

AWS Data Science

Analyze Big Data with Hadoop

Create a Hadoop Cluster and run a Hive Script to process log data

Step1: Create S3 bucket and EC2 key pair

The screenshot displays two parts of the AWS Management Console. The top part shows the 'Amazon S3' console with a list of buckets. A single bucket named 'bigdata-hadoop' is listed in the US East (Ohio) region, created on Jul 27, 2020. The bottom part shows the 'Key Pairs' console, indicating a key pair named 'key-access-aws-instance' has been successfully created with a fingerprint starting with '3e:d0:42:5f:bc:33:fe:f0:fb:9f:25:e7:74:...'.

Step2: Launch Amazon EMR Cluster

The screenshot shows the 'Amazon EMR' console with the configuration details for a cluster named 'my-emr-cluster', which is in the 'Starting' state. The 'Summary' tab is active, showing the cluster ID 'j-1UX5B7MQCCA19', creation date '2020-07-27 20:49 (UTC-5)', and elapsed time of 8 minutes. The 'Configuration details' tab shows the release label 'emr-5.30.1', Hadoop distribution 'Amazon 2.8.5', and applications including Hive, Hue, Mahout, Pig, Tez, and Oozie. The 'Network and hardware' tab shows the availability zone 'us-east-2b', subnet ID 'subnet-0677727c', and master node configuration with 1 m5.xlarge instance.

Step3: Allow SSH access

Amazon EMR

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

Termination protection: Off [Change](#)

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

Master public DNS: [ec2-18-219-153-173.us-east-2.compute.amazonaws.com](#)

[Connect to the Master Node Using SSH](#)

Log URI: [s3://aws-logs-\[redacted\]-us-east-2/elasticmapreduce/](#)

EMRFS consistent view: Disabled

Custom AMI ID: --

Application user interfaces

Persistent user interfaces: --

On-cluster user interfaces: Not Enabled [Enable an SSH Connection](#)

Network and hardware

Availability zone: us-east-2b

Subnet ID: [subnet-0677727c](#)

Master: **Bootstrapping** 1 m5.xlarge

Core: **Provisioning** 2 m5.xlarge

Task: --

Cluster scaling: Not enabled

Security and access

Key name: key-access-aws-instance

EC2 instance profile: EMR_EC2_DefaultRole

EMR role: EMR_DefaultRole

Visible to all users: All [Change](#)

Security groups for Master: [sg-0c8702d724fc31e47](#) (ElasticMapReduce-master)

Security groups for Core & Task: [sg-06aa20aa97fcb8306](#) (ElasticMapReduce-slave)

Core Instance Group: Your account is currently being verified. Verification normally takes less than 2 hours. Until your account is verified, you may not be able to launch additional instances or create additional volumes. If you are still receiving this message after more than 2 hours, please let us know by writing to [aws-verification@amazon.com](#). We appreciate your patience..

New EC2 Experience [Tell us what you think](#)

- EC2 Dashboard **New**
- Events **New**
- Tags
- Limits
- Instances
 - Instances
 - Instance Types
 - Launch Templates
 - Spot Requests
 - Savings Plans
 - Reserved Instances
 - Dedicated Hosts **New**
 - Capacity Reservations
- Images
 - AMIs
- Elastic Block Store
 - Volumes
 - Snapshots
 - Lifecycle Manager

Inbound security group rules successfully modified on security group (sg-0c8702d724fc31e47 | ElasticMapReduce-master)

Details

EC2 > Security Groups

Security Groups (1/3) [Info](#)

[Filter security groups](#)

Name	Security group ID	Security group name	VPC ID	Description	Owner
-	sg-06aa20aa97fcb8306	ElasticMapReduce-slave	vpc-614feb0a	Slave group for Elastic ...	472869819
<input checked="" type="checkbox"/>	sg-0c8702d724fc31e47	ElasticMapReduce-master	vpc-614feb0a	Master group for Elastic ...	472869819

Rules

Type	Protocol	Port range	Source	Description - optional
All TCP	TCP	0 - 65535	sg-06aa20aa97fcb8306 (ElasticMapReduce-slave)	-
All TCP	TCP	0 - 65535	sg-0c8702d724fc31e47 (ElasticMapReduce-master)	-
SSH	TCP	22	[redacted]/32	-
Custom TCP	TCP	8443	52.95.24.0/23	-

Step4: Run hive script to Process Data

Hive_CloudFront.q

-- Summary: This sample shows you how to analyze CloudFront logs stored in S3 using Hive

```
-- Create table using sample data in S3. Note: you can replace this S3 path with your own.
```

```
CREATE EXTERNAL TABLE IF NOT EXISTS cloudfront_logs (
```

DateObject Date,

LocalTime STRING,

Location STRING,

Bytes INT,

RequestIP STRING,

Method STRING,

Host STRING,

Uri STRING,

Status INT,

Referrer STRING,

OS String,

Browser String,

BrowserVersion String

)

ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.RegexSerDe'

WITH SERDEPROPERTIES (

[illegible]

```
) LOCATION '${INPUT}/cloudfront/data';
```

-- Total requests per operating system for a given time frame

```
INSERT OVERWRITE DIRECTORY '${OUTPUT}/os_requests/' SELECT os, COUNT(*) count FROM
cloudfront_logs WHERE dateobject BETWEEN '2014-07-05' AND '2014-08-05' GROUP BY os;
```

Step5: Submit the Hive Script as a Step

Amazon EMR

- Clusters
- Notebooks
- Git repositories
- Security configurations
- Block public access

Create cluster View details Clone Terminate

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

	Name	ID	Status	Creation time (UTC-5)	Elapsed time	Normalized instance hours
<input type="checkbox"/>	my-emr-cluster	j-1UX5B7MQCCA9	Waiting Cluster ready	2020-07-27 20:49 (UTC-5)	27 minutes	0

Cluster: my-emr-cluster **Waiting** Cluster ready after last step completed.

Summary Application user interfaces Monitoring Hardware Configurations Events Steps Bootstrap actions

Concurrent After last step completes: Cluster waits

Add step

Filter: A

Add step

Step type: Hive program

Name: Hive program

Script S3 location*: s3://us-east-2.elasticmapreduce.samples/cloudfront/c S3 location of your Hive script.
s3://<bucket-name>/<path-to-file>

Input S3 location: s3://us-east-2.elasticmapreduce.samples S3 location of your Hive input files.
s3://<bucket-name>/<folder>/

Output S3 location: s3://bigdata-hadoop/MyHiveQueryResults/ S3 location of your Hive output files.
s3://<bucket-name>/<folder>/

Arguments: Specify optional arguments for your script.

Action on failure: Continue What happens if the step fails

Cancel Add

Clone Terminate AWS CLI export

Cluster: my-emr-cluster **Waiting** Cluster ready after last step completed.

Summary Application user interfaces Monitoring Hardware Configurations Events Steps Bootstrap actions

Concurrency: 1 [Change](#)

After last step completes: Cluster waits

Add step Clone step Cancel step

[View Jobs in the Application History Tab](#)

Filter:	All steps	Filter steps ...	2 steps (all loaded)					
	ID	Name	Status	Start time (UTC-5)	Elapsed time	Log files		
<input type="radio"/>	s-3OZPSDS4KN4G	Hive program	Pending		--	View logs		
<input type="radio"/>	s-HPLKJW4G5V8D	Setup hadoop debugging	Completed	2020-07-27 21:16 (UTC-5)	6 seconds	View logs		

Step6: view the output of Hive script

Amazon S3 > bigdata-hadoop > MyHiveQueryResults > os_requests

bigdata-hadoop

Overview

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

Upload Create folder Download Actions

US East (Ohio)

Viewing 1 to 2

Name	Last modified	Size	Storage class
000000_0	Jul 27, 2020 9:22:07 PM GMT-0500	36.0 B	Standard
000001_0	Jul 27, 2020 9:22:07 PM GMT-0500	24.0 B	Standard

Viewing 1 to 2

```

1 LinuxSOH813
2 MacOSSOH852
3 OSXSOH799
4 iOSSOH794
5 AndroidSOH855
6 WindowsSOH883
7
  
```

Step7: Cleanup Resources

Create cluster View details Clone Terminate

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

Name	ID	Status	Creation time (UTC-5)	Elapsed time	Normalized instance hours
my-emr-cluster	j-1UX5B7MQCCA19	Terminating User request	2020-07-27 20:49 (UTC-5)	37 minutes	0

S3 buckets

[Discover the console](#)

All access types ▼

- [+ Create bucket](#)
- [Edit public access settings](#)
- [Empty](#)
- [Delete](#)

0 Buckets 0 Regions [↻](#)

You do not have any buckets. Here is how to get started with Amazon S3.



Create a new bucket



Upload your data



Set up your permissions

Operations 0 In progress 6 Success 0 Error