

CS 649 Big Data: Tools and Methods  
Spring Semester, 2022  
Doc 23 Running Spark  
Apr 7, 2022

Copyright ©, All rights reserved. 2022 SDSU & Roger Whitney,  
5500 Campanile Drive, San Diego, CA 92182-7700 USA.  
OpenContent (<http://www.opencontent.org/opl.shtml>) license  
defines the copyright on this document.

# Installing PySpark

Instructions

[https://spark.apache.org/docs/latest/api/python/getting\\_started/install.html](https://spark.apache.org/docs/latest/api/python/getting_started/install.html)

Short Version:

```
conda install pyspark
```

# Then We Can Run

In [1]:

```
1     from pyspark.sql import SparkSession
2     spark = SparkSession.builder \
3         .master("local") \
4         .appName("Print") \
5         .getOrCreate()
6     print(spark.range(5000).selectExpr("sum(id)").collect())
```

```
[Row(sum(id)=12497500)]
```

```
from pyspark.sql import SparkSession
spark = SparkSession.builder \
    .master("local") \
    .appName("Print") \
    .getOrCreate()
print(spark.range(5000).selectExpr("sum(id)").collect())
```

# Spark Properties

name <https://spark.apache.org/docs/latest/configuration.html>  
master  
logging  
memory  
etc

name - displayed in Spark Master Web page

# master

Master URL	Meaning
local	Run Spark locally with one worker thread.
local[K]	Run Spark locally with K worker threads
local[K,F]	Run Spark locally with K worker threads and F maxFailures
local[*]	Run Spark locally with as many worker threads as logical cores on your machine.
local[*],F]	Run Spark locally with as many worker threads as logical cores on your machine and F maxFailures.
<u>spark://HOST:PORT</u>	Connect to the given Spark standalone cluster master.
<u>spark:// HOST1:PORT1,HOST2:PORT2</u>	Connect to the given Spark standalone cluster with standby masters with Zookeeper.
<u>mesos://HOST:PORT</u>	Connect to the given Mesos cluster.
yarn	Connect to a YARN cluster in client or cluster mode

# Installing PySpark - Non-Notebook

Download Spark

<https://spark.apache.org/downloads.html>

Put SPARK/bin and SPARK/sbin on your path

printExample.py

**->spark-submit ./printExample.py**

```
from __future__ import print_function
def print5000():
    from pyspark.sql import SparkSession
    spark = SparkSession.builder \
        .master("local") \
        .appName("Print") \
        .getOrCreate()
    print(spark.range(5000).selectExpr("sum(id)").collect())

if __name__ == "__main__":
    print5000()
```

# Amazon Elastic Map-Reduce (EMR)

Hadoop, Hive, Spark, etc on Cluster

Predefined set of languages/tools available

Can create cluster of machines

<https://aws.amazon.com>

Create new account

Get 12 months free access

# AWS Free Tier

12 months free

EC2 - compute instances

740 hours per month

Billed in hour increments

Billed per instance

S3 - storage

5 GB

20,000 Get requests

RDS - MySQL, PostgreSQL, SQL Sever

20 GB

750 hours

EC2 Container - Docker images

500 MB



# AWS Educate

~~<https://aws.amazon.com/education/awseducate/>~~

~~SDSU is an institutional member~~

~~Students get \$100 credit~~

# EC2 Pricing

	Price Per Hour	
	On Demand	Spot
a1.medium	\$0.0255	\$0.0050
t3.nano	\$0.0047	\$0.0016
m5.large	\$0.0960	\$0.0202
c5.large	\$0.0850	\$0.0200
p3.2xlarge (GPU)	\$3.0600	\$0.9413

# Basic Outline

Develop & test Spark locally

Upload program file & data to S3

Configure & launch cluster

- AWS Management Console

- AWS CLI

- SDKs

Monitor cluster

Make sure you terminate cluster when done

# Simple Storage System - S3

Files are stored in buckets

Bucket names are global

Supports

- s3 - files divided in to block

- s3n

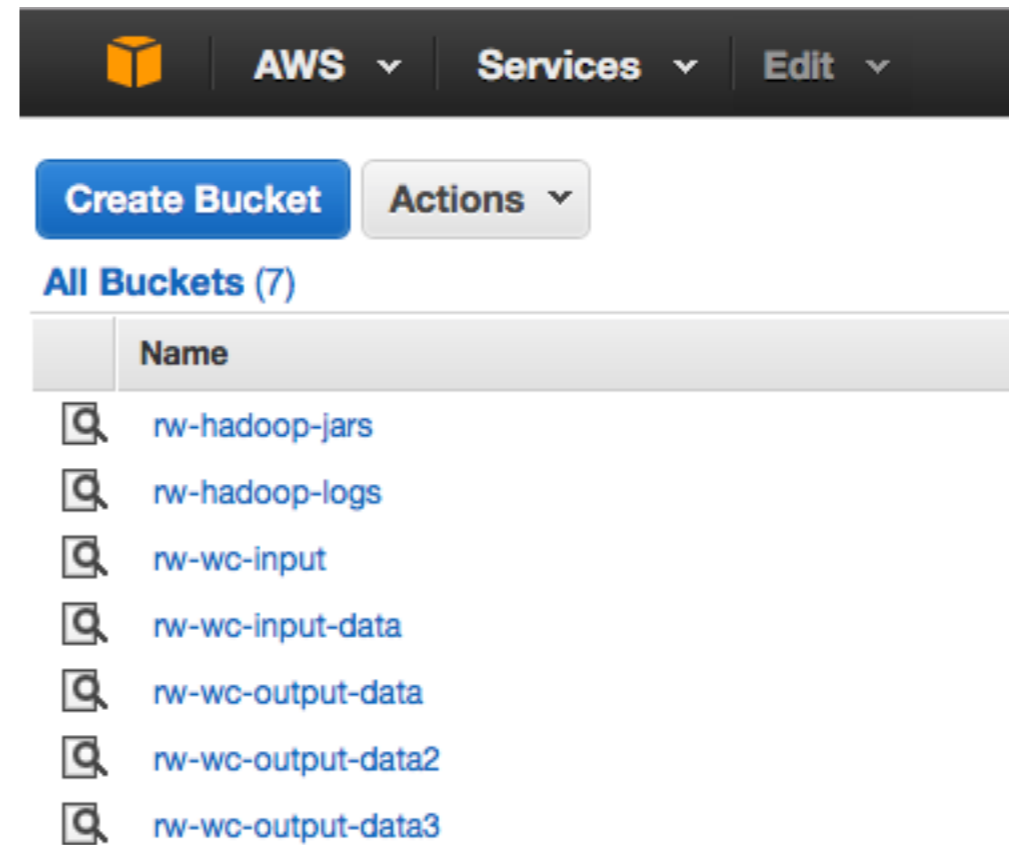
Accessing files

- S3 console

- Third party

- REST

- Java, C#, etc



# Amazon S3

## Buckets (13)


Buckets are containers for data stored in S3. [Learn more](#)

< 1 >

	Name ▲	AWS Region ▼	Access ▼	Creation date ▼
<input type="radio"/>	<a href="#">aws-emr-resources-834365227482-us-west-2</a>	US West (Oregon) us-west-2	Objects can be public	March 26, 2019, 14:47:39 (UTC-07:00)
<input type="radio"/>	<a href="#">aws-logs-834365227482-us-west-2</a>	US West (Oregon) us-west-2	Objects can be public	December 8, 2016, 15:27:27 (UTC-08:00)
<input type="radio"/>	<a href="#">aws-sam-cli-managed-default-samclisourcebucket-busdmbpxv8ww</a>	US West (Oregon) us-west-2	Objects can be public	February 25, 2021, 16:19:47 (UTC-08:00)
<input type="radio"/>	<a href="#">elasticbeanstalk-us-west-2-834365227482</a>	US West (Oregon) us-west-2	Objects can be public	January 27, 2021, 21:50:54 (UTC-08:00)
<input type="radio"/>	<a href="#">rew-648-test</a>	US West (Oregon) us-west-2	Objects can be public	February 22, 2021, 20:31:11 (UTC-08:00)
<input type="radio"/>	<a href="#">rw-648-data</a>	US West (Oregon) us-west-2	Objects can be public	February 7, 2021, 14:47:53 (UTC-08:00)
<input type="radio"/>	<a href="#">rw-696-flight</a>	US West (Oregon) us-west-2	Objects can be public	March 26, 2019, 09:56:35 (UTC-07:00)
<input type="radio"/>	<a href="#">rw-cs696-data</a>	US West (Oregon) us-west-2	Objects can be public	November 6, 2017, 14:47:10 (UTC-08:00)


# S3 Creating a Bucket

## Create bucket


Buckets are containers for data stored in S3. [Learn more](#) 

### General configuration

Bucket name

Bucket name must be unique and must not contain spaces or uppercase letters. [See rules for bucket naming](#) 

AWS Region

Copy settings from existing bucket - *optional*  
Only the bucket settings in the following configuration are copied.

### Block Public Access settings for this bucket

Public access is granted to buckets and objects through access control lists (ACLs), bucket policies, access point policies, or all. In order to ensure that public access to this bucket and its objects is blocked, turn on Block all public

# S3 Costs

## AWS Free Usage Tier

New AWS customers receive each month for one year  
5 GB of Amazon S3 storage in the Standard Storage class,  
20,000 Get Requests,  
2,000 Put Requests, and  
15 GB of data transfer out

	Standard Storage	Standard - Infrequent Access Storage	Glacier Storage
First 50 TB / month	\$0.023 per GB	\$0.0125 per GB	\$0.004 per GB
Next 450 TB / month	\$0.022 per GB	\$0.0125 per GB	\$0.004 per GB
Over 500 TB / month	\$0.021 per GB	\$0.0125 per GB	\$0.004 per GB

# S3 Objects

Objects contain

- Object data

- Metadata

Size

- 1 byte to 5 gigabytes per object

Object data

- Just bytes

- No meaning associated with bytes

Metadata

- Name-value pairs to describe the object

- Some http headers used

  - Content-Type



# S3 Buckets

Namespace for objects

No limitation on number of object per bucket

Only 100 buckets per account

Each bucket has a name

- Up to 255 bytes long

- Cannot be same as existing bucket name by any S3 user

# Bucket Names

Bucket names must

- Contain lowercase letters, numbers, periods (.), underscores (\_), and dashes (-)

- Start with a number or letter

- Be between 3 and 255 characters long

- Not be in an IP address style (e.g., "192.168.5.4")

To conform with DNS requirements, Amazon recommends

- Bucket names should not contain underscores (\_)

- Bucket names should be between 3 and 63 characters long

- Bucket names should not end with a dash

- Bucket names cannot contain dashes next to periods (e.g.,

  - "my-.bucket.com" and "my.-bucket" are invalid

# Key

Unique identifier for an object within a bucket

Object Url

`http://bucketName.s3.amazonaws.com/Key`

`http://doc.s3.amazonaws.com/2006-03-01/AmazonS3.wsdl`

Bucket = doc

Key = 2006-03-01/AmazonS3.wsdl

# Access Control Lists (ACL)

Each Bucket has an ACL

- Determines who has read/write access

Each Object can have an ACL

- Determines who has read/write access

ACL consists of a list of grants

Grant contains

- One grantee

- One permission

# S3 Data Consistency Model

Updates to a single object at a key in a bucket are atomic

But a read after a write may return the old value

Changes may take time to propagate

No object locking

If two writes to same object occur at the same time

The one with later timestamp wins

# Running Program on AWS EMR

Make sure program runs locally

Create program file containing code

Create s3 bucket(s) for

- program file

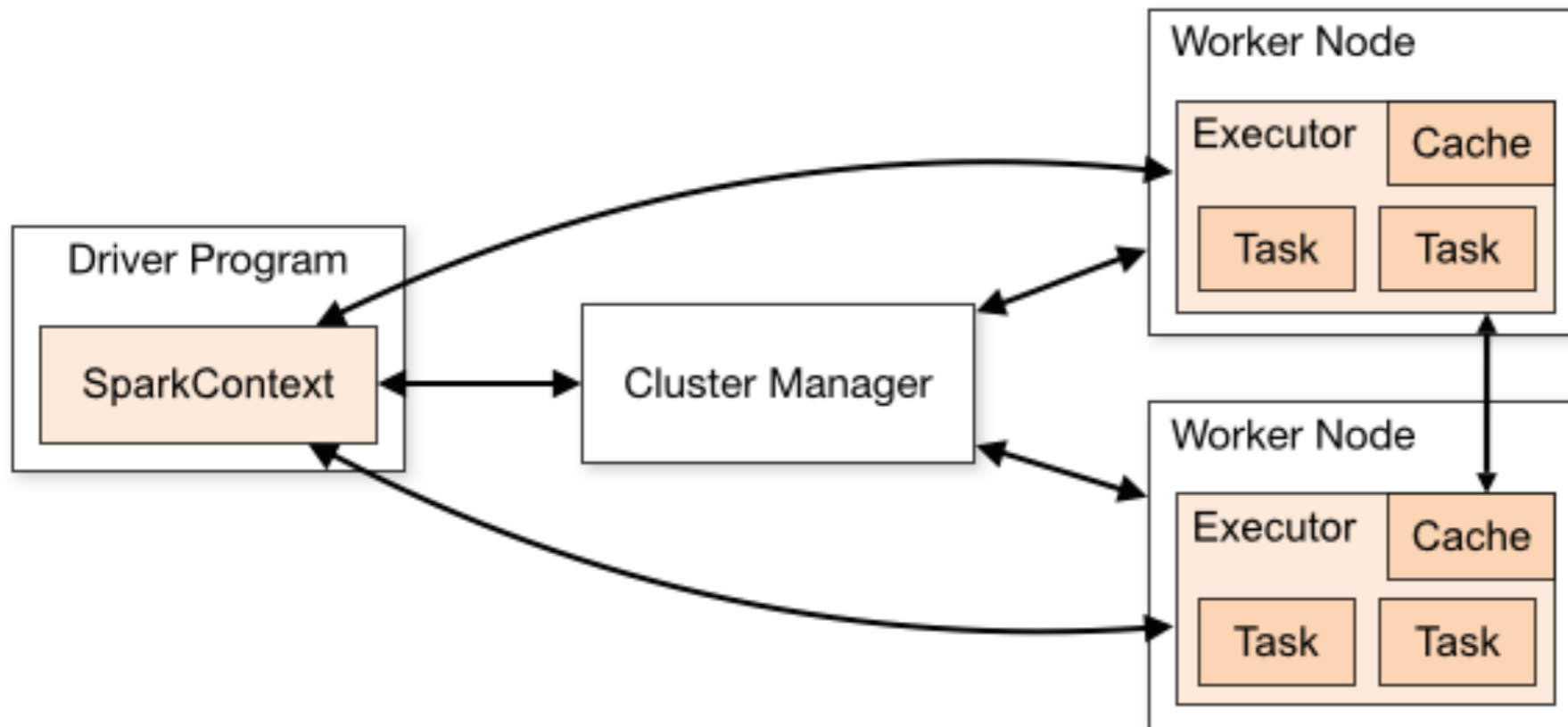
- logs

- input

- output

Upload program & data files to s3

# Spark Components



# Terms

## Application

User program built on Spark

Driver program + executors

## Driver program

The process running the main() function of the application and creating the SparkContext

## Cluster manager

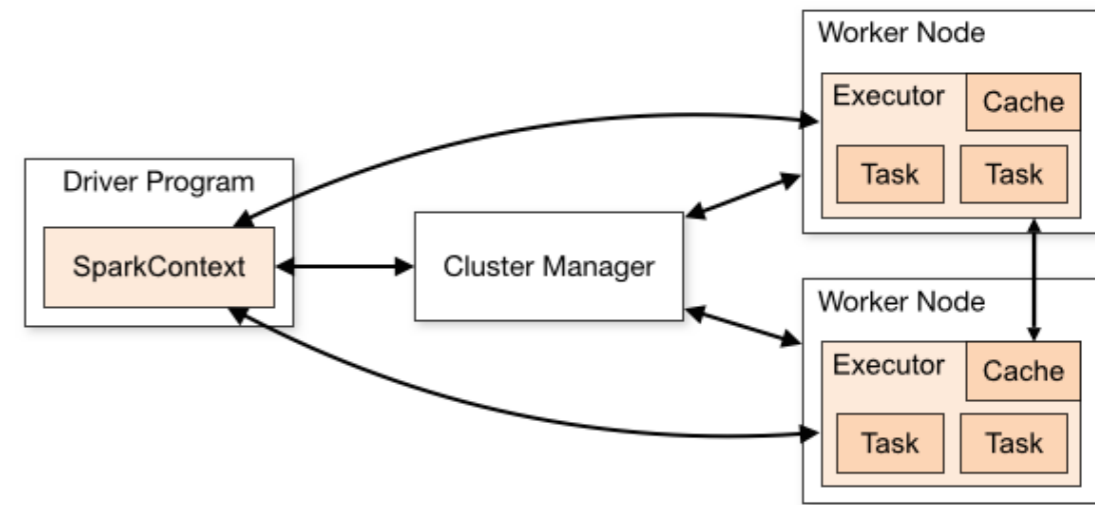
External service for acquiring resources on the cluster

## Deploy mode

Where the driver process runs

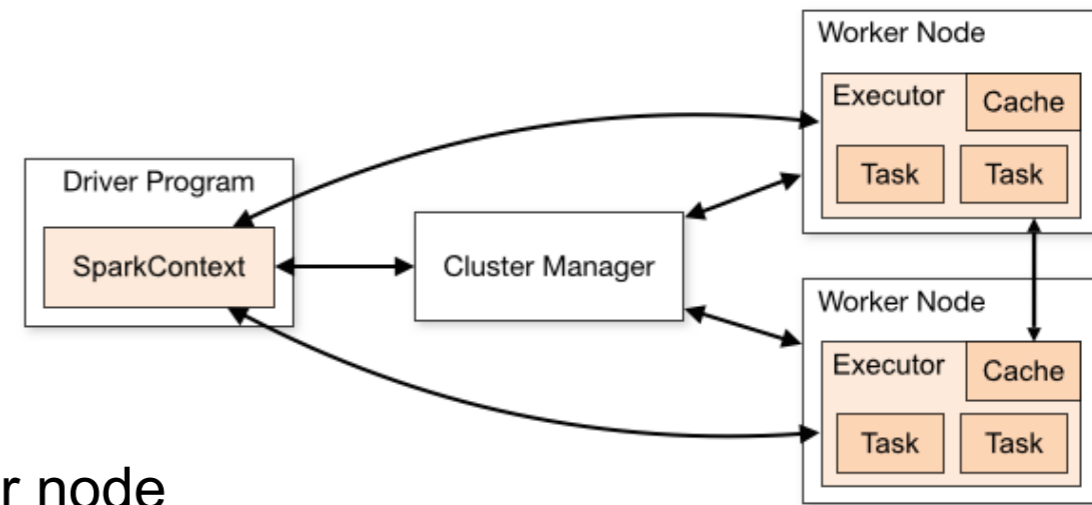
"cluster" - the driver inside of the cluster

"client" - the driver outside of the cluster





# Terms



## Executor

A process launched for an application on a worker node

Runs tasks and keeps data in memory or disk storage across them.

Each application has its own executors

## Task

A unit of work that will be sent to one executor

## Job

A parallel computation consisting of multiple tasks

Gets spawned in response to a Spark action (e.g. save, collect)

## Stage

Job divided into smaller tasks called stages

Depend on each other

# Test Program 1 - Pi

```
from random import random
from operator import add
```

```
from pyspark import SparkContext
```

```
if __name__ == "__main__":
    sc = SparkContext(appName="PythonPi")
    partitions = 3
    n = 100000 * partitions
```

```
def f(_):
    x = random() * 2 - 1
    y = random() * 2 - 1
    return 1 if x ** 2 + y ** 2 < 1 else 0
```

```
count = sc.parallelize(range(1, n + 1), partitions).map(f).reduce(add)
print("Pi is roughly %f" % (4.0 * count / n))
```

```
sc.stop()
```

Designed to have no  
Command line dependancies  
No input or output files

# My S3 Buckets



**Buckets (9)**



	Name	Region
<input type="radio"/>	<a href="#">aws-emr-resources-834365227482-us-west-2</a>	US West (Oregon) us-west-2
<input type="radio"/>	<a href="#">aws-logs-834365227482-us-west-2</a>	US West (Oregon) us-west-2
<input checked="" type="radio"/>	<a href="#">rw-696-flight</a>	US West (Oregon) us-west-2
<input type="radio"/>	<a href="#">rw-cs696-data</a>	US West (Oregon) us-west-2
<input type="radio"/>	<a href="#">rw-hadoop-jars</a>	US West (Oregon) us-west-2
<input type="radio"/>	<a href="#">rw-hadoop-logs</a>	US West (Oregon) us-west-2
<input type="radio"/>	<a href="#">rw-output-data</a>	US West (Oregon) us-west-2
<input type="radio"/>	<a href="#">rw-ws-input-data</a>	US West (Oregon) us-west-2

# My S3 Buckets

Overview **Properties** Permissions Management Access points

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

 Upload  Create folder Download Actions ▾

<input type="checkbox"/>	Name ▾
<input type="checkbox"/>	 flight.py
<input type="checkbox"/>	 pi.py

# My S3 Buckets

pi.py Latest version ▾

Overview	<b>Properties</b>	Permissions	Select from
----------	-------------------	-------------	-------------

- Open
- Download
- Download as
- Make public
- Copy path

**Owner**  
rwhitney

**Last modified**  
Mar 27, 2019 8:58:15 PM GMT-0700

**Etag**  
2a37d19cc32e1a51c0472473e26f72d1

**Storage class**  
Standard

**Server-side encryption**  
None

**Size**  
498.0 B

**Key**  
pi.py

**Object URL**  
<https://rw-696-flight.s3-us-west-2.amazonaws.com/pi.py>

# Spark on AWS - EMR Console

- Amazon EMR
- Clusters
- Security configurations
- Block public access
- VPC subnets
- Events
- Notebooks
  - Git repositories
- Help
- What's new



## Welcome to Amazon Elastic MapReduce

Amazon Elastic MapReduce (Amazon EMR) is a web service that enables businesses, researchers, data analysts, and developers to easily and cost-effectively process vast amounts of data.

You do not appear to have any clusters. Create one now:

[Create cluster](#)

## How Elastic MapReduce Works

**Upload**



Upload your data and processing application to S3.

[Learn more](#)


**Create**



Configure and create your cluster by specifying data inputs, outputs, cluster size, security settings, etc.

[Learn more](#)

**Monitor**



Monitor the health and progress of your cluster. Retrieve the output in S3.

[Learn more](#)

# Using Quick Options

## General Configuration

Cluster name

Logging ⓘ  
S3 folder

Launch mode  Cluster ⓘ  Step execution ⓘ

## Software configuration

Release  ⓘ ←

Applications

- Core Hadoop: Hadoop 3.2.1 with Hive 3.1.2, Hue 4.9.0, Pig 0.17.0 and Tez 0.9.2
- HBase: HBase 2.4.4 with Hadoop 3.2.1, Hive 3.1.2, Hue 4.9.0, Phoenix 5.1.2, and ZooKeeper 3.5.7
- Presto: Presto 0.261 with Hadoop 3.2.1 HDFS and Hive 3.1.2 Metastore
- Spark: Spark 3.1.2 on Hadoop 3.2.1 YARN with and Zeppelin 0.10.0 ←
- Trino: Trino 360 with Hadoop 3.2.1 HDFS and Hive 3.1.2 Metastore

Use AWS Glue Data Catalog for table metadata ⓘ

## Hardware configuration

Instance type  ⓘ ←  
The selected instance type adds 64 GiB of GP2 EBS storage per instance by default. [Learn more](#) ⓘ

Number of instances  (1 master and 2 core nodes)

Cluster scaling  scale cluster nodes based on workload

Auto-termination  Enable auto-termination [Learn more](#) ⓘ  
Terminate cluster when it is idle after  hours  minutes

## Security and access

EC2 key pair  ⓘ [Learn how to create an EC2 key pair.](#)

Permissions  Default  Custom  
Use default IAM roles. If roles are not present, they will be automatically created for you with managed policies for automatic policy updates.

EMR role [EMR\\_DefaultRole](#) ⓘ  Use EMR\_DefaultRole\_V2 ⓘ

EC2 instance [EMR\\_EC2\\_DefaultRole](#) ⓘ

# Use Advanced Options

## Create Cluster - Quick Options [Go to advanced options](#)

### General Configuration

**Cluster name**

**Logging** ⓘ

**S3 folder**

**Launch mode**  **Cluster** ⓘ  **Step execution** ⓘ

### Software configuration

**Release**  ⓘ

**Applications**

- Core Hadoop: Hadoop 2.8.5 with Ganglia 3.7.2, Hive 2.3.4, Hue 4.3.0, Mahout 0.13.0, Pig 0.17.0, and Tez 0.9.1
- HBase: HBase 1.4.9 with Ganglia 3.7.2, Hadoop 2.8.5, Hive 2.3.4, Hue 4.3.0, Phoenix 4.14.1, and ZooKeeper 3.4.13
- Presto: Presto 0.214 with Hadoop 2.8.5 HDFS and Hive 2.3.4 Metastore
- Spark: Spark 2.4.0 on Hadoop 2.8.5 YARN with Ganglia 3.7.2 and Zeppelin 0.8.1

Use AWS Glue Data Catalog for table metadata ⓘ



# Advanced Options

## Software Configuration

Release  ⓘ

- |  |   |  |
|--|---|--|
| <input checked="" type="checkbox"/> Hadoop 3.2.1 | <input type="checkbox"/> Zeppelin 0.10.0        | <input type="checkbox"/> Livy 0.7.1            |
| <input type="checkbox"/> JupyterHub 1.4.1        | <input type="checkbox"/> Tez 0.9.2              | <input type="checkbox"/> Flink 1.14.0          |
| <input type="checkbox"/> Ganglia 3.7.2           | <input type="checkbox"/> HBase 2.4.4            | <input checked="" type="checkbox"/> Pig 0.17.0 |
| <input checked="" type="checkbox"/> Hive 3.1.2   | <input type="checkbox"/> Presto 0.261           | <input type="checkbox"/> ZooKeeper 3.5.7       |
| <input type="checkbox"/>                         |   |  |
| JupyterEnterpriseGateway 2.1.0                   |   |  |
| <input checked="" type="checkbox"/> Hue 4.9.0    | <input type="checkbox"/> MXNet 1.8.0            | <input type="checkbox"/> Sqoop 1.4.7           |
| <input type="checkbox"/> Oozie 5.2.1             | <input type="checkbox"/> Phoenix 5.1.2          | <input type="checkbox"/> Trino 360             |
| <input type="checkbox"/> TensorFlow 2.4.1        | <input checked="" type="checkbox"/> Spark 3.1.2 | <input type="checkbox"/> HCatalog 3.1.2        |

## Multiple master nodes (optional)

- Use multiple master nodes to improve cluster availability. [Learn more](#)

## AWS Glue Data Catalog settings (optional)

- Use for Hive table metadata ⓘ
- Use for Spark table metadata ⓘ

## Edit software settings ⓘ

- Enter configuration  Load JSON from S3

```
classification=config-file-name,properties=[myKey1=myValue1,myKey2=myValue2]
```

## Steps (optional)

A step is a unit of work you submit to the cluster. For instance, a step might contain one or more Hadoop or Spark jobs. You can also submit additional steps to a cluster after it is running. [Learn more](#)

- Concurrency:**  Run multiple steps at the same time to improve cluster utilization
- After last step completes:**  Clusters enters waiting state
- Cluster auto-terminates

Step type

Cancel

Next

# Advanced Options

## Hardware Configuration ⓘ

Specify the networking and hardware configuration for your cluster. Request Spot instances (unused EC2 capacity) to save money.

## Cluster Composition

Specify the configuration of the master, core and task nodes as an instances group or instance fleet. This choice applies to all nodes for the lifetime of the cluster. Instance fleets and instance groups cannot coexist in a cluster. [see this topic](#) ↗.

### Instance group configuration

**Uniform instance groups**  
Specify a single instance type and purchasing option for each node type.

**Instance fleets**  
Specify target capacity and how Amazon EMR fulfills it for each node type. Mix instance types and purchasing options. [Learn more](#) ↗

## Networking

Use a Virtual Private Cloud (VPC) to process sensitive data or connect to a private network. Launch the cluster into a VPC with a public, private or shared subnet. Subnets may be associated with and AWS Outpost or AWS Local Zone.

Launch the cluster into a VPC with a public, private, or shared subnet. Subnets may be associated with an AWS Outpost or AWS Local Zone.

Network  [Create a VPC](#) ⓘ

EC2 Subnet

## 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
<b>Master</b> Master - 1 ✎	<b>m5.xlarge</b> ✎ 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 32 GiB ⓘ ✎ Add configuration settings ✎	1 Instances	<input checked="" type="radio"/> <b>On-demand</b> ⓘ <input type="radio"/> <b>Spot</b> ⓘ <input type="text" value="Use on-demand as r"/>
<b>Core</b> Core - 2 ✎	<b>m5.xlarge</b> ✎ 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 32 GiB ⓘ ✎ Add configuration settings ✎	<input type="text" value="2"/> Instances	<input checked="" type="radio"/> <b>On-demand</b> ⓘ <input type="radio"/> <b>Spot</b> ⓘ <input type="text" value="Use on-demand as r"/>
<b>Task</b> Task - 3 ✎	<b>m5.xlarge</b> ✎ 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 32 GiB ⓘ ✎ Add configuration settings ✎	<input type="text" value="0"/> Instances	<input checked="" type="radio"/> <b>On-demand</b> ⓘ <input type="radio"/> <b>Spot</b> ⓘ <input type="text" value="Use on-demand as r"/>

# Advanced Options

## General Options

**Cluster name**

Logging ⓘ

S3 folder  📁

Log encryption ⓘ

Debugging ⓘ

Termination protection ⓘ

## Tags ⓘ

Key	Value (optional)
<input type="text" value="Add a key to create a tag"/>	<input type="text"/>

## Additional Options

EMRFS consistent view ⓘ

**Custom AMI ID**  ⓘ

▶ Bootstrap Actions

Cancel

Previous

Next

# Advanced Options

## Security Options

EC2 key pair **No key pairs found** ⓘ

Cluster visible to all IAM users in account ⓘ

## Permissions ⓘ

Default  Custom

Use default IAM roles. If roles are not present, they will be automatically created for you with managed policies for automatic policy updates.

EMR role [EMR\\_DefaultRole](#)  Use EMR\_DefaultRole\_V2 ⓘ

EC2 instance profile [EMR\\_EC2\\_DefaultRole](#) ⓘ

Auto Scaling role [EMR\\_AutoScaling\\_DefaultRole](#) ⓘ

▶ Security Configuration

▶ EC2 security groups

ⓘ No EC2 key pair has been selected, so you will not be able to SSH to this cluster or connect to HUE (unless you are using a VPN). [Learn how to create an EC2 Key Pair.](#)

Cancel

Previous

Create cluster

# Cluster Created - Either Quick or Advanced

[Clone](#) [Terminate](#) [AWS CLI export](#)

Cluster: CourseSlidesExample Starting

[Summary](#) [Application history](#) [Monitoring](#) [Hardware](#) [Configurations](#) [Events](#) [Steps](#) [Bootstrap actions](#)

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


## Summary

**ID:** j-173L2X4KKU7HX  
**Creation date:** 2020-03-24 12:05 (UTC-7)  
**Elapsed time:** 30 seconds  
**After last step completes:** Cluster waits  
**Termination protection:** On [Change](#)

## Configuration details

**Release label:** emr-5.29.0  
**Hadoop distribution:** Amazon 2.8.5  
**Applications:** Spark 2.4.4  
**Log URI:** s3://aws-logs-834365227482-us-west-2/elasticmapreduce/   
**EMRFS consistent view:** Disabled  
**Custom AMI ID:** --

## Network and hardware

**Availability zone:** us-west-2a  
**Subnet ID:** [subnet-0f55196b](#)   
**Master:** Provisioning 1 m5.xlarge  
**Core:** Provisioning 2 m5.xlarge  
**Task:** --

# Adding a Step

Clone Terminate AWS CLI export

Cluster: CourseSlidesExample **Starting**

Summary Application history Monitoring Hardware Configurations Events **Steps** Bootstrap actions

Concurrency: 1 [Change](#)

After last step completes: Cluster waits

**Add step** Clone step Cancel step

Filter: All steps  1 step (all loaded)

	ID	Name	Status
	s-18YQWR7FSE5XT	Setup hadoop debugging	Pending

Add step
✕

**Step type**

**Name**

**Deploy mode**


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**

**Application location\***


Path to a JAR with your application and dependencies (client deploy mode only supports a local path).

**Arguments**

Specify optional arguments for your application.

**Action on failure**


What to do if the step fails.

Cancel
Add

Modes - client or cluster

Either works

client mode gives access to standard out

Clone Terminate AWS CLI export

Cluster: CourseSlidesExample **Starting** Configuring cluster software

Summary Application history Monitoring Hardware Configurations Events **Steps** Bootstrap actions

Concurrency: 1 [Change](#)

After last step completes: Cluster waits

Add step Clone step Cancel step

Filter: All steps Filter steps ... 2 steps (all loaded)

	ID	Name	Status	Start time (UTC-7) ▼	Elapsed time	Log files [
<input type="radio"/> ▶ <input checked="" type="radio"/>	s-3E10OBDAULFO	Spark application	Pending		--	<a href="#">View logs</a>
<input type="radio"/> ▶ <input checked="" type="radio"/>	s-18YQWR7FSE5XT	Setup hadoop debugging	Pending		--	<a href="#">View logs</a>





Cluster: CourseSlidesExample **Waiting** Cluster ready after last step completed.

- Summary
- Application history
- Monitoring
- Hardware
- Configurations
- Events
- Steps**
- Bootstrap actions

Concurrency: 1 [Change](#)

After last step completes: Cluster waits

- Add step**
- Clone step
- Cancel step

Filter: All steps Filter steps ... 2 steps (all loaded)

ID	Name	Status	Start time (UTC-7)	Elapsed time	Log files
s-3E10OBDAULFO	Spark application	Completed	2020-03-24 12:10 (UTC-7)	24 seconds	<a href="#">View logs</a>
s-18YQWR7FSE5XT	Setup hadoop debugging	Completed	2020-03-24 12:09 (UTC-7)	2 seconds	<a href="#">View logs</a>

View Jobs in the

Concurrency: 1 [Change](#)

After last step completes: Cluster waits

- Add step**
- Clone step
- Cancel step

Filter: All steps Filter steps ... 2 steps (all loaded)

ID	Name	Status	Start time (UTC-7)	Elapsed time	Log files
s-3E10OBDAULFO	Spark application	Completed	2020-03-24 12:10 (UTC-7)	24 seconds	<a href="#">controller   syslog*   stderr   stdout </a>
s-18YQWR7FSE5XT	Setup hadoop debugging	Completed	2020-03-24 12:09 (UTC-7)	2 seconds	<a href="#">View logs</a>

```
Pi is roughly 3.138120
```

## Example 2

```
def flight(input, output):
    import pyspark.sql.functions as F
    from pyspark.sql import SparkSession
    spark = SparkSession.builder \
        .appName("Flight") \
        .getOrCreate()









    flight_df = spark.read.json(input)

    grouped_df = flight_df.groupBy('DEST_COUNTRY_NAME').agg(F.sum('count'))
    grouped_df.write.format('csv').save(output)

def files_from_args():
    import argparse
    parser = argparse.ArgumentParser()
    parser.add_argument('-i', '--input', default='input')
    parser.add_argument('-o', '--output', default='output')
    args = parser.parse_args()
    return (args.input, args.output)

if __name__ == "__main__":
    inputfile, outputfile = files_from_args()
    flight(inputfile, outputfile)
```

# S3 Buckets

▶	 aws-emr-resources-834365227482-us-west-2	•		
▶	 aws-logs-834365227482-us-west-2	•		
▼	 rw-696-flight	•		
	 2015-summary.json		21 KB	3/26/19
	 flight.py		2 KB	3/26/19
	 pi.py		49...tes	3/27/19
▶	 rw-cs696-data	•		
▶	 rw-hadoop-jars	•		

# Adding a Step

**Add step** ✕

**Step type**

**Name**

**Deploy mode**  Run your driver on a slave node (cluster mode) or on the master node as an external client (client mode).

**Spark-submit options**  Specify other options for spark-submit.

**Application location\***  Path to a JAR with your application and dependencies (client deploy mode only supports a local path).

**Arguments**  Specify optional arguments for your application.

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

Cancel Add

# S3 output

The screenshot shows a file explorer view of an S3 bucket. The root directory is 'rw-output-data', which contains a sub-directory 'run2'. The 'run2' directory lists 20 files. The file 'part-00013-d6b0488c-30c4-472a-9a53-0abd52629b65-c000.csv' is highlighted with a mouse cursor.

File Name	Size	Modified
rw-output-data		
run2		
_SUCCESS	Zero KB	3/28/19
part-00000-d6b0488c-30c4-472a-9a53-0a...	Zero KB	3/28/19
part-00002-d6b0488c-30c4-472a-9a53-0a...	35...tes	3/28/19
part-00008-d6b0488c-30c4-472a-9a53-0a...	11...tes	3/28/19
part-00009-d6b0488c-30c4-472a-9a53-0a...	11...tes	3/28/19
part-00010-d6b0488c-30c4-472a-9a53-0a...	12...tes	3/28/19
part-00012-d6b0488c-30c4-472a-9a53-0a...	10...tes	3/28/19
<b>part-00013-d6b0488c-30c4-472a-9a53-0abd52629b65-c000.csv</b>		
part-00014-d6b0488c-30c4-472a-9a53-0a...	11...tes	3/28/19
part-00015-d6b0488c-30c4-472a-9a53-0a...	23...tes	3/28/19
part-00016-d6b0488c-30c4-472a-9a53-0a...	8 bytes	3/28/19
part-00017-d6b0488c-30c4-472a-9a53-0a...	11...tes	3/28/19
part-00021-d6b0488c-30c4-472a-9a53-0a...	7 bytes	3/28/19
part-00022-d6b0488c-30c4-472a-9a53-0a...	13...tes	3/28/19
part-00026-d6b0488c-30c4-472a-9a53-0a...	10...tes	3/28/19
part-00029-d6b0488c-30c4-472a-9a53-0a...	9 bytes	3/28/19
part-00030-d6b0488c-30c4-472a-9a53-0a...	40...tes	3/28/19
part-00031-d6b0488c-30c4-472a-9a53-0a...	10...tes	3/28/19
part-00032-d6b0488c-30c4-472a-9a53-0a...	9 bytes	3/28/19
part-00038-d6b0488c-30c4-472a-9a53-0a...	50...tes	3/28/19
part-00040-d6b0488c-30c4-472a-9a53-0a...	10...tes	3/28/19
part-00044-d6b0488c-30c4-472a-9a53-0a...	14...tes	3/28/19
part-00045-d6b0488c-30c4-472a-9a53-0a...	22...tes	3/28/19

# Warning on AWS

It can take 5-10 minutes to start cluster

Logs do not show your logging statements

When you configure Steps incorrectly they fail  
Error messages are not very helpful

# SSH to your Master Node

Create Amazon EC2 Key pair

Instructions

<http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/ec2-key-pairs.html#having-ec2-create-your-key-pair>

Open EC2 Dashboard - Select Key Pairs

# SSH to your Master Node

In Create Cluster - Quick Options

- HIVE 2.3.0 Metastore
- Spark: Spark 2.2.0 on Hadoop 2.7.3 YARN with Ganglia 3.7.2 and Zeppelin 0.7.2
- Use AWS Glue Data Catalog for table metadata ⓘ

## Hardware configuration

- Instance type**  ⌵
- Number of instances**  (1 master and 2 core nodes)

## Security and access

- EC2 key pair**  ⌵ ⓘ [Learn how to create an EC2 key pair.](#)
- Permissions**  Default  Custom  
Use default IAM roles. If roles are not present, they will be automatically created for you with managed policies for automatic policy updates.
- EMR role** [EMR\\_DefaultRole](#) ⓘ
- EC2 instance profile** [EMR\\_EC2\\_DefaultRole](#) ⓘ

Cancel

Create cluster



# SSH to your Master Node

[Clone](#) [Terminate](#) [AWS CLI export](#)

Cluster: LargeRun **Waiting** Cluster ready after last step completed.

[Summary](#) [Application history](#) [Monitoring](#) [Hardware](#) [Events](#) [Steps](#) [Configurations](#) [Bootstrap actions](#)

**Connections:** [Enable Web Connection](#) – Zeppelin, Spark History Server, Ganglia, Resource Manager ... (View All)  
**Master public DNS:** ec2-34-211-49-97.us-west-2.compute.amazonaws.com [SSH](#)  
**Tags:** -- [View All / Edit](#)

Click for Instructions

# Command-line Tools

## Flintrock

Open-source command-line tool for launching Apache Spark clusters

<https://github.com/nchammas/flintrock>

## aws cli

Amazon's command line tool

<https://aws.amazon.com/cli/>

# Generating the Command Line

[Clone](#) [Terminate](#) [AWS CLI export](#)

Cluster: CourseSlides **Starting** Configuring cluster software

[Summary](#) [Application history](#) [Monitoring](#) [Hardware](#) [Configurations](#) [Events](#) [Steps](#) [Bootstrap actions](#)

[Add step](#) [Clone step](#) [Cancel step](#)

Steps

[View all interactive jobs](#) | [View all j](#)

Filter:   2 steps (all loaded) [Refresh](#)

	ID	Name	Status	Start time (UTC-7)	Elapsed time	Log files <a href="#">↗</a>	Actions
<input type="radio"/>	s-3Q0ZVOAZV3VR	Spark application	Pending		--	<a href="#">View logs</a>	<a href="#">View jobs</a>
<input type="radio"/>	s-KSWMDHULHSD8	Setup hadoop debugging	Pending		--	<a href="#">View logs</a>	<a href="#">View jobs</a>