

## Aims

This exercise aims to get you to practice:

- Executing Spark step on **AWS EMR**
- Using Spark cluster on AWS EMR

## Package the WordCount Jar File

1. Create a folder wcpark in your home folder:

```
$ mkdir -p wcpark/src/main/scala/comp9313/lab10
```

2. Download the WordCount.scala file in the created folder at:

<https://webcms3.cse.unsw.edu.au/COMP9313/18s1/resources/16603>

3. Change directory to ~/wcpark, create a file wc.sbt, and add the following contents:

```
name := "Word Count"
version := "1.0"
scalaVersion := "2.11.8"
libraryDependencies += "org.apache.spark" %% "spark-core" % "2.3.0"
```

4. Use sbt to package your project

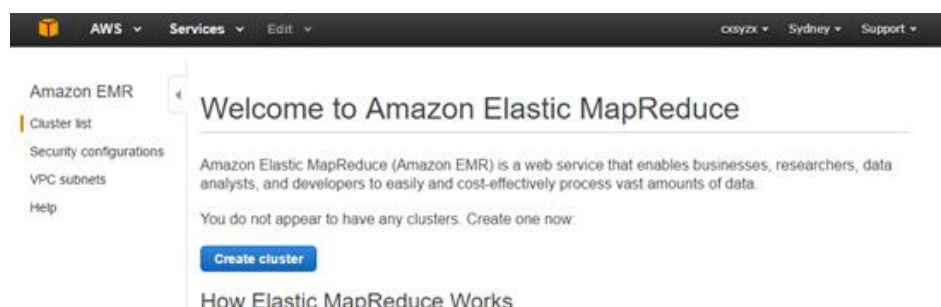
```
$ sbt package
```

The generated jar file is located at: **~/wcpark/target/scala-2.11/word-count\_2.11-1.0.jar**

5. Upload this jar file to your S3 bucket: **S3://comp9313.<YOUR\_StudentID>/**

## Run Spark Tasks on AWS EMR (Part 1)

1. Go to the AWS Management console and open the EMR console.



2. At the top right corner, select the region as “Sydney”
3. Click Create cluster. On the Create Cluster page, you need to do the following:

**In General Configuration section:**

- a) Cluster name: **comp9313.lab10**
- b) Logging: **Select**
- c) S3 folder: use default. The folder is used to store the logs.
- d) Launch mode: select **“Step execution.”**

**In Add steps section:**

**Add step**

**Step type** Spark application Run Spark application using spark-submit. [Learn more](#)

**Name** Spark application

**Deploy mode** Cluster Run your driver on a slave node (cluster mode) or on the master node as an external client (client mode). Specify other options for spark-submit.

**Spark-submit options** `--class "comp9313.lab10.WordCount"`

**Application location\*** `s3://comp9313.z3515164/word-count_2.11-1.0.jar` Path to a JAR with your application and dependencies (client deploy mode only supports a local path). Specify optional arguments for your application.

**Arguments** `s3://comp9313.z3515164/pg100.txt`  
`s3://comp9313.z3515164/output`

**Action on failure** Terminate cluster What to do if the step fails.

[Cancel](#) [Add](#)

Set the step type as Spark application, and then click “Configure”

- a) Set Name as “WordCount”
- b) Set Deploy mode as Cluster
- c) Spark-submit options: `--class “comp9313.lab10.WordCount” --master yarn`
- d) Application location: click the folder icon, select the jar file you uploaded to your s3 folder, `“s3://comp9313.<YOUR_StudentID>/ word-count_2.11-1.0.jar”`



e) Set Arguments as “s3://comp9313.<YOUR\_StudentID>/pg100.txt  
s3://comp9313.<YOUR\_StudentID>/output”

**Remember that the output folder cannot be an existing folder in your S3.**

f) Select “Terminate cluster” for Action on Failure, and finally click Add.

Then, in the Add steps section, you can see:

**Add steps**  
A step is a unit of work submitted to an application running on your EMR cluster. EMR programmatically installs the applications needed to execute the added steps. [Learn more](#)

Name	Action on failure	JAR location	Arguments		
Spark application	Terminate cluster	command-runner.jar	spark-submit --deploy-mode cluster --class comp9313.lab9.WordCount s3://comp9313.z3515164/word-count_2.11-1.0.jar s3://comp9313.z3515164/pg100.txt s3://comp9313.z3515164/output		

**Step type** Spark application **Configure**

**In Software configuration section:**

a) Release: Select the most recent version, emr-5.13.0

b) Application: Select Spark 2.3.0 on Hadoop Yarn

**In the Hardware Configuration section:**

a) Instance type: use m3.xlarge

b) Number of instances: 3

**In the Security and Access section:**

Accept the remaining default options.

6. Choose Create cluster. You should see:

Clone
Terminate
AWS CLI export

Cluster: comp9313.lab10
Starting


Summary
Application history
Monitoring
Hardware
Events
Steps
Configurations
Bootstrap actions

Connections: --
Master public DNS: --
Tags: -- [View All / Edit](#)

**Summary**

ID: j-FD3OI69OL7JY  
Creation date: 2018-05-14 11:52 (UTC+10)  
Elapsed time:  
Auto-terminate: Yes  
Termination protection: Off [Change](#)

**Configuration details**

Release label: emr-5.13.0  
Hadoop distribution: Amazon 2.8.3  
Applications: Spark 2.3.0  
Log URI: s3://aws-logs-206517396822-ap-southeast-2/elasticmapreduce/   
EMRFS consistent view: Disabled  
Custom AMI ID: --

**Network and hardware**

Availability zone: --  
Subnet ID: [subnet-749f5002](#)  
Master: Provisioning 1 m3.xlarge  
Core: Provisioning 2 m3.xlarge  
Task: --

**Security and access**

Key name: --  
EC2 instance profile: EMR\_EC2\_DefaultRole  
EMR role: EMR\_DefaultRole  
Visible to all users: All [Change](#)  
Security groups for Master:  
Security groups for Core & Task:

Later, you will see the information for Connections and Master public DNS is updated, since the cluster is already started.

7. Wait until the WordCount task is finished. **Note that this may take several minutes.**

In the meantime, you can begin working on the next section, and go back to check the results later.

8. If the **task is completed**, you should see:

Clone
Terminate
AWS CLI export

Cluster: comp9313.lab10
Terminated
Steps completed

Summary
Application history
Monitoring
Hardware
Events
Steps
Configurations
Bootstrap actions

Connections: --
Master public DNS: ec2-13-211-172-85.ap-southeast-2.compute.amazonaws.com [SSH](#)
Tags: --

**Summary**

**Configuration details**

Go to your S3 bucket, the results should be stored there. You can see that several files are in the folder. Spark automatically computes the number of partitions for you already, and each partition corresponds to one result file. You can download these files to check the results.

Amazon S3 > comp9313.z3515164 / output

Overview

Q Type a prefix and press Enter to search. Press ESC to clear.

Upload Create folder More

Asia Pacific (Sydney)

Viewing 1 to 33

<input type="checkbox"/>	Name	Last modified	Size	Storage class
<input type="checkbox"/>	_SUCCESS	May 14, 2018 12:20:48 PM GMT+1000	0 B	Standard
<input type="checkbox"/>	part-00000	May 14, 2018 12:20:47 PM GMT+1000	26.1 KB	Standard
<input type="checkbox"/>	part-00001	May 14, 2018 12:20:47 PM GMT+1000	25.5 KB	Standard
<input type="checkbox"/>	part-00002	May 14, 2018 12:20:47 PM GMT+1000	25.8 KB	Standard
<input type="checkbox"/>	part-00003	May 14, 2018 12:20:47 PM GMT+1000	26.7 KB	Standard

## Run Spark Tasks on AWS EMR (Part 2)

In this section, we will **ssh to the cluster** to do a Spark job.

1. Choose Create cluster. On the Create Cluster page, click “Go to advanced options”.
2. In Step 1, select emr-5.13.0 for Release, and select “Hadoop 2.8.3” and “Spark 2.3.0” in the cluster. Accept the other default configurations, and click “Next”.
3. In Step 2, select the default m3.xlarge as the instance type for both Master and Core. Accept all the other default configurations, and click “Next”
4. In Step 3, name your cluster and accept all default configurations and click “Next”.
5. In Step 4, use your key pair “comp9313” for the cluster. Click “EC2 Security Groups”, configure the security groups for both Master and Core as “launch-wizard-1”. Finally, click “Create Cluster”.

If you lost your key pair file “comp9313.pem”, please follow the lab9 instructions and create one key pair again. launch-wizard-1 is created in lab9 as well. If you cannot see it, please follow the EC2 instructions in lab9.

6. Waiting for the starting of the cluster. You can go back to check the results of your first cluster.

Once the information for “Connection” and “Master public DNS” is updated, your cluster is started, and you can ssh to the master node now.

Cluster: My cluster **Starting** Configuring cluster software

Summary Application history Monitoring Hardware Events Steps Configurations Bootstrap actions

Connections: [Enable Web Connection](#) – Spark History Server, Resource Manager ... (View All)

Master public DNS: ec2-13-211-32-67.ap-southeast-2.compute.amazonaws.com [SSH](#)

Tags: -- [View All](#) / [Edit](#)

Click SSH in the line of “Master public DNS:”, you will see:

SSH

Connect to the Master Node Using SSH

You can connect to the Amazon EMR master node using SSH to run interactive queries, examine log files, submit Linux commands, and so on. [Learn more.](#)

Windows Mac / Linux

1. Open a terminal window. On Mac OS X, choose Applications > Utilities > Terminal. On other Linux distributions, terminal is typically found at Applications > Accessories > Terminal.
2. To establish a connection to the master node, type the following command. Replace ~/comp9313.pem with the location and filename of the private key file (.pem) used to launch the cluster.

```
ssh -i ~/comp9313.pem hadoop@ec2-13-211-32-67.ap-southeast-2.compute.amazonaws.com
```

3. Type yes to dismiss the security warning.

Close

SSH to the master node by copying the command as shown in the dialog:

```
$ ssh -i ~/comp9313.pem hadoop@YOUR_INSTANCE
```

```
comp9313@comp9313-VirtualBox:~/wccspark$ ssh -i ~/comp9313.pem hadoop@ec2-13-211-32-67.ap-southeast-2.compute.amazonaws.com
The authenticity of host 'ec2-13-211-32-67.ap-southeast-2.compute.amazonaws.com (13.211.32.67)' can't be established.
ECDSA key fingerprint is ea:74:6f:88:67:ff:97:9d:cd:dc:dd:2c:02:4c:29:2a.
Are you sure you want to continue connecting (yes/no)?
```

Enter “yes” to connect to the cluster

7. Download the jar file from S3 by the following command:

```
$ hadoop fs -get s3://comp9313.<YOUR_StudentID>/word-count_2.11-1.0.jar
```

8. Run the Spark task. Generate the results in a different folder!

```
$ spark-submit --class "comp9313.lab10.WordCount" --master yarn word-count_2.11-1.0.jar s3://comp9313.<YOUR_StudentID>/pg100.txt
s3://comp9313.<YOUR_StudentID>/output2
```

9. Wait for the completion of the task and check the results in your S3 bucket. You should see:

Amazon S3 > comp9313.z3515164 / output2

Overview

🔍 Type a prefix and press Enter to search. Press ESC to clear.

Upload
 Create folder
 More

Asia Pacific (Sydney)

Viewing 1 to 3			
<input type="checkbox"/> Name	Last modified	Size	Storage class
<input type="checkbox"/> _SUCCESS	May 14, 2018 12:33:04 PM GMT+1000	0 B	Standard
<input type="checkbox"/> part-00000	May 14, 2018 12:33:04 PM GMT+1000	417.8 KB	Standard
<input type="checkbox"/> part-00001	May 14, 2018 12:33:04 PM GMT+1000	418.6 KB	Standard

Viewing 1 to 3

10. You can also download “pg100.txt” from S3, and put the file to HDFS, and run the Spark task by reading/writing files from/to HDFS instead of S3.

```
$ hdfs dfs -mkdir input
```

```
$ hdfs dfs -put pg100.txt input
```

```
$ spark-submit --class "comp9313.lab10.WordCount" --master yarn word-count_2.11-1.0.jar input/pg100.txt output
```

Local

**Caution:** The I/O between the cluster and S3 is also billed if your transfer exceeds the free tier limit!!!

11. You can also add a new step to this cluster to run a Spark task. Try it by yourself.

**12. Caution: Do not forget to terminate the cluster after you finish all labs!!! (Click “Terminate” and turn termination protection off)**