Atmosphere
RESEARCH GROUP

Big Data Step-by-Step

Boston Predictive Analytics Big Data Workshop

Microsoft New England Research & Development Center, Cambridge, MA

Saturday, March 10, 2012

by **Jeffrey Breen**

President and Co-Founder
Atmosphere Research Group
email: jeffrey@atmosgrp.com
Twitter: @JeffreyBreen

http://atms.gr/bigdata0310

just need a
little more RAM

~~Big Data~~ Infrastructure

Part 2: Running R + RStudio on Amazon EC2

Code & more on github:

https://github.com/jeffreybreen/tutorial-201203-big-data

Overview

- Sometimes you just need a little more RAM, CPU, or disk space than you have
- Let's try launching an instance on Amazon EC2 and configuring it to do some work
- We'll install R and RStudio and call it a day

Some details we'll skip

- Signing up (it's not that hard)

  http://aws.amazon.com/ec2/

- Pricing (it keeps dropping)

  http://aws.amazon.com/ec2/pricing/

- The alphabet soup of services (we care about EC2 computing and S3 storage)

Just look for biggest button on the page...

AWS Management Console

https://console.aws.amazon.com/ec2/home?region=us-east-1&

AWS Management Console Amazon EC2

Jeffrey Breen | Help

Elastic Beanstalk | S3 | EC2 | VPC | CloudWatch | Elastic MapReduce | CloudFront | CloudFormation | RDS | ElastiCache | SQS | IAM | SNS | SES | Route 53 | More...

Navigation

Region: US East (Virginia)

- EC2 Dashboard
- Scheduled Events
- INSTANCES
  - Instances
  - Spot Requests
  - Reserved Instances
- IMAGES
  - AMIs
  - Bundle Tasks
- ELASTIC BLOCK STORE
  - Volumes
  - Snapshots
- NETWORK & SECURITY
  - Security Groups
  - Elastic IPs
  - Placement Groups
  - Load Balancers
  - Key Pairs
  - Network Interfaces

Amazon EC2 Console Dashboard

Getting Started

To start using Amazon EC2 you will want to launch a virtual server, known as an Amazon EC2 instance.

Launch Instance

Note: Your instances will launch in the US East (Virginia) region.

Service Health

Service Status

Current Status	Details
Amazon EC2 (US East - N. Virginia)	Service is operating normally

View complete service health details

Availability Zone Status

Current Status	Details
us-east-1a	Availability zone is operating normally
us-east-1b	Availability zone is operating normally
us-east-1c	Availability zone is operating normally
us-east-1d	Availability zone is operating normally

My Resources

You are using the following Amazon EC2 resources in the US East (Virginia) region: Refresh

- 2 Running Instances
- 2 Elastic IPs
- 4 EBS Volumes
- 0 EBS Snapshots
- 2 Key Pairs
- 0 Load Balancers
- 0 Placement Groups
- 6 Security Groups

Scheduled Events

US East (Virginia): No events Refresh

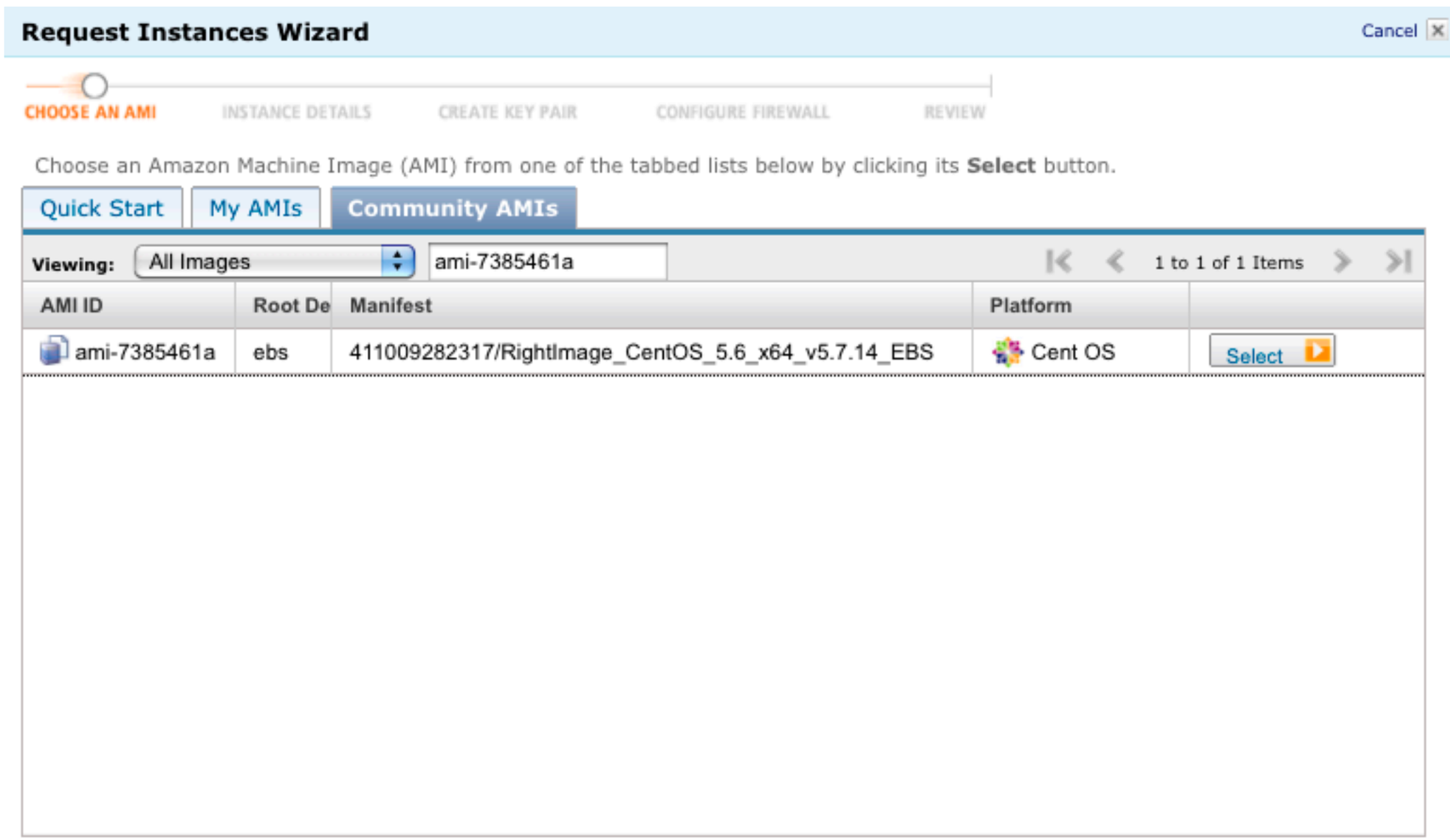
Related Links

- Getting Started Guide
- Documentation
- All EC2 Resources
- Forums
- Feedback
- Report an Issue

 | Feedback | Support | Privacy Policy | Terms of Use | An amazon.com company

Select an Amazon Machine Image

`ami-7385461a` is a good, recent 64-bit CentOS image published by RightScale



Only use EBS images

- Instance-storage machines lose their data upon shutdown (termination)
- EBS instances can be stopped and restarted, or terminated when you're done forever

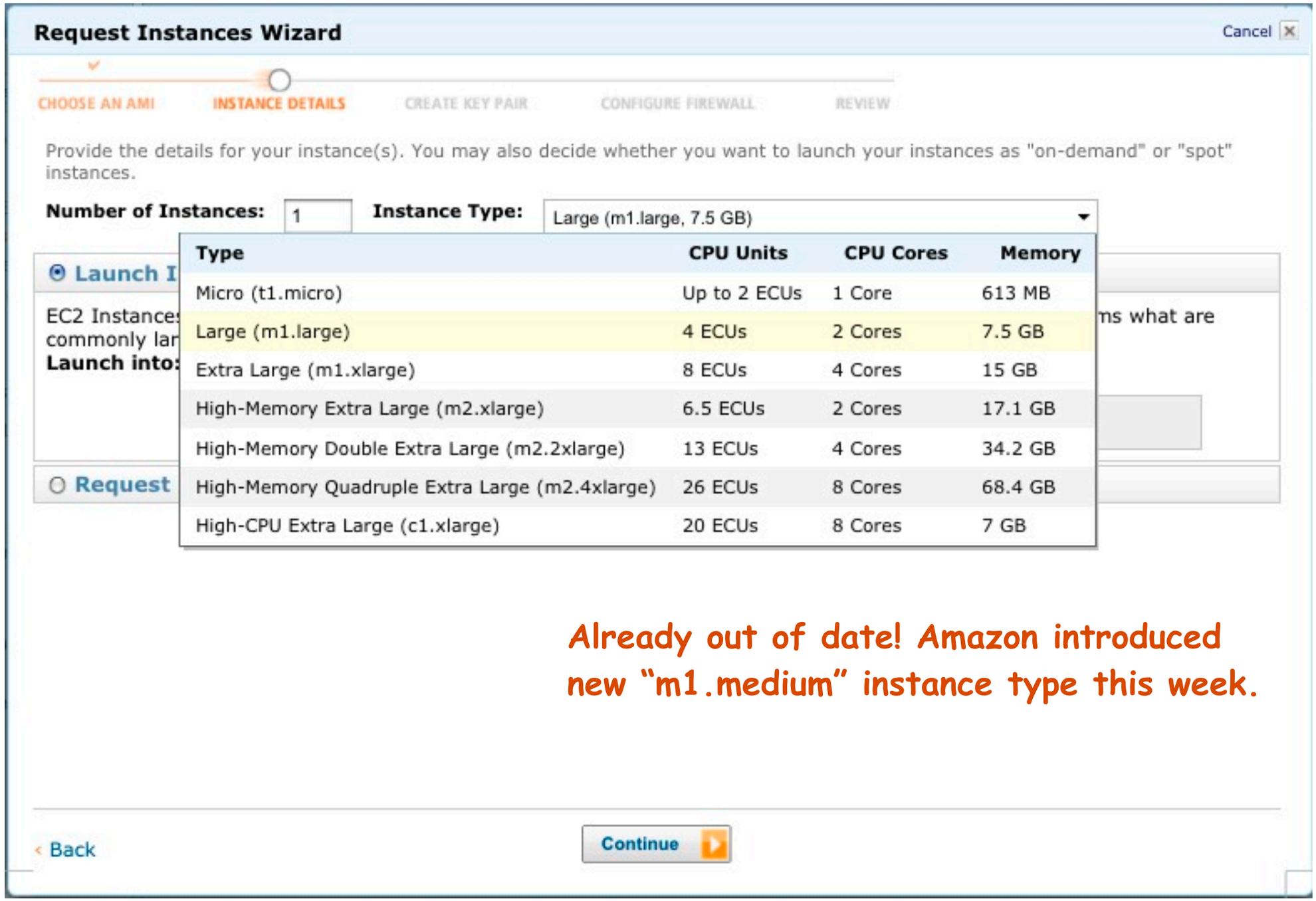
EC2 Amazon Machine Image: ami-7385461a

Description | Tags

AMI ID:	ami-7385461a				
AMI Name:	RightImage_CentOS_5.6_x64_v5.7.14_EBS				
Description:	RightImage_CentOS_5.6_x64_v5.7.14_EBS				
Source:	411009282317/RightImage_CentOS_5.6_x64_v5.7.14_EBS				
Owner:	411009282317	**Visibility:**	Public	**Product Code:**	
State:	available	**Kernel ID:**	aki-825ea7eb	**RAM Disk ID:**	-
Image Type:	machine	**Architecture:**	x86_64	**Platform:**	Cent OS
Root Device Type:	ebs	**Root Device:**	/dev/sda1	**Image Size:**	8 GiB
Block Devices:	/dev/sda1=snap-0eb8296e:8:true				
Virtualization:	paravirtual				
State Reason:	-				

Pick a size

See http://aws.amazon.com/ec2/instance-types/

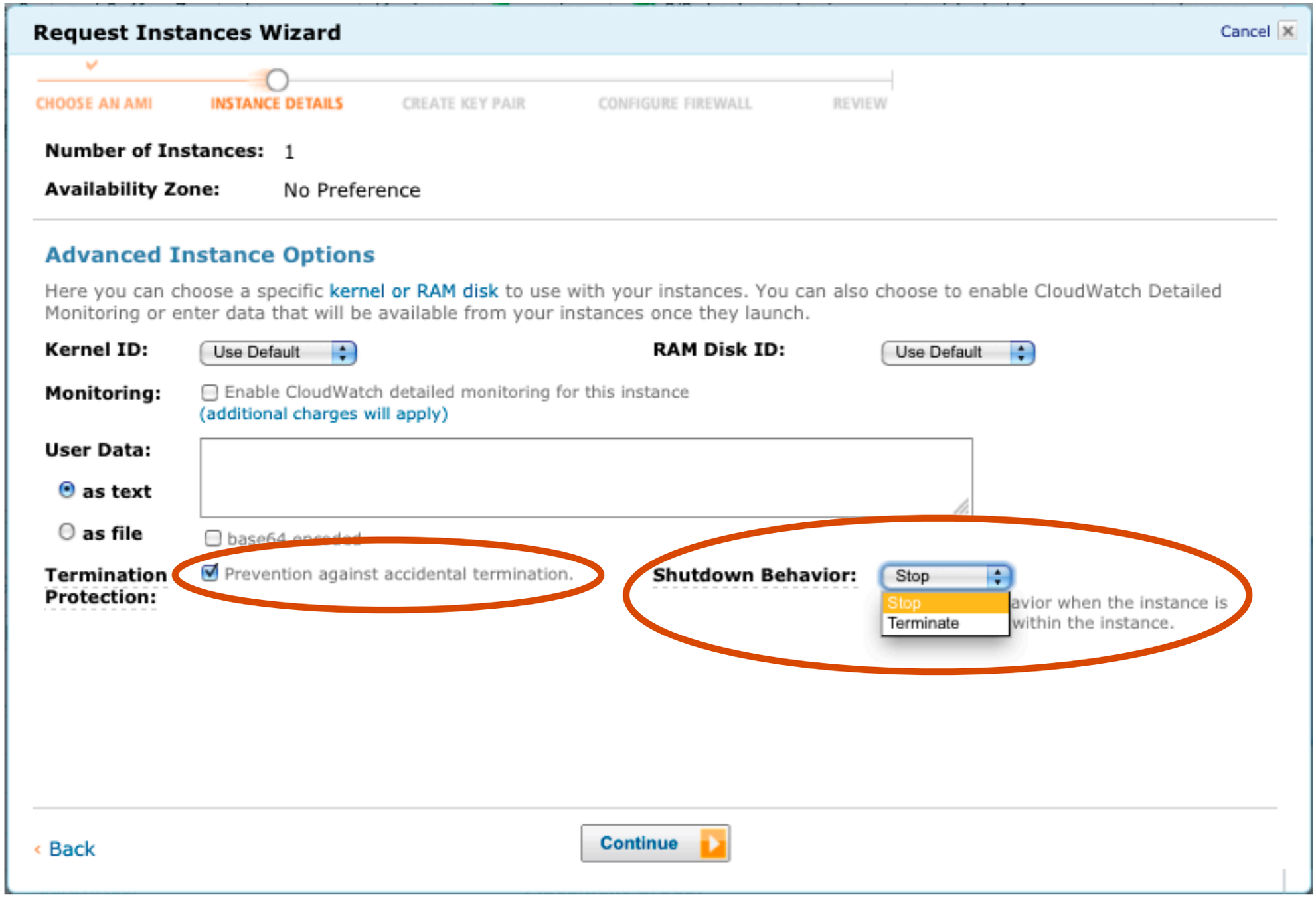


Type	CPU Units	CPU Cores	Memory
Micro (t1.micro)	Up to 2 ECUs	1 Core	613 MB
Large (m1.large)	4 ECUs	2 Cores	7.5 GB
Extra Large (m1.xlarge)	8 ECUs	4 Cores	15 GB
High-Memory Extra Large (m2.xlarge)	6.5 ECUs	2 Cores	17.1 GB
High-Memory Double Extra Large (m2.2xlarge)	13 ECUs	4 Cores	34.2 GB
High-Memory Quadruple Extra Large (m2.4xlarge)	26 ECUs	8 Cores	68.4 GB
High-CPU Extra Large (c1.xlarge)	20 ECUs	8 Cores	7 GB

Already out of date! Amazon introduced new "m1.medium" instance type this week.

Avoid Premature Termination

Set Termination Protection + Shutdown Behavior



Name your instance

Request Instances Wizard Cancel

CHOOSE AN AMI | INSTANCE DETAILS | CREATE KEY PAIR | CONFIGURE FIREWALL | REVIEW

Add tags to your instance to simplify the administration of your EC2 infrastructure. A form of metadata, tags consist of a case-sensitive key/value pair, are stored in the cloud and are private to your account. You can create user-friendly names that help you organize, search, and browse your resources. For example, you could define a tag with key = Name and value = Webserver. You can add up to 10 unique keys to each instance along with an optional value for each key. For more information, go to Using Tags in the *EC2 User Guide*.

Key (127 characters maximum)	**Value** (255 characters maximum)	**Remove**
Name	rstudio	✘
		✘

Add another Tag. (Maximum of 10)

‹ Back Continue

Create a key pair

Don’t forget to download it (and keep it safe!)

Request Instances Wizard Cancel

CHOOSE AN AMI | INSTANCE DETAILS | CREATE KEY PAIR | CONFIGURE FIREWALL | REVIEW

Public/private key pairs allow you to securely connect to your instance after it launches. To create a key pair, enter a name and click **Create & Download your Key Pair**. You will then be prompted to save the private key to your computer. Note, you only need to generate a key pair once - not each time you want to deploy an Amazon EC2 instance.

- Choose from your existing Key Pairs
- Create a new Key Pair
  - 1. **Enter a name for your key pair:*** rstudio-ec2 (e.g., jdoekey)
  - 2. **Click to create your key pair:*** Create & Download your Key Pair

    Save this file in a place you will remember. You can use this key pair to launch other instances in the future or visit the Key Pairs page to create or manage existing ones.
- Proceed without a Key Pair

‹ Back Continue

Create a Security Group

All TCP, UDP and ICMP from your IP address

Request Instances Wizard Cancel

CHOOSE AN AMI ✓ INSTANCE DETAILS ✓ CREATE KEY PAIR ✓ CONFIGURE FIREWALL REVIEW

Security groups determine whether a network port is open or blocked on your instances. You may use an existing security group, or we can help you create a new security group to allow access to your instances using the suggested ports below. Add additional ports now or update your security group anytime using the Security Groups page.

○ **Choose one or more of your existing Security Groups**

◉ **Create a new Security Group**

Group Name rstudio

Group Description only from my IP

Inbound Rules

Create a new rule: All ICMP

Source: 74.104.166.41/32

(e.g., 192.168.2.0/24, sg-47ad482e, or 1234567890/default)

Add Rule

TCP		
Port (Service)	Source	Action
0 - 65535	74.104.166.41/32	Delete

UDP		
Port (Service)	Source	Action
0 - 65535	74.104.166.41/32	Delete

‹ Back

Continue

Don’t know your IP address?

Don’t ask me. Ask Google!

(simply append “/32” when entering into firewall rules)

3... 2... 1...

Request Instances Wizard Cancel

CHOOSE AN AMI | INSTANCE DETAILS | CREATE KEY PAIR | CONFIGURE FIREWALL | REVIEW

Please review the information below, then click **Launch**.

AMI: Cent OS AMI ID ami-7385461a (x86_64) Edit AMI

Number of Instances: 1
Availability Zone: No Preference
Instance Type: Large (m1.large)
Instance Class: On Demand Edit Instance Details

Monitoring: Disabled
Termination Protection: Enabled
Tenancy: Default
Kernel ID: Use Default
Shutdown Behavior: Stop
RAM Disk ID: Use Default
User Data: Edit Advanced Details

Key Pair Name: rstudio-ec2 Edit Key Pair

Security Group(s): sg-cfe234a7 Edit Firewall

‹ Back Launch

State = running

Up and running at specified domain name

My Instances

Launch Instance | Instance Actions | Show/Hide | Refresh | Help

Viewing: All Instances | All Instance Types | Search | 1 to 4 of 4 Instances

Name	Instance	AMI ID	Root Device	Type	State	Status Checks	Monitoring	Security Groups
rstudio	i-4f59882b	ami-7385461a	ebs	t1.micro	running	2/2 checks pa	basic	rstudio

1 EC2 Instance selected.

EC2 Instance: rstudio (i-4f59882b) ec2-107-22-109-130.compute-1.amazonaws.com

Description | Status Checks | Monitoring | Tags

AMI: RightImage_CentOS_5.6_x64_v5.7.14_EBS (ami-7385461a)		Zone:	us-east-1b
Security Groups:	rstudio	Type:	t1.micro
State:	running	Scheduled Events:	No scheduled events
Owner:	581302678308	VPC ID:	-
Subnet ID:	-	Source/Dest. Check:	
Virtualization:	paravirtual	Placement Group:	
Reservation:	r-a9fbcbc9	RAM Disk ID:	-
Platform:	-	Key Pair Name:	rstudio-ec2
Kernel ID:	aki-825ea7eb	Monitoring:	basic
AMI Launch Index:	0	Elastic IP:	-
Root Device:	sda1	Root Device Type:	ebs
Tenancy:	default		
Lifecycle:	normal		
Block Devices:	sda1		

Time to get all command line

- You'll need an ssh client and the key pair we generated in order to connect with your instance
- We'll use the Cloudera VM to control versions, options, etc.
- ssh won't use your key pair if its file permissions are too lax

```
$ chmod og-rwx rstudio-ec2.pem
```

- Log in as root to your domain name

```
$ ssh -i rstudio-ec2.pem root@YOURDOMAINHERE.amazonaws.com
```

(from previous slide)

Install R and RStudio

- Create a user login for yourself (RStudio needs this)

```
# useradd jbreen
# passwd jbreen
```

- EPEL is already installed, so R is easy

```
# yum -y install R
```

- Follow RStudio's download instructions

  http://www.rstudio.org/download/server

```
# wget http://download2.rstudio.org/rstudio-server-0.95.262-x86_64.rpm
# rpm -Uvh rstudio-server-0.95.262-x86_64.rpm
```

- Browse to port 8787 and use the login and password

  e.g., http://ec2-107-22-109-130.compute-1.amazonaws.com:8787/

Success!

RStudio

http://ec2-107-22-109-130.compute-1.amazonaws.com:8787/

File Edit View Project Workspace Plots Tools Help

jbreen | Sign Out

Go to file/function

Project: (None)

mtcars

32 observations of 12 variables

	row.names	mpg	cyl	disp	hp	drat	wt	qse
1	Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.
2	Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.
3	Datsun 710	22.8	4	108.0	93	3.85	2.320	18.
4	Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.

Workspace History

Load Save Import Dataset Clear All

Data

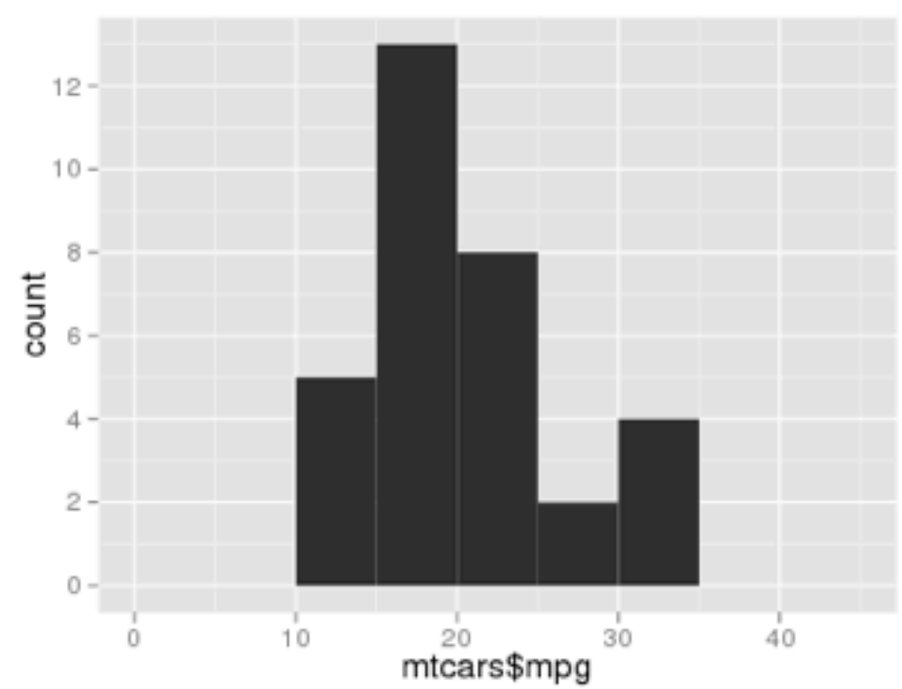
mtcars 32 obs. of 11 variables

Console ~/

```
ISBN 3-900051-07-0
Platform: x86_64-redhat-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> data(mtcars)
> View(mtcars)
> library(ggplot2)
> qplot(mtcars$mpg, binwidth=5)
>
```

Files Plots Packages Help

Zoom Export Clear All



The meter's running

- Amazon charges by the hour (or fraction thereof). So when you're done, you should probably shutdown
- via command line

```
$ sudo shutdown -h now
```

- or with the "Stop" Instance Action in the AWS Management Console
- (use "Terminate" if you never want to use it again)

Next up:

How to launch Hadoop clusters in the cloud without really trying