cloudera®

Ask Bigger Questions

CAP - Developing with Spark and Hadoop:

Homework Assignment Guide for Students

Homework: Configure a Spark Application

Files Used in This Homework:

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

Data files (HDFS)

`/loudacre/weblogs/*`

Properties files (local)

`spark.conf`

`log4j.properties`

In this exercise you will practice setting various Spark configuration options.

You will work with the CountJPGs program you wrote in the prior exercise.

Set configuration options at the command line

1. Rerun the CountJPGs Python or Scala program you wrote in the previous exercise, this time specifying an application name. For example:

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

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

2. Visit the Resource Manager UI again and note the application name listed is the one specified in the command line.

3. Optional: View the Spark Application UI. From the RM application list, follow the ApplicationMaster link (if the application is still running) or the History link to visit the Spark Application UI. View the Environment tab. Take note of the `spark.*` properties such as `master`, `appName`, and `driver` properties.

Set configuration options in a configuration file

4. Change directories to your exercise working directory. (If you are working in Scala, that is the `countjpgs` project directory.)

5. Using a text editor, create a file in the working directory called `myspark.conf`, containing settings for the properties shown below:

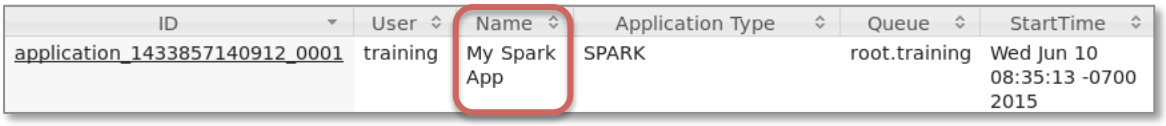
```
spark.app.name  My Spark App
spark.master    yarn-client
spark.executor.memory     400M
```

6. Re-run your application, this time using the properties file instead of using the script options to configure Spark properties:

```
$ spark-submit --properties-file myspark.conf \
  CountJPGs.py /loudacre/weblogs/*
```

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

7. While the application is running, view the YARN UI and confirm that the Spark application name is correctly displayed as "My Spark App"



ID	User	Name	Application Type	Queue	StartTime
application_1433857140912_0001	training	My Spark App	SPARK	root.training	Wed Jun 10 08:35:13 -0700 2015

Set logging levels

8. Copy the template file `/etc/spark/conf/log4j.properties.template` to `log4j.properties` in your exercise working directory.

9. Edit `log4j.properties` . The first line currently reads:

```
log4j.rootCategory=INFO, console
```

   Replace `INFO` with `DEBUG`:

```
log4j.rootCategory=DEBUG, console
```

10. Rerun your Spark application. Because the current directory is on the Java classpath, your `log4.properties` file will set the logging level to `DEBUG`.

11. Notice that the output now contains both the INFO messages it did before and DEBUG messages, e.g.:

```
15/03/19 11:40:45 INFO MemoryStore: ensureFreeSpace(154293) called
with curMem=0, maxMem=311387750
15/03/19 11:40:45 INFO MemoryStore: Block broadcast_0 stored as
values to memory (estimated size 150.7 KB, free 296.8 MB)
15/03/19 11:40:45 DEBUG BlockManager: Put block broadcast_0 locally
took  79 ms
15/03/19 11:40:45 DEBUG BlockManager: Put for block broadcast_0
without replication took  79 ms
```

   Debug logging can be useful when debugging, testing, or optimizing your code, but in most cases generates unnecessarily distracting output.

12. Edit the `log4j.properties` file to replace `DEBUG` with `WARN` and try again. This time notice that no INFO or DEBUG messages are displayed, only WARN messages.

13. You can also set the log level for the Spark Shell by placing the `log4j.properties` file in your working directory before starting the shell.

Try starting the shell from the directory in which you placed the file and note that only WARN messages now appear.

Note: Throughout the rest of the exercises, you may change these settings depending on whether you find the extra logging messages helpful or distracting.

This is the end of the Homework