Amazon Elastic MapReduce

Duke University CPS 516

Jie Li
Overview

• Check Consolidated Bill
• Sign Up EMR service
• Create S3 Bucket
• Create Key Pair
• Create Job Flow
• Configure Network Security
• SSH to the Master Node
Navigation

https://console.aws.amazon.com

We’ll need to use S3, EC2, Elastic MapReduce
Consolidated Billing

Make sure you are enrolled

https://portal.aws.amazon.com/gp/aws/manageYourAccount
Sign Up EMR

S3: Create S3 Bucket

• [https://console.aws.amazon.com/s3](https://console.aws.amazon.com/s3)
• Remember the bucket name
EC2: Create Key Pair

- [https://console.aws.amazon.com/ec2/](https://console.aws.amazon.com/ec2/)
You do not have any key pairs defined. Click the Create Key Pair button to download a new private key.
• The file “my-keypair.pem” will be downloaded automatically
• Change the file permission to 400
  $ chmod 400 ~/Downloads/my-keypair.pem
• Detailed documentation:
EMR: Create Job Flow

• https://console.aws.amazon.com/elasticmapreduce/
• Six Steps
• Pre-requisite
  – S3 Bucket created before
  – Key pair created before
Step 1

Start a dummy wordcount application
Step 2

Specify Mapper and Reducer functions to run within the Job Flow. The mapper and reducers may be either (i) class names referring to a mapper or reducer class in Hadoop or (ii) locations in Amazon S3. (Click Here for a list of available tools to help you upload and download files from Amazon S3.) The format for specifying a location in Amazon S3 is bucket_name/path_name. The location should point to an executable program, for example a python program. Extra arguments are passed to the Hadoop streaming program and can specify things such as additional files to be loaded into the distributed cache.

Input Location*: elasticmapreduce/samples/wordcount/input
Output Location*: hdfs://wordcount/output/2012-09-25
Mapper*: elasticmapreduce/samples/wordcount/wordSplitter.py
Reducer*: aggregate

Use your own S3 bucket here

Extra Args:

* Required field
Step 3

Configure the type and number of instances
Step 4

Use your key pair here so you can log into the machines later.
Step 5

Proceed with no Bootstrap Actions

I do not want to associate any Bootstrap Actions with this Job Flow.

**NOTE:** Bootstrap Actions must be associated with a Job Flow upon creation. You will not be able to add these later without creating a new Job Flow.

Configure your Bootstrap Actions

* Back  Continue  Required field
## Confirmation

Please review the details of your job flow and click "Create Job Flow" when you are ready to launch your Hadoop Cluster.

<table>
<thead>
<tr>
<th>Job Flow Name:</th>
<th>My Job Flow</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type:</td>
<td>Word Count (Streaming)</td>
</tr>
</tbody>
</table>

**Input Location:**
- s3n://elasticmapreduce/samples/wordcount/input

**Output Location:**
- s3n://jie.li/wordcount/output/2012-09-25

**Mapper:**
- s3n://elasticmapreduce/samples/wordcount/wordSplitter.py

**Reducer:**
- aggregate

**Extra Args:**

**Master Instance Type:**
- m1.small

**Core Instance Type:**
- m1.small

**Instance Count:**
- 1

**Instance Count:**
- 2

**Amazon EC2 Key Pair:**
- my-keypair

**Amazon Subnet Id:**

**Amazon S3 Log Path:**

**Enable Debugging:**
- No

**Keep Alive:**
- Yes

**Termination Protected:**
- No

**Bootstrap Actions:**
- No Bootstrap Actions created for this Job Flow

### Note:
- Once you click "Create Job Flow," instances will be launched and you will be charged accordingly.

[Create Job Flow]

[Edit Job Flow Definition]

[Edit Job Flow Parameters]

[Edit EC2 Configs]

[Edit Advanced Options]

[Edit Bootstrap Actions]
EMR: Check job flow

Wait for a few minutes

https://console.aws.amazon.com/elasticmapreduce/
Watch Out!

You’ll need to wait for a day or so. Don’t panic!

<table>
<thead>
<tr>
<th>State</th>
<th>Creation Date</th>
<th>Elapsed Time</th>
<th>Normalized Instance Hours</th>
</tr>
</thead>
<tbody>
<tr>
<td>STARTING</td>
<td>2012-09-25 18:26 EDT</td>
<td>0 hours 0 minutes</td>
<td>0</td>
</tr>
<tr>
<td>FAILED</td>
<td>2012-09-25 17:49 EDT</td>
<td>0 hours 0 minutes</td>
<td>0</td>
</tr>
</tbody>
</table>

Inge Reason: Unable to launch EC2 instances because your account is being reviewed by our team and verified as a valid new account.
The Machines are Ready!
EC2: Configure the Network Security

To start using Amazon EC2 you will want to launch a virtual server, known as an Amazon EC2 instance. Launch Instance

Note: Your instances will launch in the US East (N. Virginia) region.

Service Health

Service Status

<table>
<thead>
<tr>
<th>Current Status</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amazon EC2 (US East - N. Virginia)</td>
<td>Service is operating normally</td>
</tr>
</tbody>
</table>

Availability Zone Status

<table>
<thead>
<tr>
<th>Current Status</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>us-east-1a</td>
<td>Availability zone is operating normally</td>
</tr>
<tr>
<td>us-east-1b</td>
<td>Availability zone is operating normally</td>
</tr>
</tbody>
</table>

Related Links

- Getting Started Guide
- Documentation
- All EC2 Resources
- Find software on AWS MarketPlace
- Forums
Expose 9100 port
EMR: Get the Master Public DNS

```
<table>
<thead>
<tr>
<th>Name</th>
<th>State</th>
<th>Creation Date</th>
<th>Elapsed Time</th>
<th>Normalized Instance Hours</th>
</tr>
</thead>
<tbody>
<tr>
<td>My Job Flow</td>
<td>WAITING</td>
<td>2012-09-25 18:26 EDT</td>
<td>46 hours 4 minutes</td>
<td>141</td>
</tr>
</tbody>
</table>
```

State Change Reason: Waiting after step completed

```
| Master Public DNS Name: | ec2-204-236-208-31.compute-1.amazonaws.com |
```

Check the JobTracker Page

Visit http://<public_dns>:9100
e.g. http://ec2-204-236-208-31.compute-1.amazonaws.com:9100

ip-10-118-178-241 Hadoop Map/Reduce Administration

Quick Links

- Scheduling Info
- Running Jobs
- Retired Jobs
- Local Logs

State: RUNNING
Started: Tue Sep 25 22:44:27 UTC 2012
Version: 1.0.3, r2c6d566e34d439ac7c792ed9546c074c16cf74f4
Compiled: Sat Aug 25 00:10:28 UTC 2012 by Elastic MapReduce
Identifier: 201209252244

Cluster Summary (Heap Size is 25.38 MB/556.81 MB)

<table>
<thead>
<tr>
<th>Running Map Tasks</th>
<th>Running Reduce Tasks</th>
<th>Total Submissions</th>
<th>Nodes</th>
<th>Occupied Map Slots</th>
<th>Occupied Reduce Slots</th>
<th>Reserved Map Slots</th>
<th>Reserved Reduce Slots</th>
<th>Map Task Capacity</th>
<th>Reduce Task Capacity</th>
<th>Avg. Tasks/Node</th>
<th>Blacklisted Nodes</th>
<th>Graylisted Nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>2</td>
<td>3.00</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
SSH to the Master Node

- Get the public DNS and the keypair file

```bash
ssh -i my-keypair.pem hadoop@ec2-204-236-208-31.compute-1.amazonaws.com
```

```
jay23jack@w1-10-190-95-82: ~/Downloads$ ssh -i my-keypair.pem hadoop@ec2-204-236-208-31.compute-1.amazonaws.com
The authenticity of host 'ec2-204-236-208-31.compute-1.amazonaws.com (204.236.208.31)' can't be established.
Are you sure you want to continue connecting (yes/no)?
yes
Warning: Permanently added 'ec2-204-236-208-31.compute-1.amazonaws.com,204.235.208.31' (RSA) to the list of known hosts.
Linux (none) 2.6.35.11-83.9.amzn1.i686 #1 SMP Sat Feb 19 23:41:56 UTC 2011 i686

Welcome to Amazon Elastic MapReduce running Hadoop and Debian/Squeeze.

Hadoop is installed in /home/hadoop. Log files are in /mnt/var/log/hadoop. Check /mnt/var/log/hadoop/steps for diagnosing step failures.

The Hadoop UI can be accessed via the following commands:

- `JobTracker` lynx http://localhost:9100/
- `NameNode` lynx http://localhost:9101/

hadoop@ip-10-118-178-241:~$ 
```
Launch Hadoop Jobs

Hadoop is installed in /home/hadoop. Log files are in /mnt/var/log/hadoop. Check /mnt/var/log/hadoop/steps for diagnosing step failures.

The Hadoop UI can be accessed via the following commands:

- JobTracker: lynx http://localhost:9000/
- NameNode: lynx http://localhost:9101/

```
hadoop@ip-10-118-178-241:~$ hadoop jar hadoop-examples.jar sleep 2
12/09/27 20:58:55 INFO mapred.JobClient: Default number of map tasks: 1
12/09/27 20:58:55 INFO mapred.JobClient: Default number of reduce tasks: 1
12/09/27 20:58:56 INFO mapred.JobClient: Setting group to hadoop
12/09/27 20:58:57 INFO mapred.JobClient: map 0% reduce 0%
12/09/27 20:59:18 INFO mapred.JobClient: map 100% reduce 0%
12/09/27 20:59:30 INFO mapred.JobClient: map 100% reduce 33%
12/09/27 20:59:33 INFO mapred.JobClient: map 100% reduce 100%
12/09/27 20:59:38 INFO mapred.JobClient: Job Counters
12/09/27 20:59:38 INFO mapred.JobClient: Launched reduce tasks=1
12/09/27 20:59:38 INFO mapred.JobClient: Total time spent by all reduces waiting after reserving slots (ms)=0
12/09/27 20:59:38 INFO mapred.JobClient: Total time spent by all maps waiting after reserving slots (ms)=0
12/09/27 20:59:38 INFO mapred.JobClient: Launched map tasks=1
12/09/27 20:59:38 INFO mapred.JobClient: Bytes Read=0
12/09/27 20:59:38 INFO mapred.JobClient: Bytes Written=0
12/09/27 20:59:38 INFO mapred.JobClient: FILE_BYTES_READ=0
```
Store Data on S3

- You can write and read data on S3
- The data will be on S3 so you can shutdown EMR without losing data
- Use `s3://<bucket_name>/path` for S3 paths, e.g.
  
  ```bash
  $ hadoop jar hadoop-examples.jar teragen 100 s3://jieli/tmp/teragen
  ```
Terminate Job Flow

Terminate when you don’t need it for the short term

429 roughly means $42.9
Check Your Budget

This Month's Activity as of October 1, 2012

The statement period for this report is October 1 - October 31, 2012. The charges on this page currently show activity through approximately 10/01/2012 15:43 GMT.

Select a different statement: Current Statement

<table>
<thead>
<tr>
<th>Summary</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>AWS Service Charges</td>
<td>$5.25</td>
</tr>
<tr>
<td>Usage Charges and Recurring Fees</td>
<td>$5.25</td>
</tr>
</tbody>
</table>

Total new charges for this statement $5.25

All charges this month will be paid for by AWS Account 4957-2372-8936.

Details

<table>
<thead>
<tr>
<th>AWS Service Charges</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Amazon Elastic Compute Cloud</td>
<td>$4.08</td>
</tr>
<tr>
<td>Amazon SimpleDB</td>
<td>$0.00</td>
</tr>
</tbody>
</table>

Clean Up

• When you finish the class, your account will be removed from consolidated bill. You’ll need to:
  – Shut down any EC2/EMR machine
  – Delete any S3 bucket
Enjoy!

Question? Email to jieli@cs.duke.edu