

INTRO TO AWS (AMAZON WEB SERVICES)

Sept 6, 2012

i290-1, Analyzing Big Data With Twitter



UC Berkeley School of Information



WHAT IS AWS?

- **Very Large Shared “Cloud” Computing and Storage Resource**
- **Pre-Loaded Hadoop Clusters and Pig Environment**
- **Many Other Tools and Services**

USING THE AWS TOOLS SUPPLIED BY THIS CLASS

- This is a PRIVILEGE; we are covering the cost for you.
- It can be EXPENSIVE; we all have to work together to control for this and prevent accidental over expenditures.
- We also will have to control when you have access to the clusters. (We are charged by cluster access time, not cycles.)
- You are only to use the services we describe and for the purposes we describe. You are NOT to use extra time for your own interests. (You can always get your own account.)

RUNNING PIG ON AWS: METHOD I

- We assign each student or student pair user ID and password just for this class session.
- You can run a script, or run in interactive mode.
- When the script terminates, it stops the charges. But what if your script does not terminate? You need to monitor it.
- When you use interactive mode, you have to terminate the cluster activity yourself.

RUNNING PIG ON AWS: METHOD 2

- We have to do this at a set time so we can turn the clusters on and off.
- We create a cluster for each student or student pair.
- We “launch” that cluster. That means it is active and ready for you to use as much as you like.
- You have to access it via an ssh interface however.
- When the time is up, we terminate the cluster and the charges stop.

RUNNING WITH METHOD 1A



Innovation.

Powered by Amazon Web Services.



Low Cost

Pay-as-you-go, no upfront expenses or long-term commitments.



Instant Elasticity

Instantly deploy your application. Scale resources up or down based on demand.



Open & Flexible

If it runs in a data center, it can run on AWS. You have full control.



Secure

Utilize a secure technology platform built and managed by Amazon.

[Sign Up Now »](#)[Learn more about the AWS Free Tier »](#)

What is AWS?

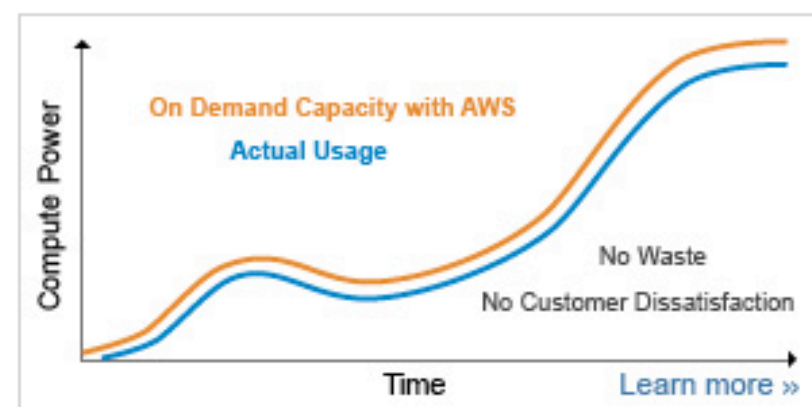
Customer Applications / AWS Marketplace

[Deployment & Management](#)[Application Services](#)[Foundation Services](#)[Global Infrastructure](#)

Amazon Web Services infrastructure enable you to run your applications in the cloud: from enterprise projects to social

One of the key benefits of cloud computing is the opportunity to replace up-front capital infrastructure expenses with low variable costs

Cost Savings with AWS



Recent News

[Announcements](#)[Media Coverage](#)

Digital Media in the AWS Cloud

Los Angeles, CA | September 19



[AWS Elastic Beanstalk Now Available in Los Angeles](#)

[AWS Support](#)

AWS.AMAZON.COM

[Learn the 7 reasons AWS customers are saving money »](#)

31

AUG

[Amazon S3 announces Cross-Origin Resource Sharing \(CORS\) support](#)



Welcome

The AWS Management Console provides a graphical interface to Amazon Web Services. Learn more about how to use our services to meet your needs, or get started by selecting a service.

[Getting started guides](#)[Reference architectures](#)[Free Usage Tier](#)

Set Start Page

[Console Home](#) ▾**AWS
re:Invent**


November 27-29, 2012 Las Vegas
[Register Now](#)

Amazon Web Services

Compute & Networking

 **Direct Connect** NEW
Dedicated Network Connection to AWS **EC2**
Virtual Servers in the Cloud **Elastic MapReduce**
Managed Hadoop Framework **Route 53**
Scalable Domain Name System **VPC**
Isolated Cloud Resources

Storage & Content Delivery

 **CloudFront**
Global Content Delivery Network **Glacier** NEW
Archive Storage in the Cloud **S3**
Scalable Storage in the Cloud **Storage Gateway**
Integrates on-premises IT environments with Cloud storage


Database

 **DynamoDB**
Predictable and Scalable NoSQL Data Store **ElastiCache**
In-Memory Cache

Deployment & Management

 **CloudFormation**
Templated AWS Resource Creation **CloudWatch**
Resource & Application Monitoring **Elastic Beanstalk**
AWS Application Container **IAM**
Secure AWS Access Control


App Services

 **CloudSearch**
Managed Search Service **SES**
Email Sending Service **SNS**
Push Notification Service **SQS**
Message Queue Service **SWF**
Workflow Service for Coordinating Application Components




MANAGEMENT CONSOLE



Buckets

 Create Bucket

Actions ▾

 test-st-1-bucket test-st-2-bucket tucson2 wordcount.ek

← Select one of your buckets to the left to look at the objects it contains, or to upload objects into it.

S3 CONSOLE

Buckets

Create Bucket

Actions ▾

test-st-1-bucket

test-st-2-bucket

tucson2

wordcount.ek

Objects and Folders

Upload

Create Folder

Actions ▾

Refresh

Properties

Transfers

Help

tucson2

Name	Size	Last Modified
log	--	--
practice.pig	179 bytes	Wed Sep 05 15:43:49 GMT-700 2012
tucson-test.pig	203 bytes	Wed Sep 05 22:10:12 GMT-700 2012
tucson.txt	15.3 KB	Wed Sep 05 15:43:50 GMT-700 2012
tucson_sentences_with_hers	--	--

tucson-test (1).pig

```
sentences = LOAD 's3://tucson2/tucson.txt' USING PigStorage() AS (sentence:chararray);  
hers = FILTER sentences BY sentence MATCHES '.* her .*';  
STORE hers INTO 's3://tucson2/tucson_sentences_with_hers';
```

THIS IS AN EXAMPLE: USE YOUR OWN BUCKET TO AVOID CONFLICTS
REMEMBER NOT TO WRITE TO THE SAME DIRECTORY TWICE.

Buckets

Create Bucket

Actions ▾

test-st-1-bucket

test-st-2-bucket

tucson2

wordcount.ek

Objects and Folders

Upload

Create Folder

Actions ▾

Refresh

Properties

Transfers

Help

tucson2

Name	Size	Last Modified
log	--	--
practice.pig	179 bytes	Wed Sep 05 15:43:49 GMT-700 2012
tucson-test.pig	203 bytes	Wed Sep 05 22:10:12 GMT-700 2012
tucson.txt	15.3 KB	Wed Sep 05 15:43:50 GMT-700 2012
tucson_sentences_with_hers	--	--

S3 CONSOLE

Buckets

Create Bucket

Actions ▾

test-st-1-bucket

test-st-2-bucket

tucson2

wordcount.ek

Objects and Folders

Upload

Create Folder

Actions ▾

Refresh

Properties

Transfers

Help

tucson2 > tucson_sentences_with_hers

Name	Size	Last Modified
_SUCCESS	0 bytes	Wed Sep 05 23:53:32 GMT-700 2012
part-m-00000	3.8 KB	Wed Sep 05 23:53:27 GMT-700 2012

part-m-00000 (1)

Scripture tells us, "There is a river whose streams make glad the city of God, the holy place where the most high dwells. God is within her, she will not fall; God will help her at break of day."

On Saturday morning, Gabby, her staff, and many of her constituents gathered outside a supermarket to exercise their right to peaceful assembly and free speech. George and Dorothy Morris -- "Dot" to her friends -- were high school sweethearts who got married and had two daughters. They did everything together, traveling the open road in their R.V., enjoying what their friends called a 50-year honeymoon.

A New Jersey native, Phyllis Schneck retired to Tucson to beat the snow. But in the summer, she would return east, where her world revolved around her three children, her seven grandchildren, and two- year-old great-granddaughter. A gifted quilter, she'd often work under her favorite tree, or sometimes she'd sew aprons with the logos of the Jets and the Giants...

... to give out at the church where she volunteered. A Republican, she took a liking to Gabby and wanted to get to know her better.

... but his true passion was helping people. As Gabby's outreach director, he made the cares of thousands of her constituents his own, seeing to it that seniors got the Medicare benefits that they had earned, that veterans got the medals and the care that they deserved, that government was working for ordinary folks.

And then there is nine-year-old Christina-Taylor Green. Christina was an A student. She was a dancer. She was a gymnast. She was a swimmer. She decided that she wanted to be the first woman to play in the Major Leagues, and as the only girl on her Little League team, no one put it past her.

She showed an appreciation for life uncommon for a girl her age. She'd remind her mother, "We are so blessed. We have the best life." And she'd pay those blessings back by participating in a charity that helped children who were less fortunate.

And I want to tell you -- her husband, Mark, is here, and he allows me to share this with you. Right after we went to visit, a few minutes after we left her room and some of her colleagues from Congress were in the room, Gabby opened her eyes for the first time. Gabby opened her eyes for the first time.

Buckets

Create Bucket

Actions ▾

test-st-1-bucket

test-st-2-bucket

tucson2

wordcount.ek

Objects and Folders

Upload

Create Folder

Actions ▾

Refresh

Properties

Transfers

Help

tucson2 > tucson_sentences_with_hers

Name	Size	Last Modified
_SUCCESS	0 bytes	Wed Sep 05 23:53:32 GMT-700 2012
part-m-00000	3.8 KB	Wed Sep 05 23:53:27 GMT-700 2012

S3 CONSOLE

Buckets

Create Bucket

Actions ▾

test-st-1-bucket

test-st-2-bucket

tucson2

wordcount.ek

Objects and Folders

Upload

Create Folder

Actions ▾

Refresh

Properties

Transfers

Help

tucson2

Name

Size

Last Modified

log

practice.pig

179 bytes

Wed Sep 05 15:43:49 GMT-700 2012

tucson-test

tucson.txt

tucson_ser

Create a Bucket - Select a Bucket Name and Region

Cancel X

A bucket is a container for objects stored in Amazon S3. When creating a bucket, you can choose a Region to optimize for latency, minimize costs, or address regulatory requirements. For more information regarding bucket naming conventions, please visit the [Amazon S3 documentation](#).

Bucket Name:

Region:

US Standard ▾

WARNING: BUCKET NAMES MUST USE LOWER CASE LETTERS

Set Up Logging >

Create

Cancel



Welcome

The AWS Management Console provides a graphical interface to Amazon Web Services. Learn more about how to use our services to meet your needs, or get started by selecting a service.

[Getting started guides](#)

[Reference architectures](#)

[Free Usage Tier](#)

Set Start Page

Console Home ▾

AWS
re:Invent

November 27-29, 2012 Las Vegas
[Register Now](#)

Amazon Web Services

Compute & Networking



Direct Connect NEW
Dedicated Network Connection to AWS



EC2
Virtual Servers in the Cloud



Elastic MapReduce
Managed Hadoop Framework



Route 53
Scalable Domain Name System



VPC
Isolated Cloud Resources

Storage & Content Delivery



CloudFront
Global Content Delivery Network



Glacier NEW
Archive Storage in the Cloud



S3
Scalable Storage in the Cloud



Storage Gateway
Integrates on-premises IT environments with Cloud storage

Database



DynamoDB
Predictable and Scalable NoSQL Data Store



ElastiCache
In-Memory Cache

Deployment & Management



CloudFormation
Templated AWS Resource Creation



CloudWatch
Resource & Application Monitoring



Elastic Beanstalk
AWS Application Container



IAM
Secure AWS Access Control

App Services



CloudSearch
Managed Search Service



SES
Email Sending Service



SNS
Push Notification Service



SQS
Message Queue Service




SWF
Workflow Service for Coordinating Application Components


MANAGEMENT CONSOLE

Your Elastic MapReduce Job Flows


Region:  US East (Virginia) ▼

 Create New Job Flow

 Terminate

 Debug










 Show/Hide

 Refresh

 Help

Viewing: All ▼







1 to 9 of 9 Job Flows



	Name	State	Creation Date	Elapsed Time	Normalized Instance Hours
<input type="checkbox"/>	tucson pig interactive	 TERMINATED	2012-09-05 23:37 PDT	0 hours 12 minutes	3
<input type="checkbox"/>	tucson pig 3	 TERMINATED	2012-09-05 22:11 PDT	0 hours 10 minutes	3
<input type="checkbox"/>	Tucson Pig 2	 FAILED	2012-09-05 15:45 PDT	0 hours 4 minutes	3
<input type="checkbox"/>	My Job Flow	 FAILED	2012-09-05 15:41 PDT	0 hours 0 minutes	0
<input type="checkbox"/>	My Job Flow test 1	 TERMINATED	2012-09-05 12:57 PDT	1 hour 29 minutes	6
<input type="checkbox"/>	111	 FAILED	2012-09-05 02:30 PDT	0 hours 0 minutes	0
<input type="checkbox"/>	pigExample3	 FAILED	2012-08-31 02:19 PDT	0 hours 3 minutes	3
<input type="checkbox"/>	pigExample2	 FAILED	2012-08-30 15:33 PDT	0 hours 3 minutes	3
<input type="checkbox"/>	pigExample1	 FAILED	2012-08-30 14:42 PDT	0 hours 3 minutes	3

EMR CONSOLE

Your Elastic MapReduce Job Flows

Region:  US East (Virginia)   Create New Job Flow  Terminate  Debug  Show/Hide  Refresh  Help

Viewing: All     1 to 9 of 9 Job Flows  

	Name	State	Creation Date	Elapsed Time	Normalized Instance Hours
<input type="checkbox"/>	tucson pig interactive	 TERMINATED	2012-09-05 23:37 PDT	0 hours 12 minutes	3
<input checked="" type="checkbox"/>	tucson pig 3	 TERMINATED	2012-09-05 22:11 PDT	0 hours 10 minutes	3

1 Job Flow selected

Job Flow: j-30FF6EONHJTH9

Last State Change Reason: Terminated by user request

Description

Steps

Bootstrap Actions

Instance Groups

Monitoring

Name: tucson pig 3

Start Date: 2012-09-05 22:17 PDT

Availability Zone: us-east-1a

Master Instance Type: -

Key Name: -

Ami Version: latest

Hadoop Version: 1.0.3

Termination Protected: false

Supported Products: -

Creation Date: 2012-09-05 22:11 PDT

End Date: 2012-09-05 22:26 PDT

Instance Count: -

Slave Instance Type: -

Log URI: s3n://tucson2/log/

Master Public DNS Name: ec2-174-129-129-0.compute-1.amazonaws.com

Keep Alive: true

Subnet Id: -

EMR CONSOLE

Your Elastic MapReduce Job Flows

Region:

US East (Virginia)

Create New Job Flow

Terminate

Debug

Show/Hide

Refresh

Help

Viewing: All

1 to 9 of 9 Job Flows

	Name	State	Creation Date	Elapsed Time	Normalized Instance Hours
<input type="checkbox"/>	tucson pig interactive	TERMINATED	2012-09-05 23:37 PDT	0 hours 12 minutes	3
<input checked="" type="checkbox"/>	tucson pig 3	TERMINATED	2012-09-05 22:11 PDT	0 hours 10 minutes	3

1 Job Flow selected

Job Flow: j-30FF6EONHJTH9

Last State Change Reason: Terminated by user request

Description

Steps

Bootstrap Actions

Instance Groups

Monitoring

Step Name	State	Start Date	End Date	JAR	Main Class	Args
Setup Hadoop Debugging	COMPLETED	2012-09-05 22:17 PDT	2012-09-05 22:17 PDT	s3://elasticmapreduce/libs/script-runner/script-runner.jar	-	s3://elasticmapreduce/libs/state-pusher/0.1/fetch
Setup Pig	COMPLETED	2012-09-05 22:17 PDT	2012-09-05 22:18 PDT	s3://elasticmapreduce/libs/script-runner/script-runner.jar	-	s3://elasticmapreduce/libs/pig/pig-script --base-path s3://elasticmapreduce/libs/pig/ --install-pig --pig-versions latest
Run Pig Script	COMPLETED	2012-09-05 22:18 PDT	2012-09-05 22:20 PDT	s3://elasticmapreduce/libs/script-runner/script-runner.jar	-	s3://elasticmapreduce/libs/pig/pig-script --run-pig-script --pig-versions latest --args -p INPUT=s3://tucson2/ -p OUTPUT=s3://tucson2/output s3://tucson2/tucson-test.pig

EMR CONSOLE

Your Elastic MapReduce Job Flows

Region: US East (Virginia) ▼

Create New Job Flow

Terminate

Debug

Show/Hide

Refresh

Help

Viewing: All

1 to 9 of 9 Job Flows

	Name	State	Creation Date	Elapsed Time	Normalized Instance Hours
<input type="checkbox"/>	tucson pig interactive	TERMINATED	2012-09-05 23:37 PDT	0 hours 12 minutes	3
<input checked="" type="checkbox"/>	tucson pig 3	TERMINATED	2012-09-05 22:11 PDT	0 hours 10 minutes	3

1 Job Flow selected

Job Flow: j-30FF6EONHJTH9

Last State Change Reason: Terminated by user request

Description

Steps

Bootstrap Actions

Instance Groups

Monitoring

Instance Group Id	Role	Instance Type	State	Market	Bid Price	Running Count	Request Count	Creation DateTime	Last State Change Reason
ig-2BIIITM88P87	MASTER	m1.small	ENDED	ON_DEMAND	-	0	1	2012-09-05 22:11 PDT	Job flow terminated
ig-XVXO4GHGRNOA	CORE	m1.small	ENDED	ON_DEMAND	-	0	2	2012-09-05 22:11 PDT	Job flow terminated


EMR CONSOLE


Your Elastic MapReduce Job Flows


Region:  US East (Virginia) ▼

 Create New Job Flow

 Terminate

 Debug



 Show/Hide

 Refresh

 Help

Viewing: All ▼

1 to 9 of 9 Job Flows < >

	Name	State	Creation Date	Elapsed Time	Normalized Instance Hours
<input type="checkbox"/>	tucson pig interactive	 TERMINATED	2012-09-05 23:37 PDT	0 hours 12 minutes	3
<input checked="" type="checkbox"/>	tucson pig 3	 TERMINATED	2012-09-05 22:11 PDT	0 hours 10 minutes	3

1 Job Flow selected

 Job Flow: j-30FF6EONHJTH9

Last State Change Reason: Terminated by user request

Description

Steps

Bootstrap Actions

Instance Groups

Monitoring

Times are displayed in UTC.

Time Range: Last 24 Hours ▼

 Refresh

Avg Map Tasks Running
(Count)



Avg Map Tasks Remaining
(Count)



Avg Map Slots Open (Count)



Avg Remaining Map Tasks / Slot (Count)



Avg Reduce Tasks Running
(Count)



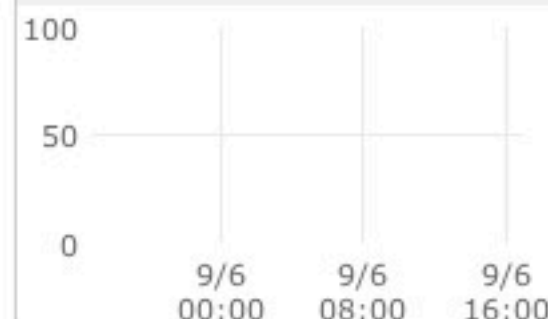
Avg Reduce Tasks Remaining
(Count)



Avg Reduce Slots Open (Count)



Avg HDFS Utilization (Percent)



Avg Missing Blocks (Count)



Avg Jobs Running (Count)



Avg Jobs Failed (Count)

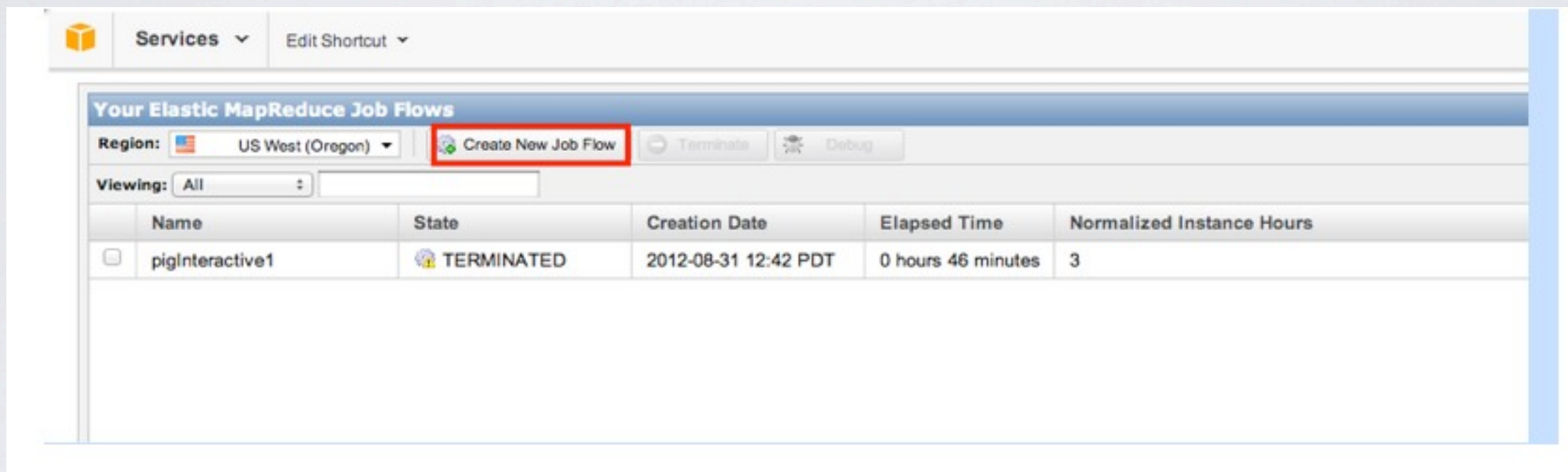


Avg Is Idle? (Count)



I. Setting up Elastic MapReduce

1. You will receive a user ID, password, and private key file(*.pem). You MUST keep it securely. Do NOT share it with others.
2. Log in to the amazon URL you are supplied with, and enter the user ID and password you were given.
3. Go to the Elastic MapReduce Console Menu and click Create New Job Flow.



RUNNING A PIG SCRIPT

I. Setting up The Job Flow

4. Create a Job Flow Name, and Choose Pig Program at Job Type. Then continue.

Create a New Job Flow Cancel

1

2

3

4

5

6

DEFINE JOB FLOW SPECIFY PARAMETERS CONFIGURE EC2 INSTANCES ADVANCED OPTIONS BOOTSTRAP ACTIONS REVIEW

Creating a job flow to process your data using Amazon Elastic MapReduce is simple and quick. Let's begin by giving your job flow a name and selecting its type. If you don't already have an application you'd like to run on Amazon Elastic MapReduce, samples are available to help you get started.

Job Flow Name*:

Job Flow Name doesn't need to be unique. We suggest you give it a descriptive name.

Hadoop Version*:

Create a Job Flow*: ☒ Run your own application
☐ Run a sample application

Run your own application: specify your own parameters for your applications using Hive Program, Custom JAR, Streaming or Pig Program

Run a sample application: by selecting a sample application, parameters will be filled with the necessary data to create a sample Job Flow.

Continue

* Required field

I. Interactive or Script (Batch) Mode?

5a. Now you can choose either to run a script or to run an interactive session. If you are running a script, you need to be sure that the file and directory names in the script refer to the right places in the S3 file system. For the sample script shown above, we use the bucket `s3://tucson2/` and make all our file and directory names relative to that.

Cancel X

DEFINE JOB FLOW SPECIFY PARAMETERS CONFIGURE EC2 INSTANCES ADVANCED OPTIONS BOOTSTRAP ACTIONS REVIEW

Choose between either executing an existing Pig script or starting an interactive Pig session.

☒ Execute a Pig Script

Run a Pig script which has been uploaded to S3. With this option the job flow starts, automatically executes the script, then terminates the job flow automatically when the script has completed.

Script Location*:

The location of your Pig script in Amazon S3.

Input Location:

The URL of the Amazon S3 Bucket that contains the input files.

Output Location:

The URL of the Amazon S3 Bucket to store output files. Should be unique.

Extra Args:

☐ Start an Interactive Pig Session

Start a job flow with Pig setup for interactive use. Interactive use requires you to have an SSH client to access the master host via the user "hadoop". When you are finished your session, manually terminate the job flow from the list of running jobs.

< Back

* Required field

THIS IS AN EXAMPLE: USE YOUR OWN BUCKET TO AVOID CONFLICTS
REMEMBER NOT TO WRITE TO THE SAME DIRECTORY TWICE.

I. Interactive or Script (Batch) Mode?

5b. To choose interactive mode, select “Start an Interactive Pig Session”. Then continue.

Create a New Job Flow

Cancel X

DEFINE JOB FLOW **SPECIFY PARAMETERS** CONFIGURE EC2 INSTANCES ADVANCED OPTIONS BOOTSTRAP ACTIONS REVIEW

Choose between either executing an existing Pig script or starting an interactive Pig session.

☐ Execute a Pig Script

Run a Pig script which has been uploaded to S3. With this option the job flow starts, automatically executes the script, then terminates the job flow automatically when the script has completed.

Script Location*:

The location of your Pig script in Amazon S3.

Input Location:

The URL of the Amazon S3 Bucket that contains the input files.

Output Location:


The URL of the Amazon S3 Bucket to store output files. Should be unique.

Extra Args:

☒ Start an Interactive Pig Session

Start a job flow with Pig setup for interactive use. Interactive use requires you to have an SSH client to access the master host via the user "hadoop". When you are finished your session, manually terminate the job flow from the list of running jobs.

< Back

Continue 

* Required field

I. Configure instance type

6. The instance type is the type of cluster. Bigger instances cost more money!
NEVER choose a type larger than Large and
LEAVE COUNT SET TO 2.

Create a New Job Flow

Cancel X

DEFINE JOB FLOW

SPECIFY PARAMETERS

CONFIGURE EC2 INSTANCES

ADVANCED OPTIONS

BOOTSTRAP ACTIONS

REVIEW

Specify the **Master, Core and Task Nodes** to run your job flow. For more than 20 instances, complete the [limit request form](#).

Master Instance Group: This EC2 instance assigns Hadoop tasks to Core and Task Nodes and monitors their status.

Instance Type: Small (m1.small) ☐ Request Spot Instance

Core Instance Group: These EC2 instances run Hadoop tasks and store data using the Hadoop Distributed File System (HDFS). Recommended for capacity needed for the life of your job flow.

Instance Count: 2

Instance Type: Small (m1.small) ☐ Request Spot Instances

Task Instance Group (Optional): These EC2 instances run Hadoop tasks, but do not persist data. Recommended for capacity needed on a temporary basis.

Instance Count: 0

Instance Type: Small (m1.small) ☐ Request Spot Instances

< Back

Continue

* Required field

I. Select the key pair

7. This is only needed for interactive mode. Enter your private key file from Step 1. If you want a log directory, enter a log path with the s3 pathname, and Enable Debugging. Then continue.

Create a New Job Flow

Cancel X

DEFINE JOB FLOW SPECIFY PARAMETERS CONFIGURE EC2 INSTANCES **ADVANCED OPTIONS** BOOTSTRAP ACTIONS REVIEW

Here you can select an EC2 key pair, configure your cluster to use VPC, set your job flow debugging options, and enter advanced job flow details such as whether it is a long running cluster.

Amazon EC2 Key Pair: Proceed without an EC2 Key Pair

Use an existing Key Pair to SSH into the master node of the Amazon EC2 cluster as the user "hadoop".

Amazon VPC Subnet Id: Proceed without a VPC Subnet ID

Select a Subnet to run this job flow in a Virtual Private Cloud. [Create a VPC](#)

Configure your logging options. [Learn more.](#)

Amazon S3 Log Path (Optional): s3://pig-test-1/log-0831-2

The URL of the Amazon S3 bucket in which your job flow logs will be stored.

Enable Debugging: Yes No

An index of your logs will be stored in Amazon SimpleDB. An Amazon S3 Log Path is required.

Set advanced job flow options.

Keep Alive Yes No

You have selected an interactive session, this job flow will run until manually terminated.

Termination Protection Yes No

< Back

Continue

* Required field

I. No bootstrap.

8. Choose No bootstrap, then Continue.

Create a New Job Flow

Cancel

✓

✓

✓

✓

DEFINE JOB FLOWSPECIFY PARAMETERSCONFIGURE EC2 INSTANCESADVANCED OPTIONSBOOTSTRAP ACTIONSREVIEW

☒ Proceed with no Bootstrap Actions

I do not want to associate any Bootstrap Actions with this Job Flow.

NOTE: Bootstrap Actions must be associated with a Job Flow upon creation. You will not be able to add these later without creating a new Job Flow.

☐ Configure your Bootstrap Actions

[< Back](#)

Continue

* Required field

I. Run the job!

9. Now you'll see the summary of your settings, and if all looks correct, start the job flow!

The screenshot shows the 'Create a New Job Flow' wizard in the AWS Management Console, specifically the 'REVIEW' step. The wizard has six steps: DEFINE JOB FLOW, SPECIFY PARAMETERS, CONFIGURE EC2 INSTANCES, ADVANCED OPTIONS, BOOTSTRAP ACTIONS, and REVIEW. The REVIEW step is currently active, indicated by a highlighted circle and a checkmark. Below the step indicators, a message reads: 'Please review the details of your job flow and click "Create Job Flow" when you are ready to launch your Hadoop Cluster.'

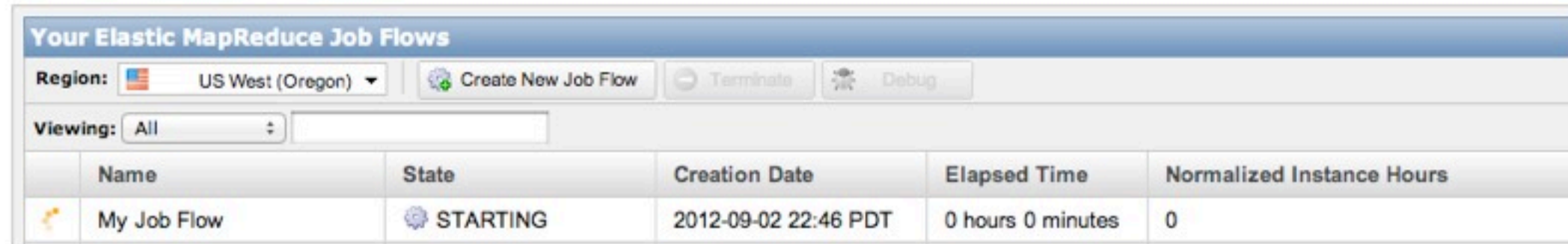
The configuration details are as follows:

- Job Flow Name:** My Job Flow
- Type:** Interactive Pig Session
- Parameters:** Interactive Pig Session has no parameters
- Master Instance Type:** m1.small
- Core Instance Type:** m1.small
- Instance Count:** 1 (for Master), 2 (for Core)
- Amazon EC2 Key Pair:** [Redacted]
- Amazon Subnet Id:** [Redacted]
- Amazon S3 Log Path:** s3://[Redacted]
- Enable Debugging:** Yes
- Keep Alive:** Yes
- Termination Protected:** No
- Bootstrap Actions:** No Bootstrap Actions created for this Job Flow

At the bottom of the wizard, there is a 'Back' link, a 'Create Job Flow' button with a right-pointing arrow, and a note: 'Note: Once you click "Create Job Flow," instances will be launched and you will be charged accordingly.'

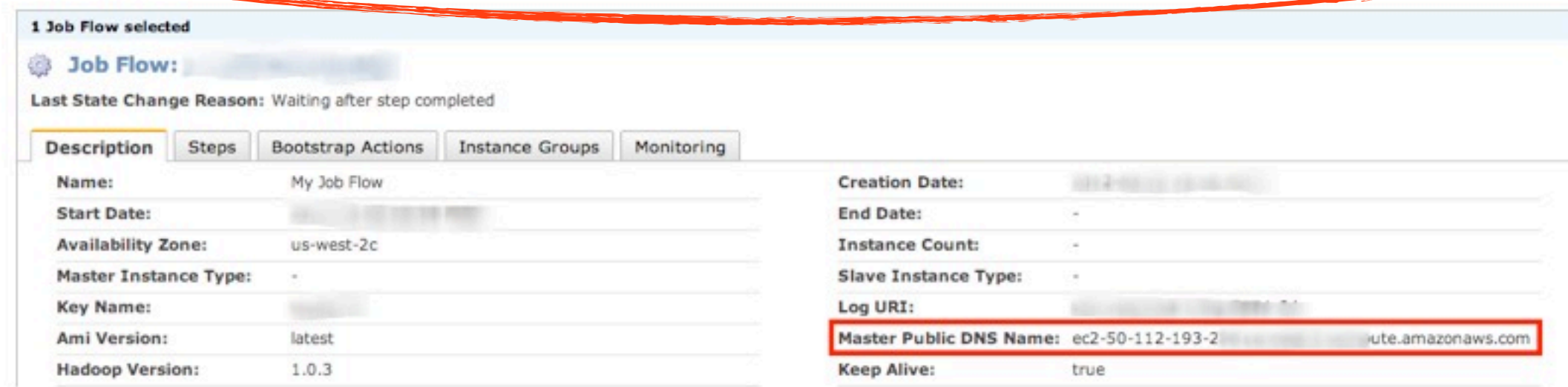
II. Running the Pig Script.

1. Wait for the server to start. You can watch it in the EMR console.



Your Elastic MapReduce Job Flows					
Region: US West (Oregon) Create New Job Flow Terminate Debug					
Viewing: All <input type="text"/>					
	Name	State	Creation Date	Elapsed Time	Normalized Instance Hours
	My Job Flow	STARTING	2012-09-02 22:46 PDT	0 hours 0 minutes	0

2. When the state changes to Waiting, click the Job Flow, and copy Master Public DNS Name from Description.



1 Job Flow selected

Job Flow:

Last State Change Reason: Waiting after step completed

Description Steps Bootstrap Actions Instance Groups Monitoring

Name:	My Job Flow	Creation Date:	2012-09-02 22:46 PDT
Start Date:	2012-09-02 22:46 PDT	End Date:	-
Availability Zone:	us-west-2c	Instance Count:	-
Master Instance Type:	-	Slave Instance Type:	-
Key Name:	my-key	Log URI:	s3://my-bucket-123456789012-us-west-2-logs/
Ami Version:	latest	Master Public DNS Name:	ec2-50-112-193-2-123456789012.us-west-2.compute.amazonaws.com
Hadoop Version:	1.0.3	Keep Alive:	true

II. Running the Pig Script.

3. Now, back on your own computer, open a shell. (For example, terminal on mac, putty on windows.) You can now access the instance using ssh and the Master Public DNS Name from Step II.2 above, using your private key file and the user ID “hadoop”.

A terminal window with a dark background. The prompt is 'amour:~ jooddang\$'. The command entered is 'ssh -i asdf.pem hadoop@ec2-58-112-1west-2.compute.amazonaws.com'. The background of the terminal shows a blurred view of the AWS Management Console with tabs for 'Actions', 'Instance Groups', and 'Monitoring'.

```
amour:~ jooddang$  
amour:~ jooddang$  
amour:~ jooddang$ ssh -i asdf.pem hadoop@ec2-58-112-1west-2.compute.amazonaws.com
```

II. Running the Pig Script.

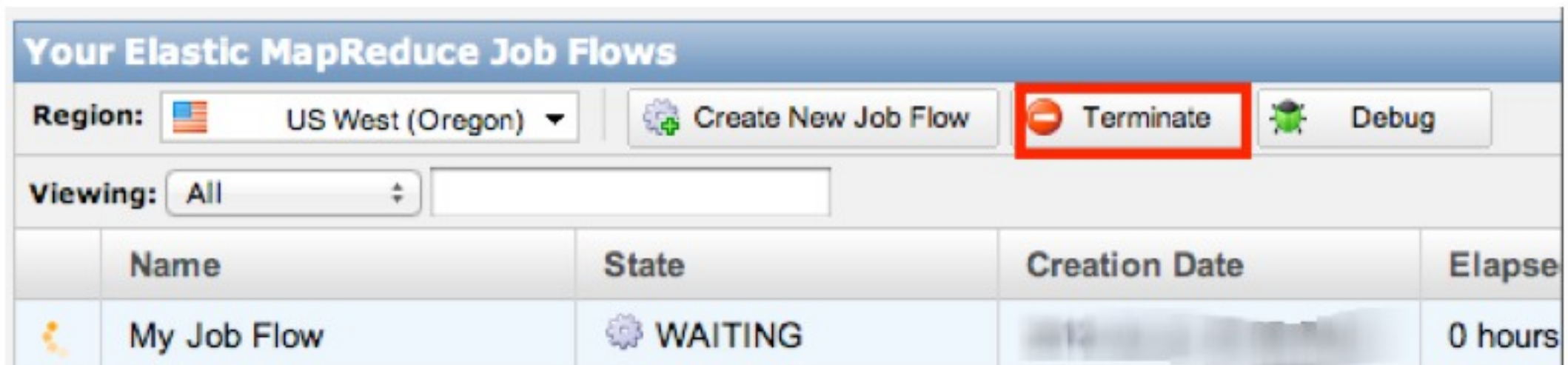
4. Now, you can run the script from your machine, on the AWS cluster.

```
$ pig -p INPUT=s3://<your-bucket>/<your-file-name> -p  
      OUTPUT=s3://<your-bucket>/<output-file-name> s3://<your-bucket>/<your-script-name>
```




III. Terminating the Interactive Session (IMPORTANT!)

1. It is very important that you terminate your interactive session, so we are not charged for time that is not being used!!!



When you finish running the job flow, terminate the Job Flow from the EMR console.



Your Elastic MapReduce Job Flows

Region:  US West (Oregon) ▼  Create New Job Flow **Terminate**  Debug

Viewing: All

	Name	State	Creation Date	Elapse
	My Job Flow	 WAITING		0 hours

USING THE AWS TOOLS SUPPLIED BY THIS CLASS

- This is a PRIVILEGE; we are not making you pay for it.
- It can be EXPENSIVE; we all have to work together to control for this and prevent accidental over expenditures.
- We also will have to control when you have access to the clusters. (We are charged by cluster access time, not cycles.)
- You are only to use the services we describe and for the purposes we describe. You are NOT to use extra time for your own interests. (You can always get your own account.)