Activity 03– AWS MapReduce

Purpose

1. To be able describe the MapReduce computational model
2. To be able to solve simple problems with MapReduce
3. To be able to run MapReduce programs with the AWS Elastic MapReduce service

Activity

Please complete the following and submit a report.

1. **Simple Map Reduce computation** (10 points). Follow the tutorial below and implement a MapReduce calculation for counting words. As you work through the tutorial keep a log of your work in the following format:

Time	Activity	Notes and Reflections
start	login	I logged into the console. My password is really hard to remember. I wonder if it is possible to change the password for my account. Will try that later …
+5 min	Browsing documentation	Spending time browsing the tutorial for Elastic Map Reduce. I've been wondering how Hadoop and AWS EMR are related?
…	…	…

   Please make an entry in your log at least every 5 minutes. (If you spend more than 3 hours on this please stop.)

2. **Reflection** (5 points). As a programmer, write a brief reflection (200-400 words) on your experience with AWS Elastic Map Reduce. You may want to address such question as: What was the process like? What was easy? What was difficult? What mistakes did you make? How might the process of submitting a job be made easier?

3. **What is LongValueSum?** (2 points). Inspect the Map script **wordSplitter.py**. Why does each record begin with LongValueSum? Briefly discuss. (Hint: You may need to conduct a web search to find a satisfactory answer to this question.)

4. **Power Frequency Distribution**. The directories **info445sp2013common/inputs/power** and **info445sp2013common/inputs/small_power** contain data is the same format as the Work_Log table from assignment 02. Note: The second file is considerably smaller than the first.

   Suppose you wanted to compute a frequency distribution of the amount of power used in in 15 minutes across all devices. The distribution might look like the following:

kWh	Frequency
0	100
1	101
2	3000
…	

   a. What is the SQL statement of computing this distribution? (2 points)
   b. With pseudo code, write a Map and a Reduce function for computing this distribution (4 points).
   c. Implement the Map and Reduce functions for AWS Elastic Map Reduce (2 points). Please include your code and a brief summary if you got it to work correctly or not. As you do this question, keep of log as in Question #1 above. **Note: Please do not spend a huge amount of time on this implementation activity unless you really want to.**

5. **Power Demand**. In a 24hr day there are ninety-six 15 min time slices. Suppose you wanted to compute the minimum, average, and maximum demand for power in each of these time slices for each building.

   a. With pseudo code, write Map and Reduce functions to compute these results. (Hint: You may need to compute two MapReduce calculations in sequence.) (6 points)
   b. Implement the Map and Reduce functions in AWS Elastic Map Reduce (2 points). Please include code and a brief summary if you got it to work correctly or not. As you do this question, keep of log as in Question #1 above. **Note: Please do not spend a huge amount of time on this implementation activity unless you really want to.**

Your Amazon Web Service Account

To login into your AWS account you need the following:

1. Your user ID, which is your UW Net ID
2. Your AWS password, which will be given out to you on a piece of paper.

To login go this URL and enter your user ID and your AWS password:

https://info445sp2013.signin.aws.amazon.com/console

If should have received your AWS password from the teaching assistant or the instructor. If not, please ask.

Services

You will see a large number of AWS services. You may want to explore some of these services to see what is available.

Elastic Map Reduce on AWS: Your Frist Program

We are going to start with Elastic MapReduce, which provides access to a large installation of Hadoop, an open source implementation of MapReduce (see Dean & Ghemawat, 2008).

A key document for learning about Elastic MapReduce is:

Amazon Elastic MapReduce: Developer Guide
http://s3.amazonaws.com/awsdocs/ElasticMapReduce/latest/emr-dg.pdf

Turn to the Get Started section of the Developer Guide on p. 13, which presents a tutorial for getting started with Elastic MapReduce . In the following steps I will guide you through the tutorial for our particular class setup. Note: This is complex, technical work so I recommend that you go slow and make sure each step is working before you take the next step.

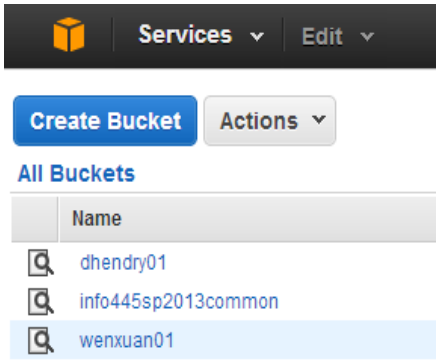
Step 1: Data Set-up

Before we can run MapReduce computation we have set-up a data environment. We will be using the Amazon Simple Storage System.

Do the following:

1. With your user id and password sign-in here: https://info445sp2013.signin.aws.amazon.com/console

2. Click on the services menu and then click on S3 (Amazon Simple Storage System). You should see something like this:



3. Click into the bucket called **info445sp2013common**. Then click into the folder called **inputs**. Browse around the subfolders and notice that some data has already been loaded into the S3. You will use this data in your first MapReduce programs.
4. Click into the folder **info445sp2013common/scripts.** You will find some scripts there that provide implementations of Map and Reduce functions. Inspect the python script wordSplitter.py careful. Note: This is map function. What does it do?
5. Now create a bucket for your own work. So that we can keep track of our buckets I suggest that you call it userid01 where userid is the User ID that you used to login to the system.
6. Within your bucket, say **userid01**, create the following directories:

    userid01/output -- for holding the output of MapReduce programs
    userid01/logs -- for logging and debugging information
    userid01/scripts -- for your own map and reduce scripts
    userid01/inputs -- for your own input data

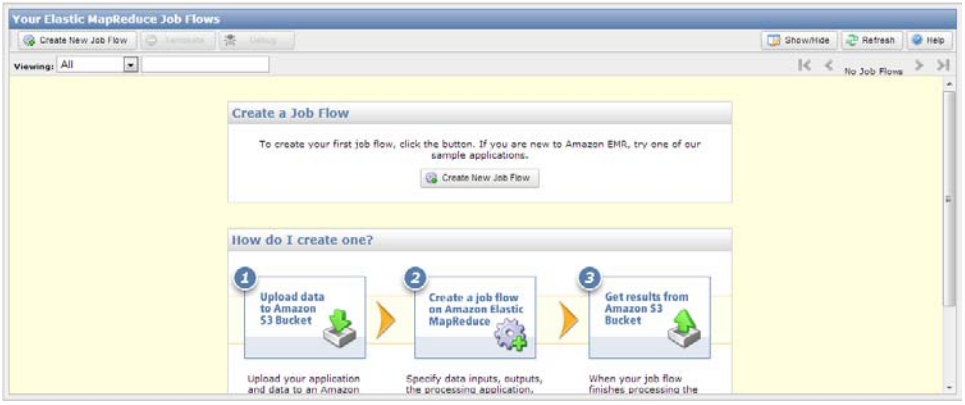
Step 2: Data Set-up

1. Within S3 create an output folder for the output of your first program. The output folder should be named as follows:

   userid01/output/data02

Step 3: Creating an Elastic MapReduce Job

1. From the services menu click Elastic MapReduce. You should get something like:



2. Click on Create New Job Flow (top left). We are now back on track with the Amazon tutorial (see p. 18).
3. Fill in the Create a New Job Flow screen just as done in the tutorial and click continue (see p. 18 of tutorial).
4. Fill in the Input location, Output location, Mapper script, and Reducer script as follows (see p. 19 of tutorial):

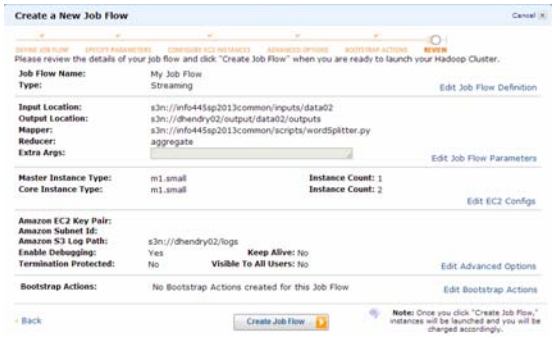
   Input Location: **info445sp2013common/inputs/data02**
   Output Location: **userid01/output/data02/outputs**
   (*Note: This location must NOT alredy exist*.)
   Mapper: **info445sp2013common/scripts/wordSplitter.py**
   Reducer: **aggregate**

5. Select all default settings on the next screen (see p. 20 of the tutorial)
6. Fill in the Amazon S3 Log Path as follows (see p. 22 of the tutorial):

   Amazon S3 Log Path: **userid01/logs**

7. On the Bootstrap options screen select the defaults (see p. 23 of the tutorial)

8. The Review screen should look something like this:



9. Click Create Job Flow – this starts the MapReduce computation. It will take several minutes to complete, that is, to be scheduled and executed.
10. When the computation completes the output will be found in the output file location: **userid01/output/data02/outputs**