students

Homework: Write and Run a Spark Application

Files and Data Used in This Homework:

Exercise Directory: `$DEV1/exercises/spark-application`

Data files (HDFS): `/loudacre/weblogs`

Scala Project: `countjpgs`

Scala Classes (professor VM only):

`stubs.CountJPGs`

`solution.CountJPGs`

Python Stub (professor VM only):

`CountJPGs.py`

Python Solution (professor VM only):

`$DEV1/exercises/spark-application/solutions/CountJPGs.py`

In this Exercise you will write your own Spark application instead of using the interactive Spark Shell application.

Write a simple program that counts the number of JPG requests in a web log file. The name of the file should be passed into the program as an argument.

This is the same task you did earlier in the "Use RDDs to Transform a Dataset" exercise. The logic is the same, but this time you will need to set up the SparkContext object yourself.

Depending on which programming language you are using, follow the appropriate set of instructions below to write a Spark program.

Before running your program, be sure to exit from the Spark Shell.

Write a Spark application in Python

> You may use any text editor you wish. If you don't have an editor preference, you may wish to use gedit, which includes language-specific support for Python.

1. A simple stub file to get started has been provided in the exercise project (professor VM only): `$DEV1/exercises/spark-application/CountJPGs.py`. This stub imports the required Spark class and sets up your main code block.

2. Set up a SparkContext using the following code:

```
sc = SparkContext()
```

3. In the body of the program, load the file passed into the program, count the number of JPG requests, and display the count. You may wish to refer back to the "Getting Started with RDDs" exercise for the code to do this.

4. Run the program, passing the name of the log file to process, e.g.:

```
$ spark-submit CountJPGs.py /loudacre/weblogs/*
```

Write a Spark application in Scala

> You may use any text editor you wish. If you don't have an editor preference, you may wish to use gedit, which includes language-specific support for Scala. If you prefer to work in an IDE, Eclipse is included and configured for the Scala projects in the course. However, teaching use of Eclipse is beyond the scope of this course.

A Maven project to get started has been provided (professor VM only): `$DEV1/exercises/spark-application/countjpgs`

1. Before starting this exercise, start the Archiva service; this service provides a local maven repository that has been pre-cached with the libraries you will need for these exercises:

```
$ sudo service archiva start
```

2. Edit the Scala class defined in `CountJPGs.scala` in `src/main/scala/stubs/` (professor VM only).

3. Set up a SparkContext using the following code:

```
val sc = new SparkContext()
```

4. In the body of the program, load the file passed in to the program, count the number of JPG requests, and display the count. You may wish to refer back to the "Use RDDs to Explore and Transform a Dataset" exercise for the code to do this.

5. From the `countjpgs` project directory, build your project using the following command:

```
$ mvn package
```

6. If the build is successful, it will generate a JAR file called `countjpgs-1.0.jar` in `countjpgs/target`. Run the program using the following command:

```
$ spark-submit \
--class stubs.CountJPGs \
target/countjpgs-1.0.jar /loudacre/weblogs/*
```

Submit a Spark application to the cluster

In the previous section, you ran a Python or Scala Spark application using `spark-submit`. By default, `spark-submit` runs the application locally. In this section, run the application on the YARN cluster instead.

1. Re-run the program, specifying the cluster master in order to run it on the cluster. Use one of the commands below depending on whether your application is in Python or Scala.

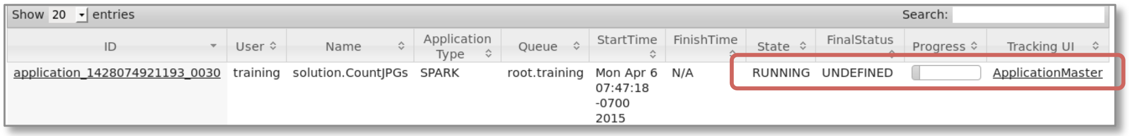
   To run Python:

   ```
   $ spark-submit \
     --master yarn-client \
     CountJPGs.py /loudacre/weblogs/*
   ```

   To run Scala:

   ```
   $ spark-submit \
     --class stubs.CountJPGs \
     --master yarn-client \
     target/countjpgs-1.0.jar /loudacre/weblogs/*
   ```

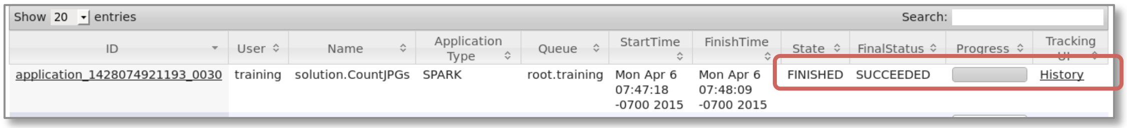
2. After starting the application, open Firefox and visit the YARN Resource Manager UI using the provided bookmark (or going to URL `http://localhost:8088`). While the application is running, it appears in the list of applications something like this:



Show 20 entries Search:

ID	User	Name	Application Type	Queue	StartTime	FinishTime	State	FinalStatus	Progress	Tracking UI
application_1428074921193_0030	training	solution.CountJPGs	SPARK	root.training	Mon Apr 6 07:47:18 -0700 2015	N/A	RUNNING	UNDEFINED		ApplicationMaster

   After the application has completed, it will appear in the list like this:



Show 20 entries Search:

ID	User	Name	Application Type	Queue	StartTime	FinishTime	State	FinalStatus	Progress	Tracking UI
application_1428074921193_0030	training	solution.CountJPGs	SPARK	root.training	Mon Apr 6 07:47:18 -0700 2015	Mon Apr 6 07:48:09 -0700 2015	FINISHED	SUCCEEDED		History

3. Take note of the your application's ID (e.g. `application_12345`…). (You may wish to copy it.) In a terminal window, enter

```
$ yarn logs -applicationId <your-app-id>
```

This is the end of the Homework