CLOUD ANALYTICS

- Cloud Computing Overview
- AWS Services
- Amazon EMR
- Amazon EMR Exercise

AWS ACCOUNT

- URL: https://ets-apps.ucsd.edu/dse230_sp21-custom-aws/
- Log in using UCSD Active Directory credentials

EMR NOTEBOOKS

- Jupyter notebooks on Amazon EMR
- Pre-configured with following kernels:
  - PySpark
  - Python3
  - Spark
  - SparkR
- Note
  - EMR notebook contents are saved in S3, separate from EMR cluster that executes code
  - Attach cluster to notebook to execute

OUTLINE OF STEPS FOR EMR

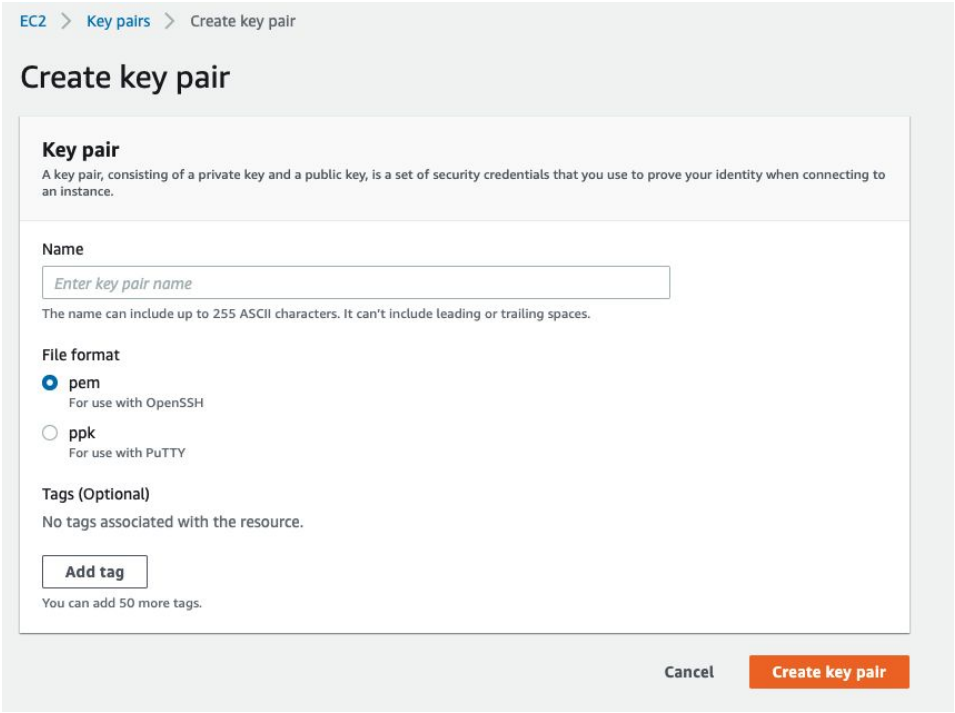
- Create EC2 key pair
  - Security credentials for connecting to EC2 instance
- Create S3 bucket
  - Store notebooks and instance logs
- Create EMR cluster
  - Specify software configuration
  - Add step to copy data from public S3 bucket for class
  - Specify number of master and worker nodes
  - Specify EC2 key pair
- Create EMR notebook
  - Connect to cluster
  - Specify notebook location
  - Create notebook
  - Read data from your S3 bucket or from public S3 bucket

OUTLINE OF STEPS FOR EMR

- Create EC2 key pair
  - Security credentials for connecting to EC2 instance
- Create S3 bucket
  - Store notebooks and instance logs
- Create EMR cluster
  - Specify software configuration
  - Add step to copy data from public S3 bucket for class
  - Specify number of master and worker nodes
  - Specify EC2 key pair
- Create EMR notebook
  - Connect to cluster
  - Specify notebook location
  - Create notebook
  - Read data from your S3 bucket or from public S3 bucket

CREATE KEY PAIR

- Services -> EC2 (under Compute)
- Under ‘Network & Security’, click on ‘Key Pairs’
- Click on ‘Create key pair’
- Enter name. Click on ‘Create key pair’
- Save <key-pair-name>.pem on your computer



OUTLINE OF STEPS FOR EMR

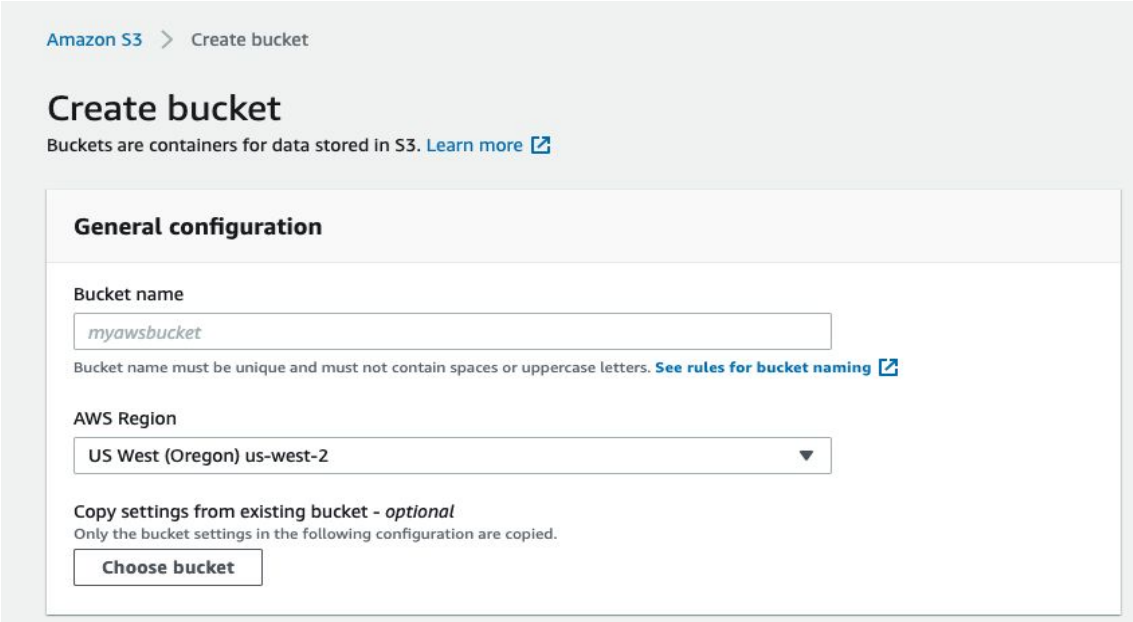
- Create EC2 key pair
  - Security credentials for connecting to EC2 instance
- Create S3 bucket
  - To store notebooks and instance logs
- Create EMR cluster
  - Specify software configuration
  - Add step to copy data from public S3 bucket for class
  - Specify number of master and worker nodes
  - Specify EC2 key pair
- Create EMR notebook
  - Connect to cluster
  - Specify notebook location
  - Create notebook
  - Read data from your S3 bucket or from public S3 bucket

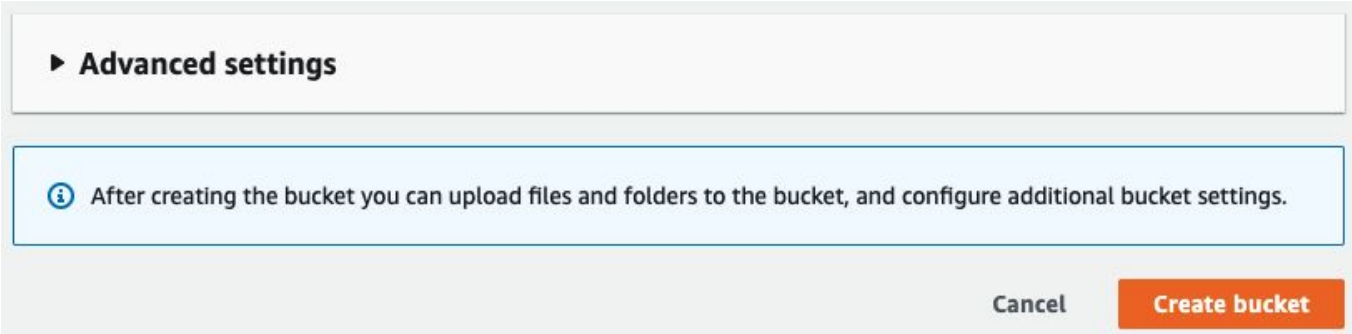
CREATE S3 BUCKET

- Services -> S3 (under Storage)
- Bucket name
  - Name must be unique across all S3 buckets
  - At least 3 characters and no more than 63 characters long
  - Cannot contain uppercase characters or underscores
  - Must start with lowercase letter or number
  - Cannot change name after bucket has been created
  - More on bucket naming
    - https://docs.aws.amazon.com/AmazonS3/latest/dev//BucketRestrictions.html#bucketnamingrules
- Click ‘Create Bucket’

CREATE S3 BUCKET

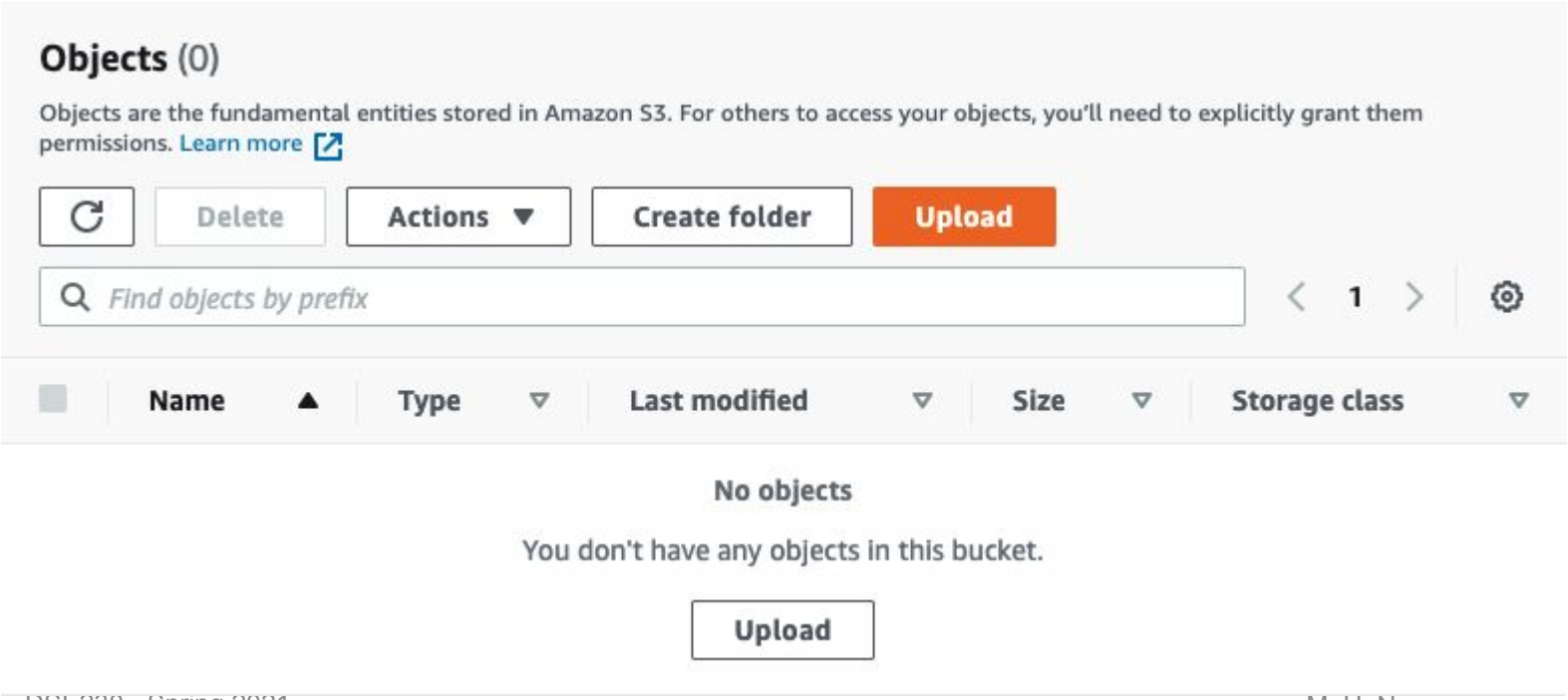
- Enter bucket name
- Click ‘Create bucket’





CREATE FOLDER IN BUCKET

- Create folder in S3 bucket
  - Click on bucket name
  - Select ‘Create folder’
  - Enter name for folder. Click ‘Create folder’
  - Create folder for EMR notebooks called ‘Notebooks’



UPLOAD DATA FILES TO S3

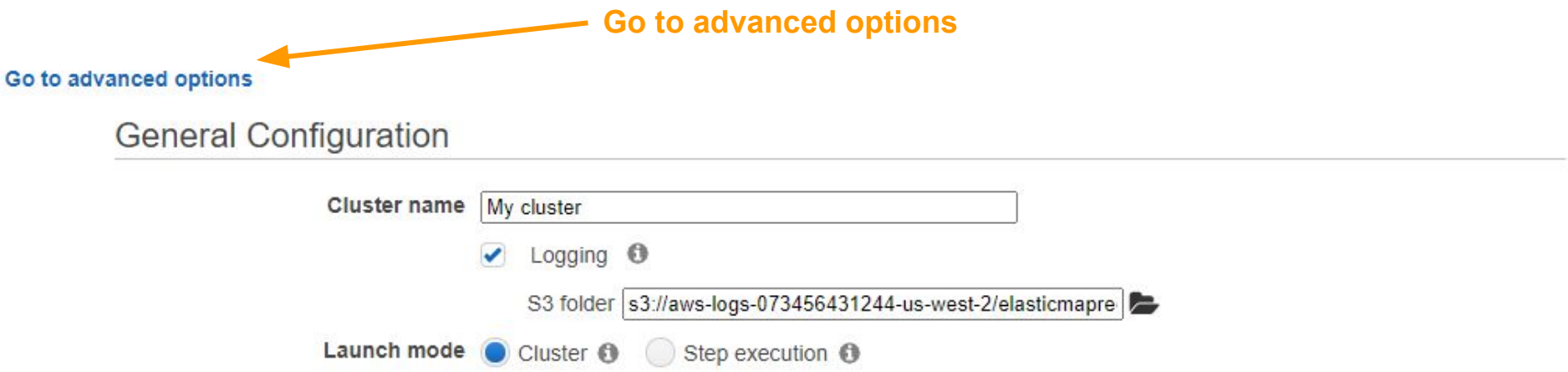
- Services -> S3
- Click into bucket and folder
- Click ‘Upload’
- Select file to upload
  - 2 ways
    - Drag and Drop
    - Click on ‘Add files’ and select file
- Click ‘Upload’
- Upload progress is shown at bottom of webpage.
  - Large files may take several minutes.

OUTLINE OF STEPS FOR EMR

- Create EC2 key pair
  - Security credentials for connecting to EC2 instance
- Create S3 bucket
  - Upload data to bucket
- Create EMR cluster
  - Specify software configuration
  - Specify number of master and worker nodes
  - Specify EC2 key pair
- Create EMR notebook
  - Connect to cluster
  - Specify notebook location
  - Create notebook
  - Read data from your S3 bucket or from public S3 bucket

CREATE CLUSTER

- Services -> EMR
- Click on ‘Create cluster’
- Click on ‘Go to advanced options’



CREATE CLUSTER - SOFTWARE CONFIGURATION

Select emr-6.2.0
Then select software options shown here

Software Configuration

Release emr-6.2.0

- [x] Hadoop 3.2.1
- [ ] JupyterHub 1.1.0
- [ ] Ganglia 3.7.2
- [ ] Hive 3.1.2
- [ ] ZooKeeper 3.4.14
- [ ] Sqoop 1.4.7
- [ ] Oozie 5.2.0
- [ ] TensorFlow 2.3.1
- [ ] Zeppelin 0.9.0
- [ ] Tez 0.9.2
- [ ] HBase 2.2.6-amzn-0
- [ ] Presto 0.238.3
- [x] JupyterEnterpriseGateway 2.1.0
- [ ] Hue 4.8.0
- [x] Spark 3.0.1
- [ ] Livy 0.7.0
- [ ] Flink 1.11.2
- [ ] Pig 0.17.0
- [ ] PrestoSQL 343
- [ ] MXNet 1.7.0
- [ ] Phoenix 5.0.0
- [ ] HCatalog 3.1.2

CREATE CLUSTER - CONFIGURE NODES

- Specify instance count for Master and Core (worker)
- https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-master-core-task-nodes.html

Cluster Nodes and Instances

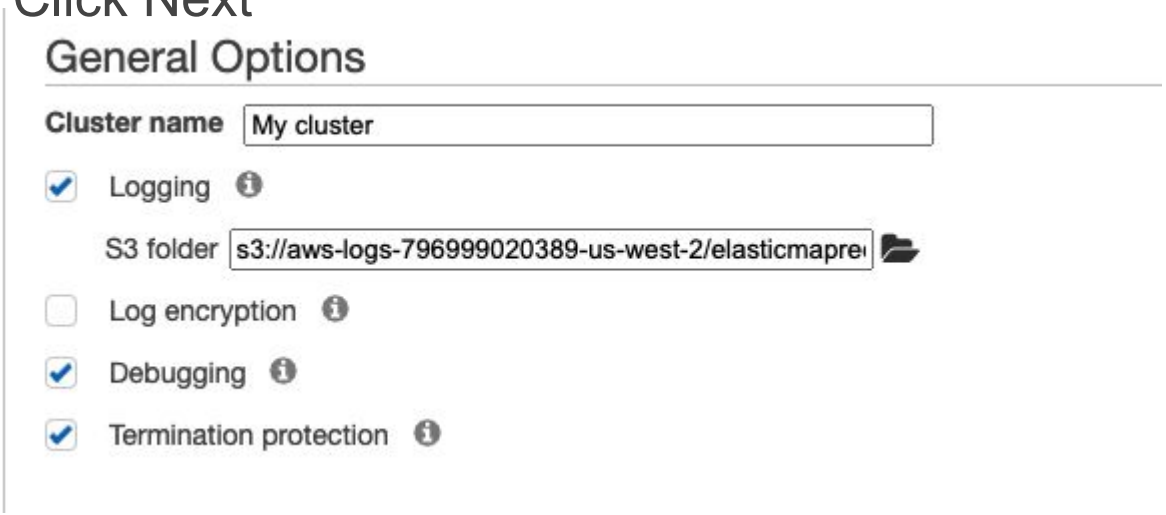
Choose the instance type, number of instances, and a purchasing option. Learn more about instance purchasing options

Console options for automatic scaling have changed. Learn more

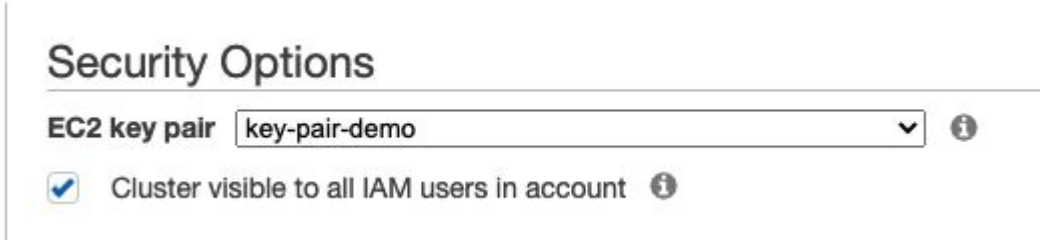
Node type	Instance type	Instance count	Purchasing option
Master Master - 1	m5.xlarge 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 64 GiB Add configuration settings	1 Instances	On-demand Spot Use on-demand as max price
Core Core - 2	m5.xlarge 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 64 GiB Add configuration settings	1 Instances	On-demand Spot Use on-demand as max price
Task Task - 3	m5.xlarge 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 64 GiB Add configuration settings	0 Instances	On-demand Spot Use on-demand as max price

CREATE CLUSTER - NAME AND EC2 KEY PAIR

- Enter name for ‘Cluster name’
- Click Next

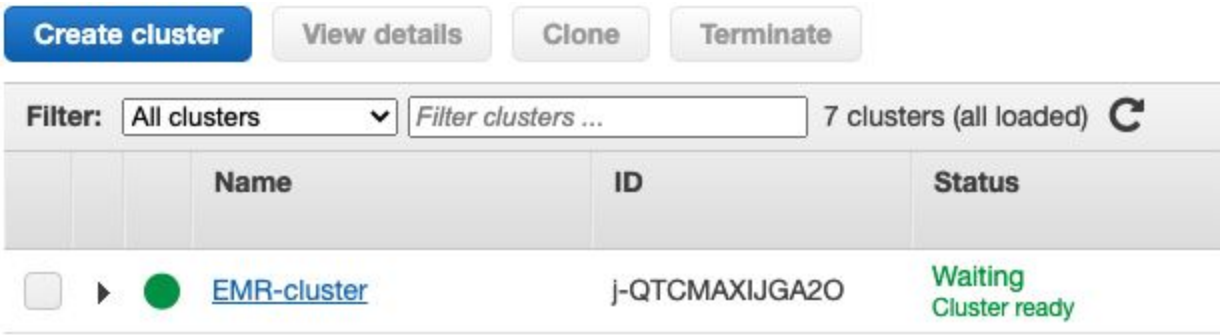


- Select EC2 key pair created earlier
- Click ‘Create cluster’



CREATE CLUSTER

- Creating cluster may take several minutes
  - To provision resources for master and worker nodes
- Click on Steps tab to monitor status of cluster creation and view logs
- Cluster is ready when Status shows ‘Waiting. Cluster ready’

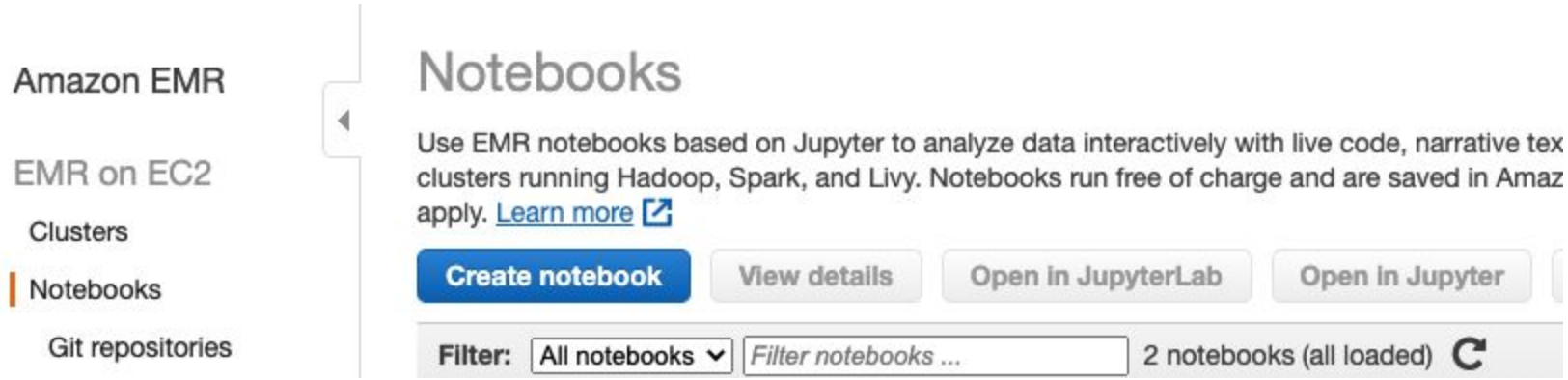


OUTLINE OF STEPS FOR EMR

- Create EC2 key pair
  - Security credentials for connecting to EC2 instance
- Create S3 bucket
  - Upload data to bucket
- Create EMR cluster
  - Specify software configuration
  - Add step to copy data from public S3 bucket for class
  - Specify number of master and worker nodes
  - Specify EC2 key pair
- Create EMR notebook
  - Connect to cluster
  - Specify notebook location
  - Create notebook
  - Read data from your S3 bucket or from public S3 bucket

CREATE EMR NOTEBOOK

- Services -> EMR
- Click on Notebooks in menu on left
- Click on ‘Create notebook’



CREATE EMR NOTEBOOK

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* line-count

Names may only contain alphanumeric characters, hyphens (-), or underscores (_).

Description

256 characters max.

Cluster* Choose an existing cluster

Choose

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 a location that EMR creates

Choose an existing S3 location in us-west-2

s3://dse230-notebooks/Notebooks

Git repository Link to a Git repository

Tags

* Required

Cancel Create notebook

Add name for notebook

Choose cluster previously created

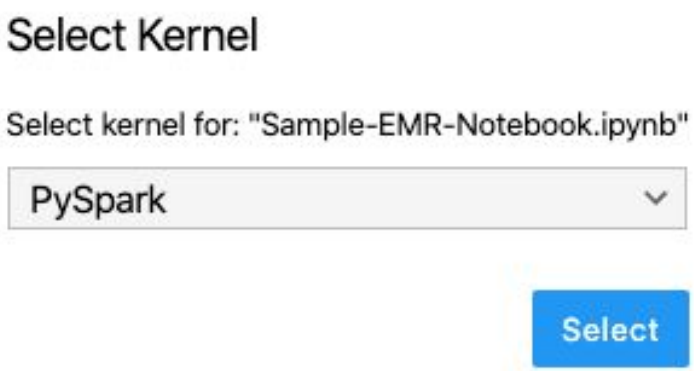
Choose Notebooks folder of S3 bucket previously created

OUTLINE OF STEPS FOR EMR

- Create EC2 key pair
  - Security credentials for connecting to EC2 instance
- Create S3 bucket
  - Upload data to bucket
- Create EMR cluster
  - Specify software configuration
  - Add step to copy data from public S3 bucket for class
  - Specify number of master and worker nodes
  - Specify EC2 key pair
- Create EMR notebook
  - Connect to cluster
  - Specify notebook location and create notebook
  - Read data from S3 bucket

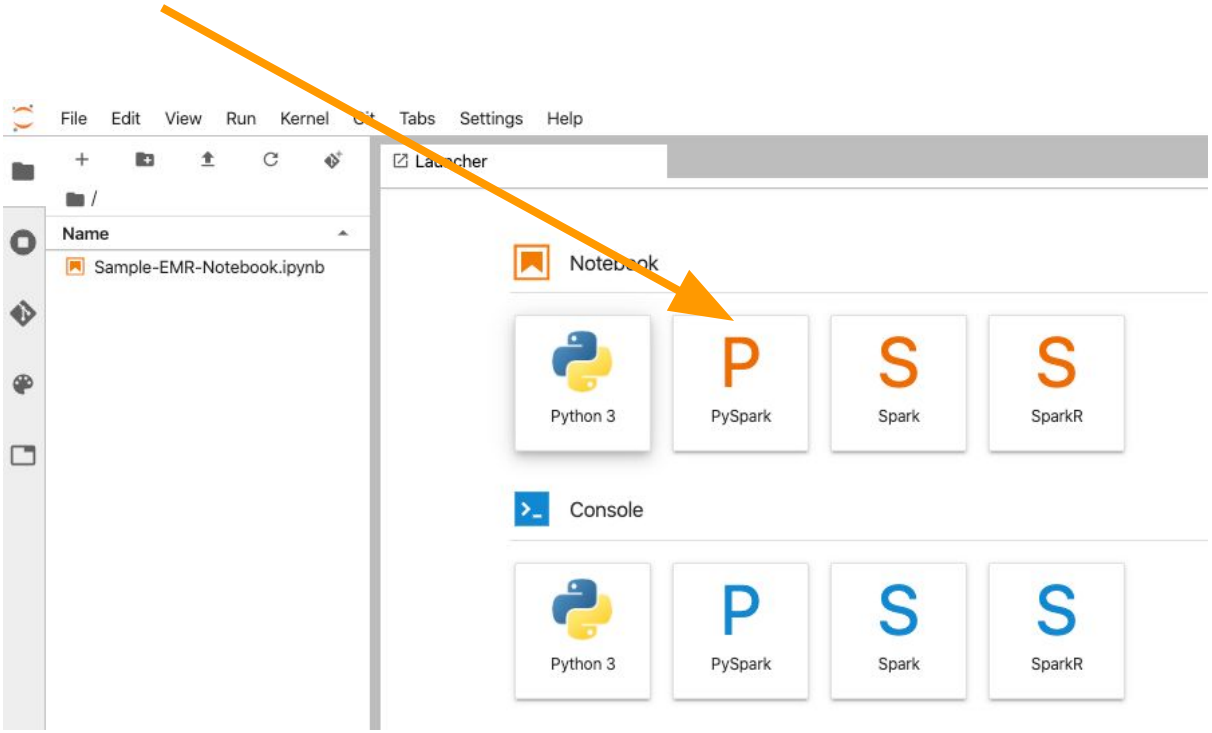
WORKING WITH EMR NOTEBOOKS

- Wait for Notebook status to be ‘Pending’ or ‘Ready’
- Click ‘Open in JupyterLab’
- In JupyterLab,
  - Double click notebook to open it
  - Select ’PySpark’ as kernel (not ‘Python 3’)



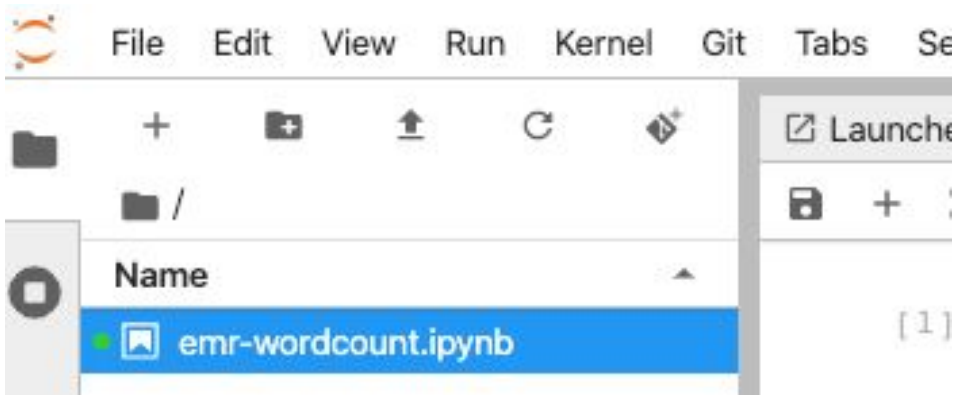
WORKING WITH EMR NOTEBOOKS

- To create new notebook from within JupyterLab
  - Select ’PySpark’ as kernel (not ‘Python 3’)



EMR NOTEBOOK

- Rename notebook from ‘Untitled.ipynb’
- Entering code
  - Can either copy and paste code from notebook shell or upload file directly.
  - Ensure kernel is PySpark



RUNNING EMR NOTEBOOK

- Enter code in notebook, then run as normal
- Print versions

```
import pyspark
print(spark.version, pyspark.version.__version__)
```

- Read data

```
dataFileName = "<file-location>"
textDF = spark.read.text(dataFileName).cache()
```

- Get number of rows

```
textDF.count()
```

- Stop Spark

```
spark.stop()
```

- File location - https://dse230-emr.s3-us-west-1.amazonaws.com/Shakespeare.txt

DOWNLOAD NOTEBOOK AND RESULTS

- Download notebook from JupyterLab
  - File -> Download as Notebook (.ipynb)
  - File -> Download as HTML (.html)
- Download results from S3
  - Navigate to bucket
  - Navigate to folder
  - Right click on file to download

CLEANING UP

- Do this when done with current session
  - Important: Delete cluster and notebook to avoid accumulating fees!
  - Note: There is a daily limit of $5.00 and total course limit of $50.00. You will not be able to run anything if you exceed these limits!
- In JupyterLab
  - File -> Close and Shutdown Notebook
- In EMR
  - Click on Notebooks in the left menu
    - Select Notebook, then click Stop
  - Click on Clusters in the left menu
    - Select cluster and Click Terminate to delete cluster
    - Wait until cluster status is Terminated.
- S3 - Can leave S3 bucket up until end of course

STOP NOTEBOOK

- Cleaning up: Make sure notebook is stopped

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

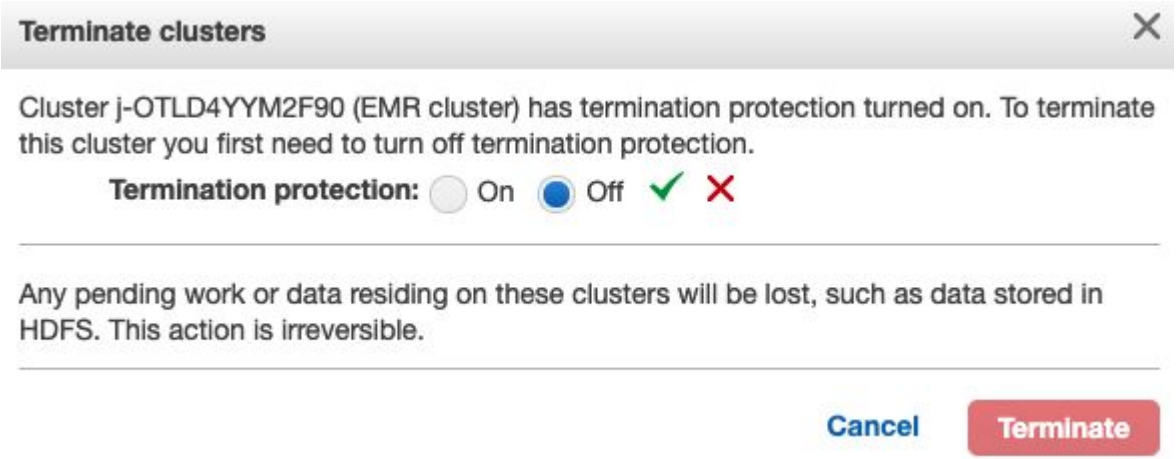
Create notebook | View details | Open in JupyterLab | Open in Jupyter | Start | Stop | Delete

Filter: All notebooks | Filter notebooks ... | 2 notebooks (all loaded)

	Name	Status	Cluster	Creation time (UTC-8)	Last modified
	EMR-Notebook	Stopped	j-332GN8XPDB86V	2021-02-26 19:58 (UTC-8)	20 seconds ago

TERMINATE CLUSTER

- Select cluster. Click ‘Terminate’
- In popup window
  - Click on ‘Change’ for Termination protection
  - Click on ‘Off’. Then click on checkmark
  - Click on ‘Terminate’
- Check that cluster’s status is ‘Terminated’



WORK ON EXISTING NOTEBOOK

- Start another cluster
  - Select ‘Clusters’ on the left
  - Select previous cluster and click Clone
  - Click ‘Create Cluster’
- Open notebook in new cluster
  - Select ‘Notebooks’ on the left
  - Select ‘Change cluster’
  - Select ‘Choose an existing cluster’ and select new cluster
  - Select ‘Change cluster and start notebook’