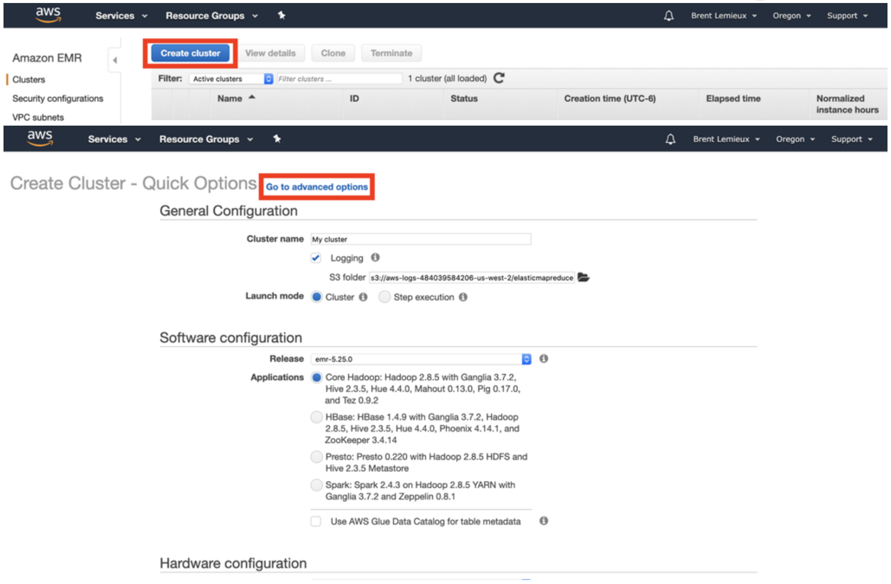
BU.330.740 Large Scale Computing with Hadoop

Lab 4. Simple text mining using Python Spark

Learning Goal: try Bag of Words model (aka. wordcount) in Python Spark on AWS Hadoop cluster and Jupyter notebook. Refer to lecture 4 notes on page 36-37.
Required Skills: understand basic Spark syntax

1. **Prepare:** download an e-book *Pride and Prejudice* by Jane Austen in txt format from http://www.gutenberg.org/ebooks/1342 and save in your local folder.
2. **S3:** login into AWS Educate account, go to AWS Management Console, create a S3 bucket for lab4. Create 1 folder for input text file. Upload the e-book into input text file folder, using all the default settings.
3. **EC2:** if you have an EC2 key pair, you can reuse it. Otherwise, create a new key pair.
4. **EMR:** Be careful this step is slightly different from Lab 2 and Lab 3. DO NOT CLONE the previous cluster.

   Go to EMR, click **Create Cluster**, then **Go to advanced options**.



Filter: Active clusters | Filter clusters ... | 1 cluster (all loaded)

Create Cluster - Quick Options

General Configuration

Cluster name: My cluster

☑ Logging

S3 folder: s3://aws-logs-484039584206-us-west-2/elasticmapreduce

Launch mode: ● Cluster ○ Step execution

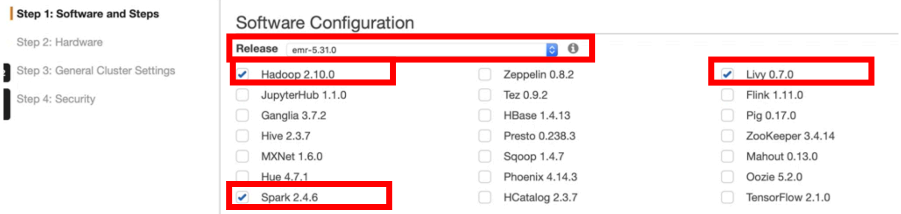
Software configuration

Release: emr-5.25.0

Applications:
- ● Core Hadoop: Hadoop 2.8.5 with Ganglia 3.7.2, Hive 2.3.5, Hue 4.4.0, Mahout 0.13.0, Pig 0.17.0, and Tez 0.9.2
- ○ HBase: HBase 1.4.9 with Ganglia 3.7.2, Hadoop 2.8.5, Hive 2.3.5, Hue 4.4.0, Phoenix 4.14.1, and ZooKeeper 3.4.14
- ○ Presto: Presto 0.220 with Hadoop 2.8.5 HDFS and Hive 2.3.5 Metastore
- ○ Spark: Spark 2.4.3 on Hadoop 2.8.5 YARN with Ganglia 3.7.2 and Zeppelin 0.8.1

☐ Use AWS Glue Data Catalog for table metadata

At Step 1, in **Release emr-5.31.0** section, make the following selections (Hadoop, Spark and Livy):



Step 1: Software and Steps
Step 2: Hardware
Step 3: General Cluster Settings
Step 4: Security

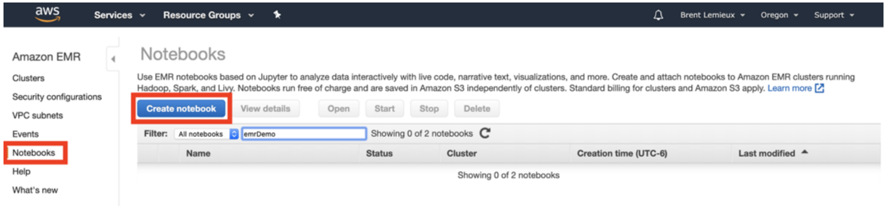
Software Configuration

Release: emr-5.31.0

- ☑ Hadoop 2.10.0
- ☐ JupyterHub 1.1.0
- ☐ Ganglia 3.7.2
- ☐ Hive 2.3.7
- ☐ MXNet 1.6.0
- ☐ Hue 4.7.1
- ☑ Spark 2.4.6
- ☐ Zeppelin 0.8.2
- ☐ Tez 0.9.2
- ☐ HBase 1.4.13
- ☐ Presto 0.238.3
- ☐ Sqoop 1.4.7
- ☐ Phoenix 4.14.3
- ☐ HCatalog 2.3.7
- ☑ Livy 0.7.0
- ☐ Flink 1.11.0
- ☐ Pig 0.17.0
- ☐ ZooKeeper 3.4.14
- ☐ Mahout 0.13.0
- ☐ Oozie 5.2.0
- ☐ TensorFlow 2.1.0

Click **Next**. Do not change anything at Step 2, and click **Next**. Name your cluster at Step 3, then click **Next**. Select the key pair you have in Step 4 and click **Create cluster**. Your cluster will take a few minutes to start, until you see the status turns to **Waiting**.

5. **Add step:** We will use Jupyter notebook to execute Python Spark scripts. The Jupyter notebook will run on the Hadoop cluster we just created.
On EMR console, navigate to **Notebooks** in the left panel. Click **Create notebook**.



Notebooks

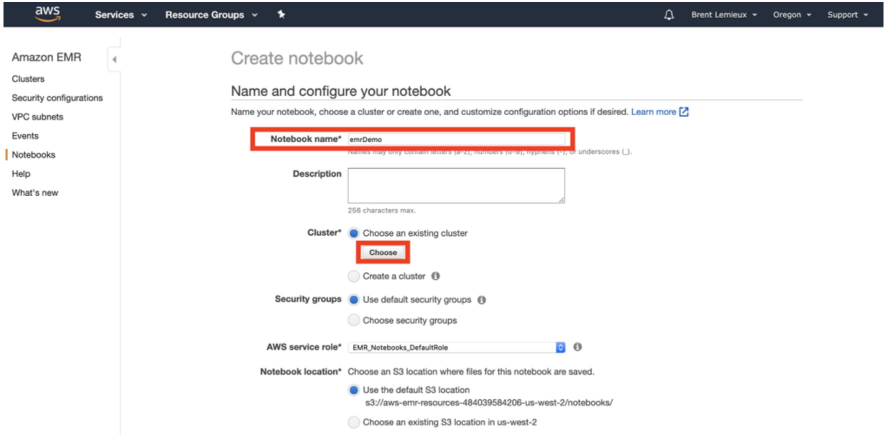
Use EMR notebooks based on Jupyter to analyze data interactively with live code, narrative text, visualizations, and more. Create and attach notebooks to Amazon EMR clusters running Hadoop, Spark, and Livy. Notebooks run free of charge and are saved in Amazon S3 independently of clusters. Standard billing for clusters and Amazon S3 apply. Learn more

Filter: All notebooks emrDemo Showing 0 of 2 notebooks

Name	Status	Cluster	Creation time (UTC-6)	Last modified

Showing 0 of 2 notebooks

Name your notebook and choose the cluster you just created.



Create notebook

Name and configure your notebook

Name your notebook, choose a cluster or create one, and customize configuration options if desired. Learn more

Notebook name* emrDemo
Names may only contain letters (a-z), numbers (0-9), hyphens (-), or underscores (_).

Description

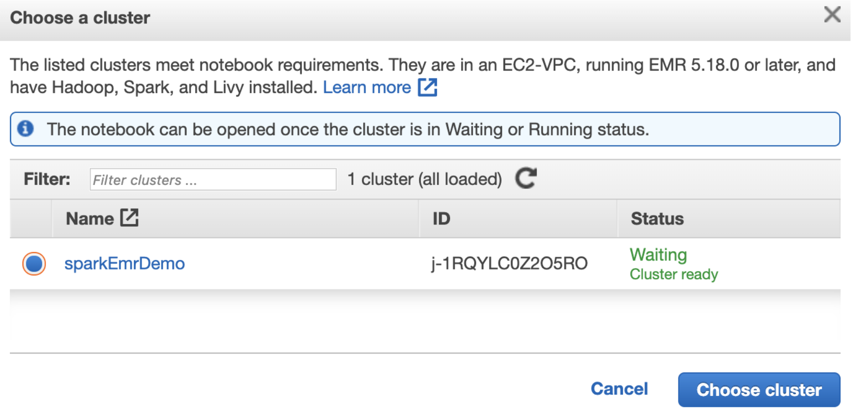
256 characters max.

Cluster* ● Choose an existing cluster
○ Create a cluster

Security groups ● Use default security groups
○ Choose security groups

AWS service role* EMR_Notebooks_DefaultRole

Notebook location* Choose an S3 location where files for this notebook are saved.
● Use the default S3 location
s3://aws-emr-resources-484039584206-us-west-2/notebooks/
○ Choose an existing S3 location in us-west-2



Choose a cluster

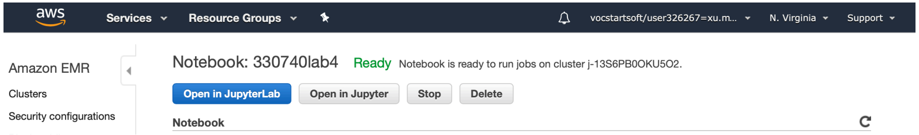
The listed clusters meet notebook requirements. They are in an EC2-VPC, running EMR 5.18.0 or later, and have Hadoop, Spark, and Livy installed. Learn more

The notebook can be opened once the cluster is in Waiting or Running status.

Filter: Filter clusters ... 1 cluster (all loaded)

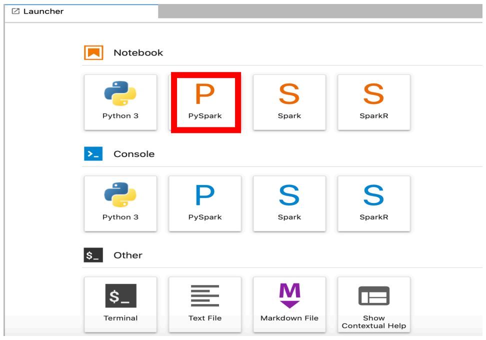
	Name	ID	Status
●	sparkEmrDemo	j-1RQYLC0Z2O5RO	Waiting Cluster ready

Then click **Create notebook**. Wait till the status shows **Ready**. Click Open in **JupyterLab**.



Notebook: 330740lab4 Ready Notebook is ready to run jobs on cluster j-13S6PB0OKU5O2.

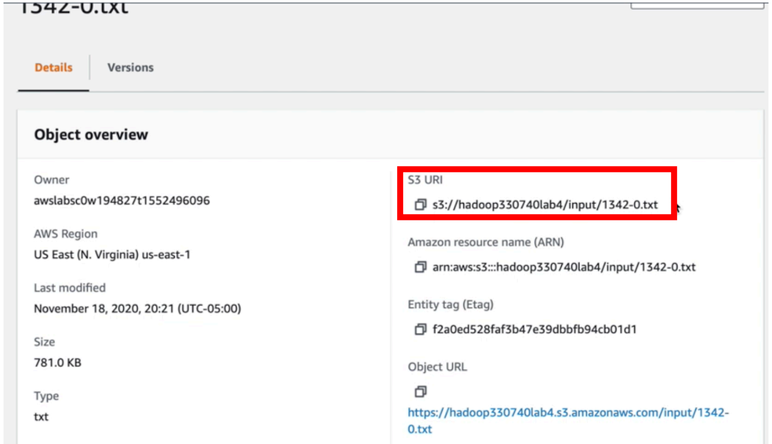
Choose **Launcher->Notebook->PySpark**.



In the notebook, type in the following scripts:

```
text_file = sc.textFile("s3://bucket/folder/ebook.txt")
counts = text_file.flatMap(lambda line: line.split(" ")) \
             .map(lambda word: (word, 1)) \
             .reduceByKey(lambda a, b: a + b)
counts.saveAsTextFile("s3://bucket/output")
```

Replace the red part with your actual S3 path. Go to S3, select your S3 bucket, then your input folder, and your e-book file. Copy the **S3 URI,** and replace the text file path in red. The output path should be a folder in your S3 bucket. Note that I did not create this output folder in step 2.



1342-0.txt

Details | Versions

Object overview

Owner
awslabsc0w194827t1552496096

AWS Region
US East (N. Virginia) us-east-1

Last modified
November 18, 2020, 20:21 (UTC-05:00)

Size
781.0 KB

Type
txt

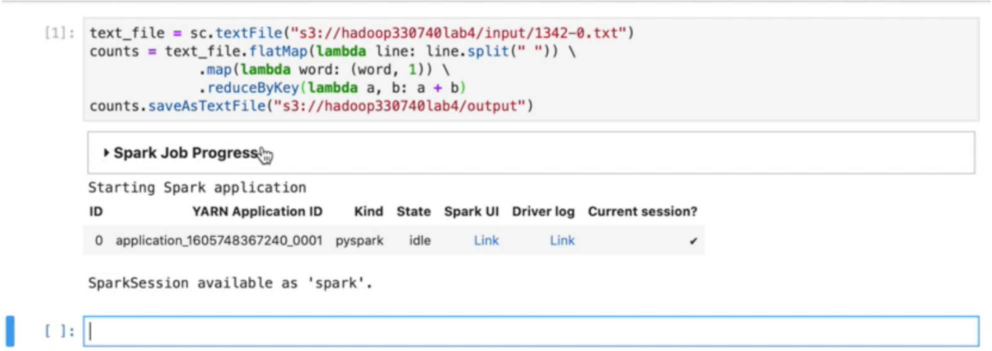
S3 URI
s3://hadoop330740lab4/input/1342-0.txt

Amazon resource name (ARN)
arn:aws:s3:::hadoop330740lab4/input/1342-0.txt

Entity tag (Etag)
f2a0ed528faf3b47e39dbbfb94cb01d1

Object URL
https://hadoop330740lab4.s3.amazonaws.com/input/1342-0.txt

Here is my notebook screen:



```
[1]: text_file = sc.textFile("s3://hadoop330740lab4/input/1342-0.txt")
     counts = text_file.flatMap(lambda line: line.split(" ")) \
                  .map(lambda word: (word, 1)) \
                  .reduceByKey(lambda a, b: a + b)
     counts.saveAsTextFile("s3://hadoop330740lab4/output")
```

▸ Spark Job Progress

```
Starting Spark application
```

ID	YARN Application ID	Kind	State	Spark UI	Driver log	Current session?
0	application_1605748367240_0001	pyspark	idle	Link	Link	✓

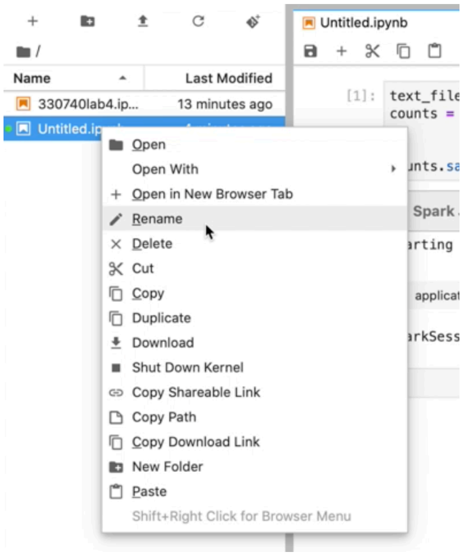
```
SparkSession available as 'spark'.
```

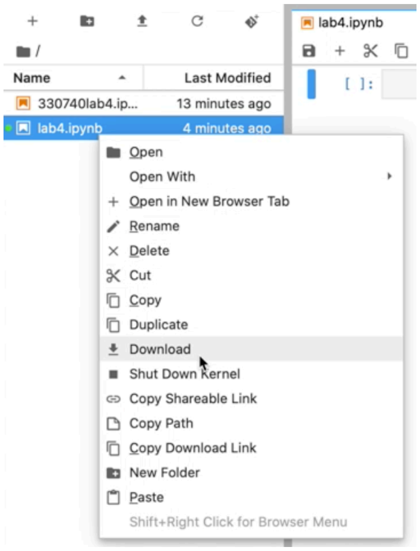
```
[ ]: |
```

Note that the input is an e-book, so it takes time to complete Spark application.

6. After the application is done, you can check and download the results in your output folder in S3.

You can also rename and download the .ipynb file for your record.





7. **DO NOT FORGET TO CLEAN UP RESOURCSES!!** Stop the notebook. Terminate the cluster (If terminate protection is on, turn it off and then terminate the cluster). Delete all S3 buckets under your account. And always double check.

Now that you are done with lab 4. You can go ahead to Assignment 2.

Reference:
https://towardsdatascience.com/getting-started-with-pyspark-on-amazon-emr-c85154b6b921