
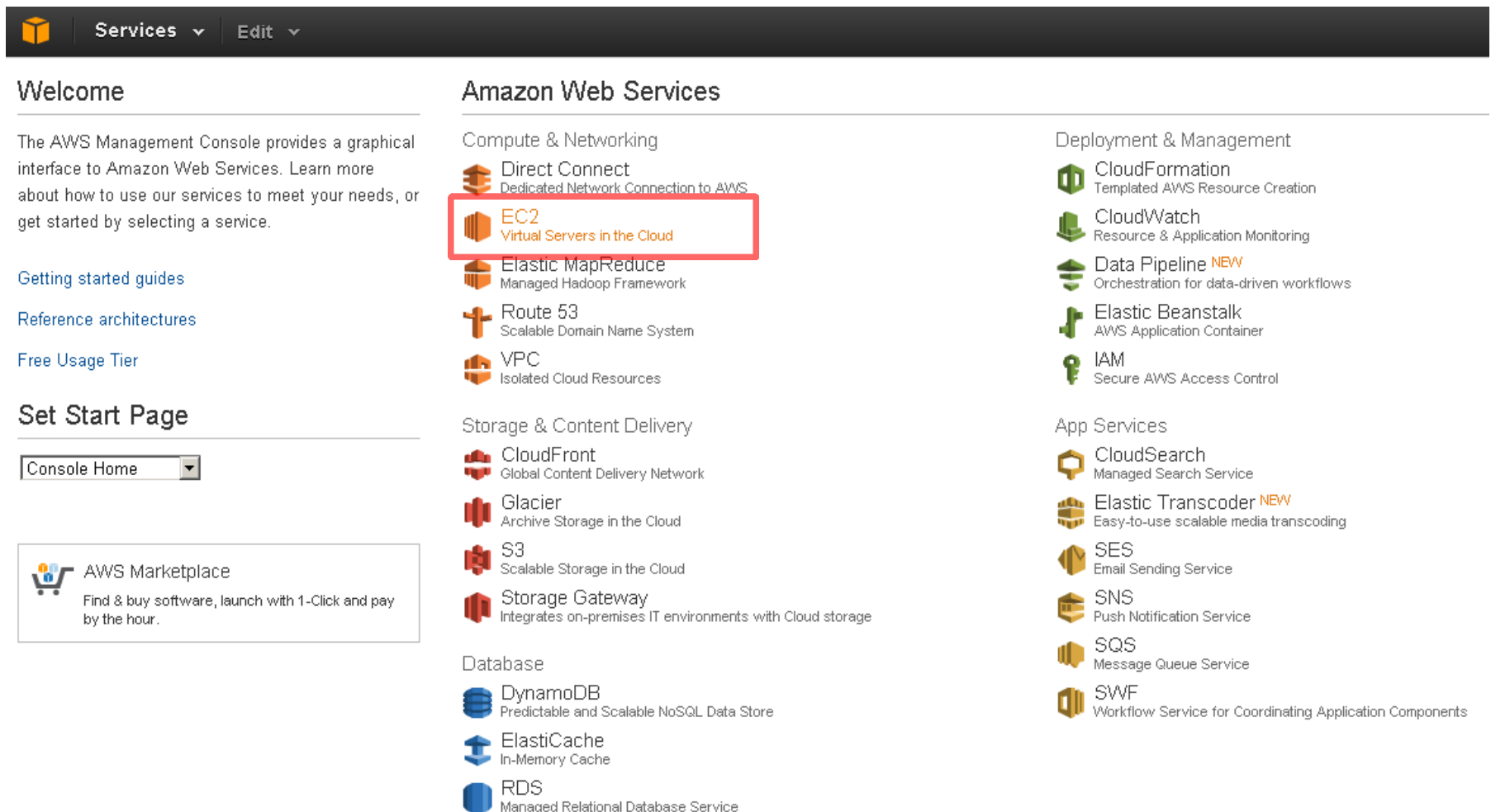


# **Hadoop and AWS**

# Developing with Hadoop in the AWS cloud

- **Hadoop** is Linux based.
- You can install Linux at home and run these examples.
- We will create a Linux instance using **AWS** and **EC2** to run our code.

- Log in to your **AWS account**. 
- Select the **EC2** service.



The screenshot shows the AWS Management Console interface. At the top, there is a navigation bar with the AWS logo, 'Services', and 'Edit' dropdown menus. Below this, the 'Welcome' section provides introductory text and links for getting started. The main area is titled 'Amazon Web Services' and is organized into several categories: Compute & Networking, Storage & Content Delivery, Database, Deployment & Management, and App Services. The 'EC2' service, described as 'Virtual Servers in the Cloud', is highlighted with a red rectangular box. Other services listed include Direct Connect, Elastic MapReduce, Route 53, VPC, CloudFormation, CloudWatch, Data Pipeline, Elastic Beanstalk, IAM, CloudSearch, Elastic Transcoder, SES, SNS, SQS, SWF, DynamoDB, ElastiCache, and RDS.

**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 Marketplace**  
Find & buy software, launch with 1-Click and pay by the hour.

**Amazon Web Services**

**Compute & Networking**

- Direct Connect  
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  
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
- RDS  
Managed Relational Database Service

**Deployment & Management**

- CloudFormation  
Templated AWS Resource Creation
- CloudWatch  
Resource & Application Monitoring
- Data Pipeline **NEW**  
Orchestration for data-driven workflows
- Elastic Beanstalk  
AWS Application Container
- IAM  
Secure AWS Access Control

**App Services**

- CloudSearch  
Managed Search Service
- Elastic Transcoder **NEW**  
Easy-to-use scalable media transcoding
- SES  
Email Sending Service
- SNS  
Push Notification Service
- SQS  
Message Queue Service
- SWF  
Workflow Service for Coordinating Application Components

- Click on **Launch Instance**

The screenshot shows the AWS Management Console interface. At the top, there is a navigation bar with the AWS logo, 'Services', and 'Edit' dropdown menus. On the left, a sidebar contains the 'EC2 Dashboard' with a list of categories: INSTANCES (Events, Instances, Spot Requests, Reserved Instances), IMAGES (AMIs, Bundle Tasks), ELASTIC BLOCK STORE (Volumes, Snapshots), and NETWORK & SECURITY (Security Groups, Elastic IPs, Placement Groups, Load Balancers, Key Pairs, Network Interfaces). The main content area is titled 'Resources' and lists EC2 resources in the US East (N. Virginia) region: 0 Running Instances, 0 Elastic IPs, 0 Volumes, 0 Snapshots, 2 Key Pairs, 0 Load Balancers, 0 Placement Groups, and 9 Security Groups. Below this is the 'Create Instance' section, which includes a 'Launch Instance' button highlighted with a red box. A note states: 'Your instances will launch in the US East (N. Virginia) region.' To the right of the 'Create Instance' section is the 'Scheduled Events' section, which shows 'US East (N. Virginia): No events'. At the bottom of the 'Create Instance' section is a 'Service Health' section, which includes a 'Service Status' (US East (N. Virginia): This service is operating normally) and an 'Availability Zone Status' (us-east-1a, us-east-1b, and us-east-1d are all operating normally). A 'Service Health Dashboard' link is provided at the bottom of the 'Service Health' section.

**Services** ▾ **Edit** ▾

**EC2 Dashboard**

- 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

**Resources**

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


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

**Create Instance**

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 (N. Virginia) region

**Service Health**  **Scheduled Events**

**Service Status:**

- ✓ US East (N. Virginia): This service is operating normally

**Availability Zone Status:**

- ✓ us-east-1a Availability zone is operating normally
- ✓ us-east-1b Availability zone is operating normally
- ✓ us-east-1d Availability zone is operating normally

[Service Health Dashboard](#)

**US East (N. Virginia):**  
No events

- ~~Click on Quick Launch Wizard~~
- Select **Ubuntu Server 14.04 LTS**

The screenshot shows the AWS IAM console interface for selecting an Amazon Machine Image (AMI). The page is titled "Step 1: Choose an Amazon Machine Image (AMI)" and includes a "Cancel and Exit" link. The main content area displays a list of AMIs with the following details:

OS/Provider	AMI Name	AMI ID	Architecture	Action
Amazon Linux	Amazon Linux AMI 2014.09.1 (HVM)	ami-6e7bd919	64-bit	Select
Red Hat	Red Hat Enterprise Linux 7.0 (HVM), SSD Volume Type	ami-8cff51fb	64-bit	Select
SUSE Linux	SUSE Linux Enterprise Server 11 SP3 (HVM), SSD Volume Type	ami-30842747	64-bit	Select
Ubuntu	Ubuntu Server 14.04 LTS (HVM), SSD Volume Type	ami-f0b11187	64-bit	Select
Windows	Microsoft Windows Server 2012 R2 Base	ami-d4228ea3	64-bit	Select
Windows	Microsoft Windows Server 2012 R2 with SQL Server Web	ami-d23d91a5	64-bit	Select

The "Select" button for the Ubuntu Server 14.04 LTS AMI is highlighted with a red box. The page also includes a "Quick Start" sidebar on the left and a "Free tier only" filter option.

- Click **Continue**

- Click on **Review and Launch**.

Services Edit Mr M Harris Ireland Support

1. Choose AMI 2. Choose Instance Type 3. Configure Instance 4. Add Storage 5. Tag Instance 6. Configure Security Group 7. Review

### Step 2: Choose an Instance Type

Amazon EC2 provides a wide selection of instance types optimized to fit different use cases. Instances are virtual servers that can run applications. They have varying combinations of CPU, memory, storage, and networking capacity, and give you the flexibility to choose the appropriate mix of resources for your applications. [Learn more](#) about instance types and how they can meet your computing needs.

Filter by: All instance types Current generation Show/Hide Columns

Currently selected: t2.micro (Variable ECUs, 1 vCPUs, 2.5 GHz, Intel Xeon Family, 1 GiB memory, EBS only)

T2 instances are VPC-only. Your T2 instance will launch into your VPC. [Learn more](#) about T2 and VPC.

	Family	Type	vCPUs	Memory (GiB)	Instance Storage (GB)	EBS-Optimized Available	Network Performance
<input checked="" type="checkbox"/>	General purpose	t2.micro Free tier eligible	1	1	EBS only	-	Low to Moderate
<input type="checkbox"/>	General purpose	t2.small	1	2	EBS only	-	Low to Moderate
<input type="checkbox"/>	General purpose	t2.medium	2	4	EBS only	-	Low to Moderate
<input type="checkbox"/>	General purpose	m3.medium	1	3.75	1 x 4 (SSD)	-	Moderate
<input type="checkbox"/>	General purpose	m3.large	2	7.5	1 x 32 (SSD)	-	Moderate
<input type="checkbox"/>	General purpose	m3.xlarge	4	15	2 x 40 (SSD)	Yes	High
<input type="checkbox"/>	General purpose	m3.2xlarge	8	30	2 x 80 (SSD)	Yes	High
<input type="checkbox"/>	Compute optimized	c3.large	2	3.75	2 x 16 (SSD)	-	Moderate
<input type="checkbox"/>	Compute optimized	c3.xlarge	4	7.5	2 x 40 (SSD)	Yes	Moderate
<input type="checkbox"/>	Compute optimized	c3.2xlarge	8	15	2 x 80 (SSD)	Yes	High

Cancel Previous **Review and Launch** Next: Configure Instance Details

- Click on **Launch** to start the instance (this can take a few seconds).

Services Edit Mr M Harris Ireland Support

1. Choose AMI 2. Choose Instance Type 3. Configure Instance 4. Add Storage 5. Tag Instance 6. Configure Security Group 7. Review

### Step 7: Review Instance Launch

Please review your instance launch details. You can go back to edit changes for each section. Click **Launch** to assign a key pair to your instance and complete the launch process.

**⚠ Improve your instance's security. Your security group, launch-wizard-7, is open to the world.**

Your instance may be accessible from any IP address. We recommend that you update your security group rules to allow access from known IP addresses only. You can also open additional ports in your security group to facilitate access to the application or service you're running, e.g., HTTP (80) for web servers. [Edit security groups](#)

▼ AMI Details [Edit AMI](#)

**Ubuntu Server 14.04 LTS (HVM), SSD Volume Type - ami-f0b11187**

Free tier eligible  
Ubuntu Server 14.04 LTS (HVM), EBS General Purpose (SSD) Volume Type. Support available from Canonical (<http://www.ubuntu.com/cloud/services>).  
Root Device Type: ebs Virtualization type: hvm

▼ Instance Type [Edit instance type](#)

Instance Type	ECUs	vCPUs	Memory (GiB)	Instance Storage (GB)	EBS-Optimized Available	Network Performance
t2.micro	Variable	1	1	EBS only	-	Low to Moderate

▼ Security Groups [Edit security groups](#)

**Security group name** launch-wizard-7  
**Description** launch-wizard-7 created 2014-10-31T11:09:52.008+00:00

Type ⓘ	Protocol ⓘ	Port Range ⓘ	Source ⓘ
SSH	TCP	22	0.0.0.0/0

▶ Instance Details [Edit instance details](#)

▶ Storage [Edit storage](#)

▶ Tags [Edit tags](#)

Cancel Previous **Launch**

- Create a new key pair.
- Give it a name.
- Click **Download Key Pair** and save the file somewhere you can find it easily.

### Select an existing key pair or create a new key pair ✕


A key pair consists of a **public key** that AWS stores, and a **private key file** that you store. Together, they allow you to connect to your instance securely. For Windows AMIs, the private key file is required to obtain the password used to log into your instance. For Linux AMIs, the private key file allows you to securely SSH into your instance.

Note: The selected key pair will be added to the set of keys authorized for this instance. Learn more about [removing existing key pairs from a public AMI](#).

Create a new key pair ▾

**Key pair name**  
mykeypair

**Download Key Pair**

 You have to download the **private key file** (\*.pem file) before you can continue. **Store it in a secure and accessible location.** You will not be able to download the file again after it's created.

[Cancel](#) [Launch Instances](#)

- Click **Launch Instance**.



- Click **View Instance**.

## Launch Status

### ✓ Your instance is now launching

The following instance launch has been initiated: [i-57d2fcb2](#) [View launch log](#)

### 💬 Get notified of estimated charges

[Create billing alerts](#) to get an email notification when estimated charges on your AWS bill exceed an amount you define (for example, if you exceed the free usage tier).

## How to connect to your instance

Your instance is launching, and it may take a few minutes until it is in the **running** state, when it will be ready for you to use. Usage hours on your new instance will start immediately and continue to accrue until you stop or terminate your instance.

Click **View Instances** to monitor your instance's status. Once your instance is in the **running** state, you can **connect** to it from the Instances screen. [Find out](#) how to connect to your instance.

### ▼ Here are some helpful resources to get you started

- [How to connect to your Linux instance](#)
- [Amazon EC2: User Guide](#)
- [Learn about AWS Free Usage Tier](#)
- [Amazon EC2: Discussion Forum](#)

While your instances are launching you can also

[Create status check alarms](#) to be notified when these instances fail status checks. (Additional charges may apply)

[Create and attach additional EBS volumes](#) (Additional charges may apply)

[Manage security groups](#)

[View Instances](#)

- Click the instance (it'll have a green light next to it), to display information about it.

The screenshot shows the AWS Management Console interface. On the left is a navigation menu with categories like INSTANCES, IMAGES, ELASTIC BLOCK STORE, NETWORK & SECURITY, and AUTO SCALING. The main area displays a table of EC2 instances. One instance is highlighted, and its details are shown below. The instance is in a 'running' state, indicated by a green dot. The public IP address is 54.171.121.255.

Name	Instance ID	Instance Type	Availability Zone	Instance State	Status Checks	Alarm Status	Public DNS	Public IP	Key Name	Monitoring
	i-57d2fcb2	t2.micro	eu-west-1a	running	Initializing	None		54.171.121.255	mykeypair	disabled

Instance: i-57d2fcb2    Public IP: 54.171.121.255

Description		Status Checks		Monitoring		Tags	
Instance ID	i-57d2fcb2	Public DNS	-				
Instance state	running	Public IP	54.171.121.255				

Our instance is now running.

This will be important in a minute

- Click on the **Security Groups** link.

- Select the '~~quicklaunch-1~~' group.

The screenshot shows the AWS Management Console interface for Security Groups. The left-hand navigation pane includes sections for INSTANCES, IMAGES, ELASTIC BLOCK STORE, NETWORK & SECURITY, and AUTO SCALING. Under NETWORK & SECURITY, the 'Security Groups' option is selected. The main content area displays a table of security groups. The group 'launch-wizard-7' with ID 'sg-955ae2f0' is highlighted. Below the table, the details for this group are shown, with the 'Description' tab selected. The details include the group name 'launch-wizard-7' and the group ID 'sg-955ae2f0'.

Name	Group ID	Group Name	VPC ID	Description
myhadoop-master	sg-57df6b20	myhadoop-master		Group for Hadoop Master.
myhadoop	sg-5bdf6b2c	myhadoop		Group for Hadoop Slaves.
ElasticMapReduce-slave	sg-7d20890a	ElasticMapReduce-slave		Slave group for Elastic MapReduce
ElasticMapReduce-master	sg-7f208908	ElasticMapReduce-master		Master group for Elastic MapReduce
hadoopy	sg-8544f1f2	hadoopy		Group for Hadoop Slaves.
launch-wizard-2	sg-89eb94fe	launch-wizard-2		launch-wizard-2 created on Friday, November 1, 2013 6:12:09 PM UTC
launch-wizard-7	sg-955ae2f0	launch-wizard-7	vpc-64519701	launch-wizard-7 created 2014-10-31T11:09:52.008+00:00
hadoopy-master	sg-a144f1d6	hadoopy-master		Group for Hadoop Master.
launch-wizard-1	sg-b3324ec4	launch-wizard-1		launch-wizard-1 created on Friday, October 25, 2013 12:57:13 PM UTC+1
launch-wizard-6	sg-ba46ffdf	launch-wizard-6	vpc-64519701	launch-wizard-6 created 2014-10-29T15:47:18.853+00:00
default	sg-c6ca6eb1	default		default group
launch-wizard-4	sg-c853edad	launch-wizard-4	vpc-64519701	launch-wizard-4 created 2014-10-27T14:31:36.734+00:00
hadooptest-master	sg-d773c6a0	hadooptest-master		Group for Hadoop Master.
hadooptest	sg-db73c6ac	hadooptest		Group for Hadoop Slaves.

**Security Group: sg-955ae2f0**

**Description** | Inbound | Outbound | Tags

<b>Group name</b>	launch-wizard-7	<b>Group description</b>	launch-wizard-7 created 2014-10-31T11:09:52.008+00:00
<b>Group ID</b>	sg-955ae2f0	<b>VPC ID</b>	vpc-64519701

- Select the 'Inbound' tab.

Services Edit Mr M Harris Ireland Support

EC2 Dashboard  
Events  
Tags  
Reports  
Limits

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

AUTO SCALING  
Launch Configurations  
Auto Scaling Groups

Create Security Group Actions

Filter by tags and attributes or search by keyword

Name	Group ID	Group Name	VPC ID	Description
<input type="checkbox"/>	sg-57d6b20	myhadoop-master		Group for Hadoop Master.
<input type="checkbox"/>	sg-5bd6b2c	myhadoop		Group for Hadoop Slaves.
<input type="checkbox"/>	sg-7d20890a	ElasticMapReduce-slave		Slave group for Elastic MapReduce
<input type="checkbox"/>	sg-7f208908	ElasticMapReduce-master		Master group for Elastic MapReduce
<input type="checkbox"/>	sg-8544f1f2	hadoopy		Group for Hadoop Slaves.
<input type="checkbox"/>	sg-89eb94fe	launch-wizard-2		launch-wizard-2 created on Friday, November 1, 2013 6:12:09 PM UTC
<input checked="" type="checkbox"/>	sg-955ae2f0	launch-wizard-7	vpc-64519701	launch-wizard-7 created 2014-10-31T11:09:52.008+00:00
<input type="checkbox"/>	sg-a144f1d6	hadoopy-master		Group for Hadoop Master.
<input type="checkbox"/>	sg-b3324ec4	launch-wizard-1		launch-wizard-1 created on Friday, October 25, 2013 12:57:13 PM UTC+1
<input type="checkbox"/>	sg-ba46ffdf	launch-wizard-6	vpc-64519701	launch-wizard-6 created 2014-10-29T15:47:18.853+00:00
<input type="checkbox"/>	sg-c6ca6eb1	default		default group
<input type="checkbox"/>	sg-c853edad	launch-wizard-4	vpc-64519701	launch-wizard-4 created 2014-10-27T14:31:36.734+00:00
<input type="checkbox"/>	sg-d773c6a0	hadooptest-master		Group for Hadoop Master.
<input type="checkbox"/>	sg-db73c6ac	hadooptest		Group for Hadoop Slaves.

Security Group: sg-955ae2f0

Description Inbound Outbound Tags

Edit

Type	Protocol	Port Range	Source
SSH	TCP	22	0.0.0.0/0

Make sure you have this rule. We'll be logging in through **port 22** in a minute.



EC2 Dashboard

Events

Tags

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

[Launch Instance](#) [Connect](#) [Actions ▾](#)



Filter: [All instances ▾](#) [All instance types ▾](#)  ×

1 to 1 of 1 Instances

<input type="checkbox"/>	Name	Instance ID	Instance Type	Availability Zone	Instance State	Status Checks	Alarm Status
<input type="checkbox"/>		i-3193ce49	t1.micro	us-east-1b	<span style="color: green;">●</span> running	Initializing	None

[Create Status Check Alarm](#)

### System Status Checks ⓘ

These checks monitor the AWS systems required to use this instance and ensure they are functioning properly.

### Instance Status Checks ⓘ

These checks monitor your software and network configuration for this instance.

### Additional Resources

[Feedback](#)

- Select the Java SSH Client option.
- Enter the path to the key pair file you downloaded, i.e. right-click on the file if you're not sure.

Filter: All instances ▾ All instance types ▾ SSHD installers

### Connect To Your Instance

I would like to connect with

A standalone SSH client

A Java SSH Client directly from my browser (Java required)

Enter the required information in the fields below to connect to your instance. AWS automatically detects the key pair name, and Public DNS for your instance. You need to enter the location and name of the .pem file containing your private key.

**Public DNS** ec2-54-234-227-244.compute-1.amazonaws.com

**User name**

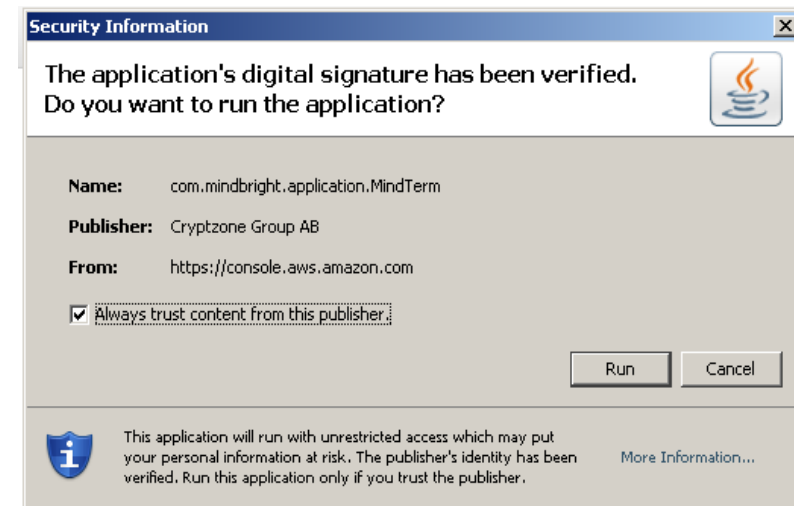
**Key name** HadoopTest.pem

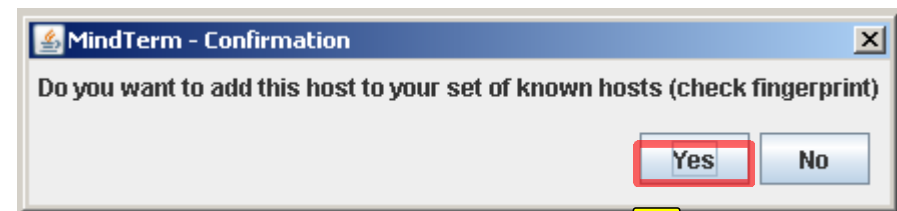
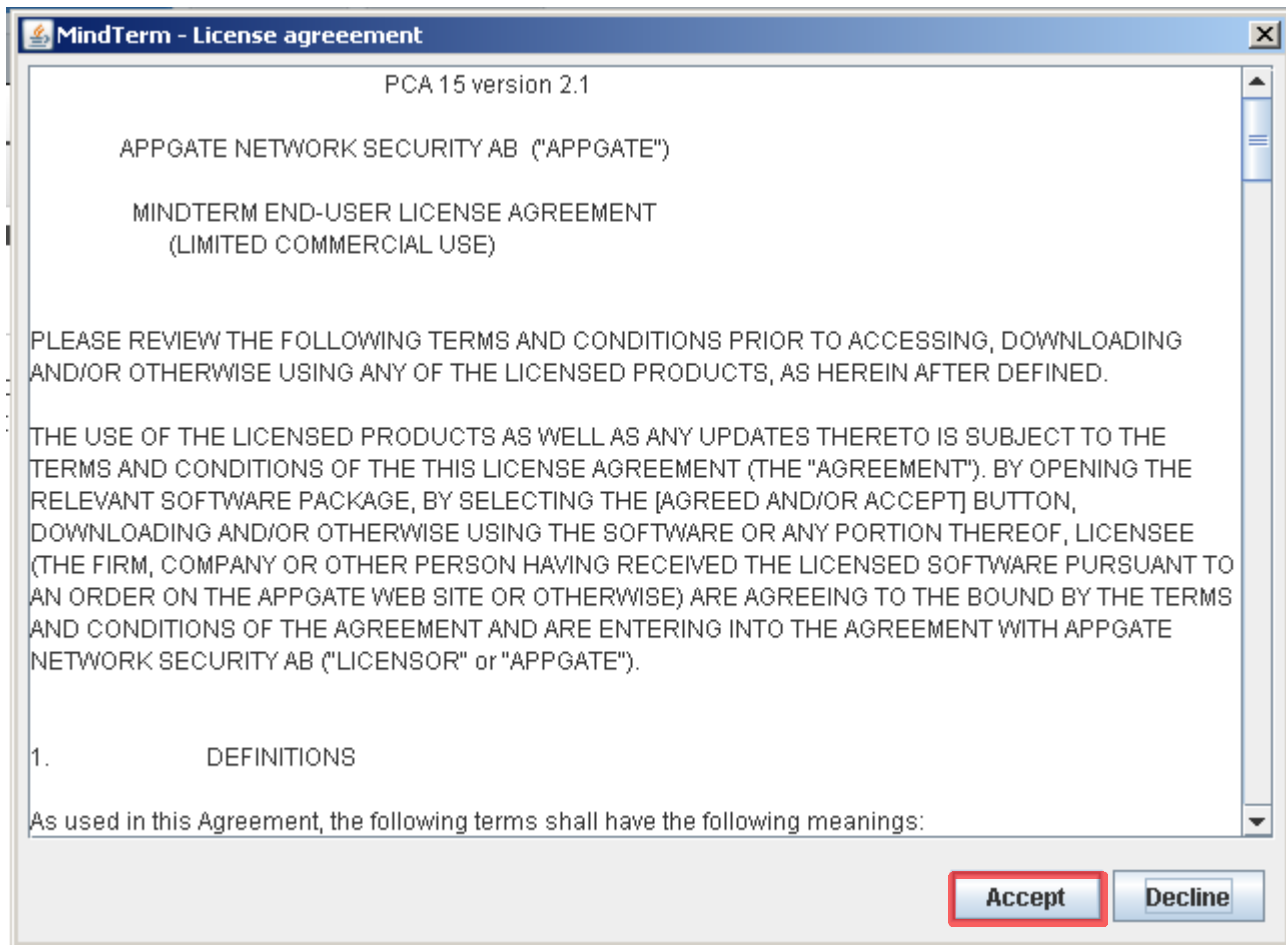
**Private key path**

**Save key location**  Store in browser cache

**Launch SSH Client**

**Close**

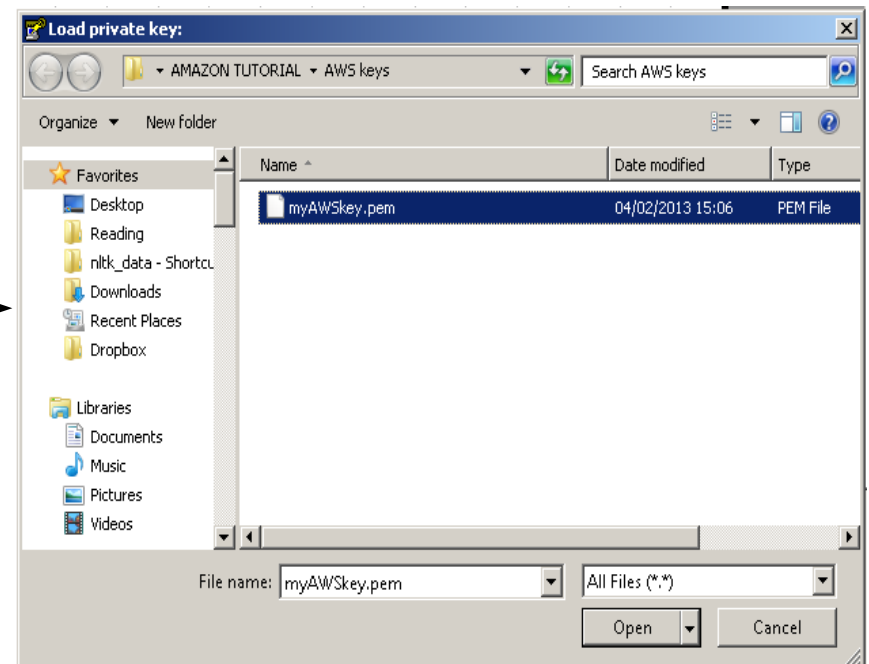
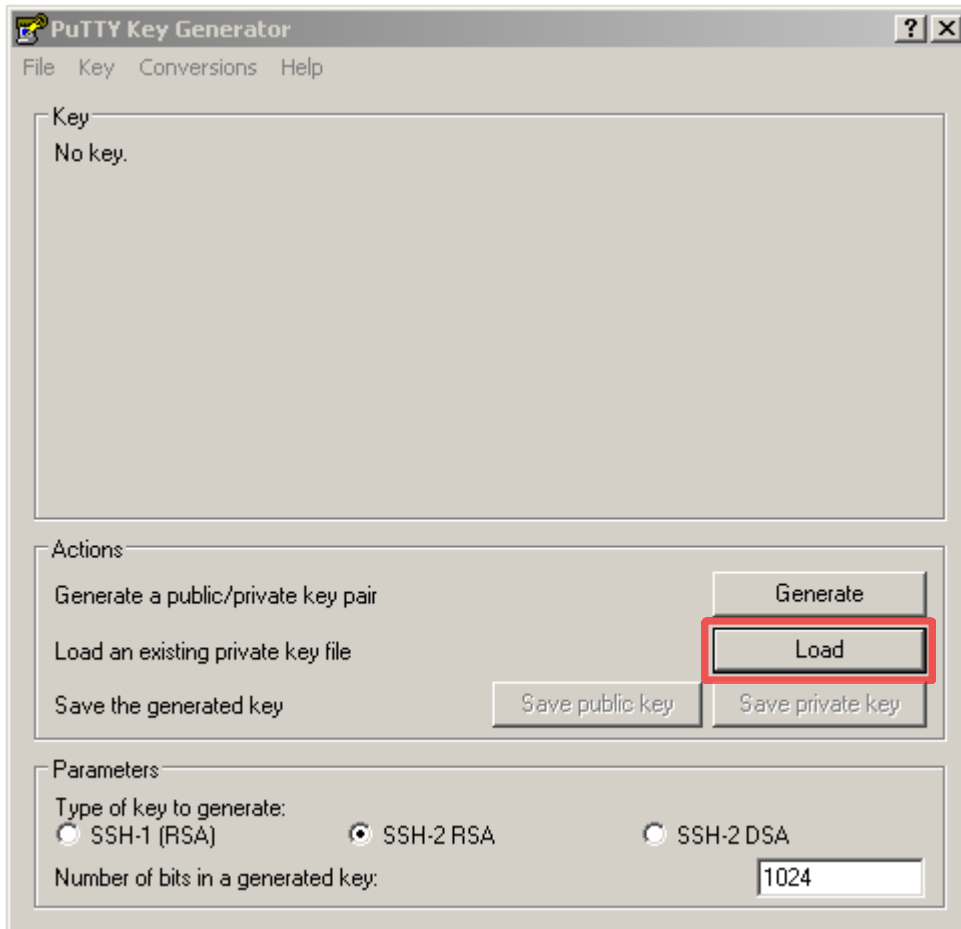




# Setting up PuTTY for AWS instance connection

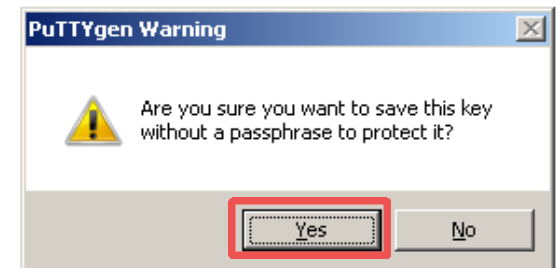
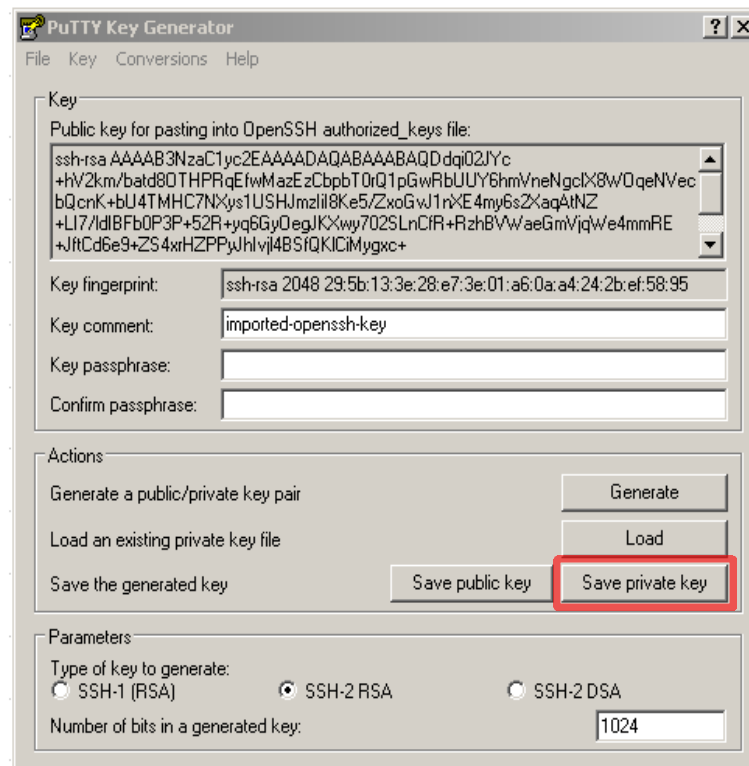


- Start PuTTYgen (Start menu, click All Programs > PuTTY > **PuTTYgen**).
- Click on **Load button**
- Find the folder with your **\*.pem** key in.
- Select **All Files \*.\*** and click on your AWS **.pem** key.



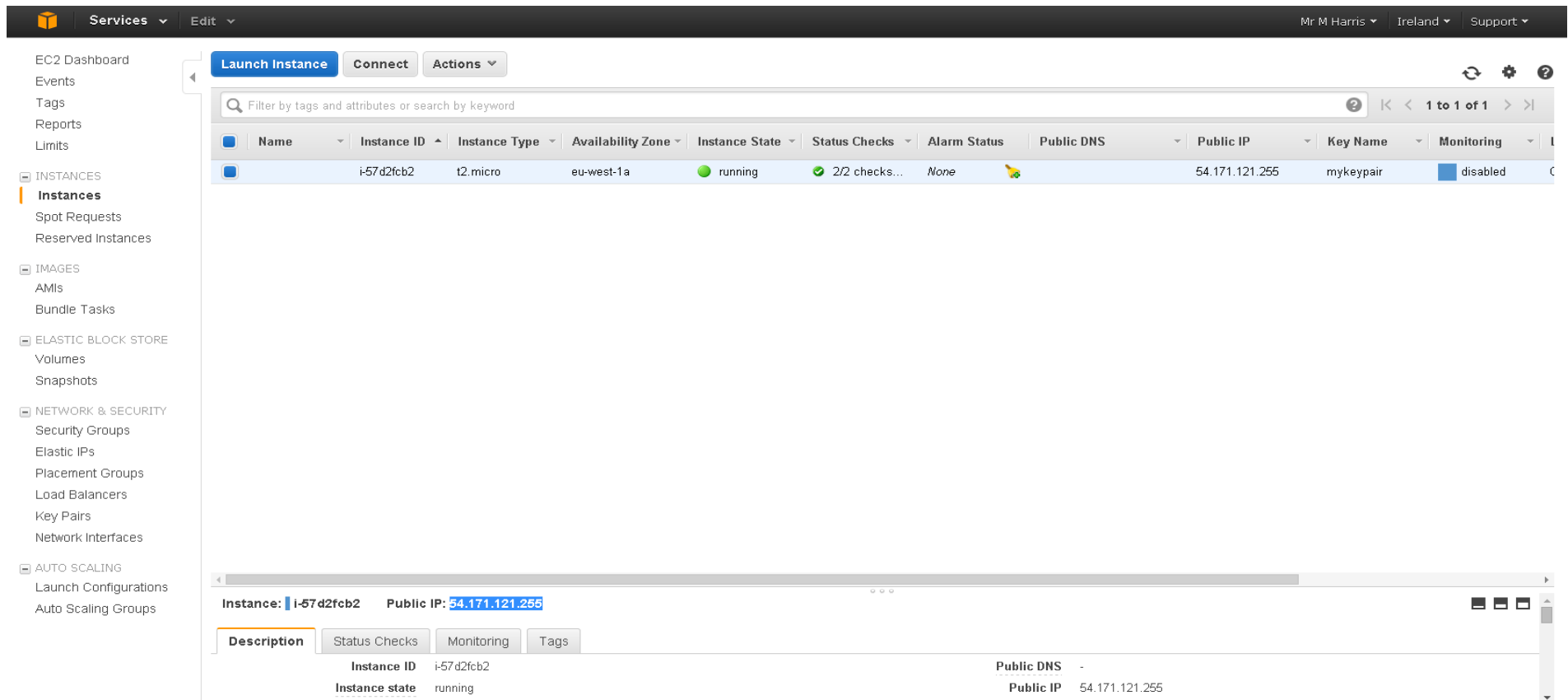


- A success message should appear, now we need to save the key in PUTTY's own format.
- Click on **Save private key**.
- Confirm you wish to save **without** a passphrase, and save in the same directory.



# Connecting to our instance using PuTTY SSH

- Go to Start > All Programs > PuTTY > PuTTY to load up **PUTTY SSH**.
- Switch back to the AWS console, and copy the address of your instance, it'll look something like **54.171.121.255**



The screenshot displays the AWS Management Console interface for the EC2 service. The top navigation bar shows 'Services' and 'Edit'. The left sidebar contains a navigation menu with categories like INSTANCES, IMAGES, ELASTIC BLOCK STORE, NETWORK & SECURITY, and AUTO SCALING. The main content area shows a table of EC2 instances. The table has columns for Name, Instance ID, Instance Type, Availability Zone, Instance State, Status Checks, Alarm Status, Public DNS, Public IP, Key Name, and Monitoring. A single instance is listed with Instance ID 'i-57d2fcb2', Instance Type 't2.micro', Availability Zone 'eu-west-1a', Instance State 'running', Status Checks '2/2 checks...', Alarm Status 'None', Public IP '54.171.121.255', and Key Name 'mykeypair'. Below the table, the instance details are shown, including the Instance ID 'i-57d2fcb2' and Public IP '54.171.121.255'. The 'Description' tab is selected, showing the Instance ID and Instance state 'running'.

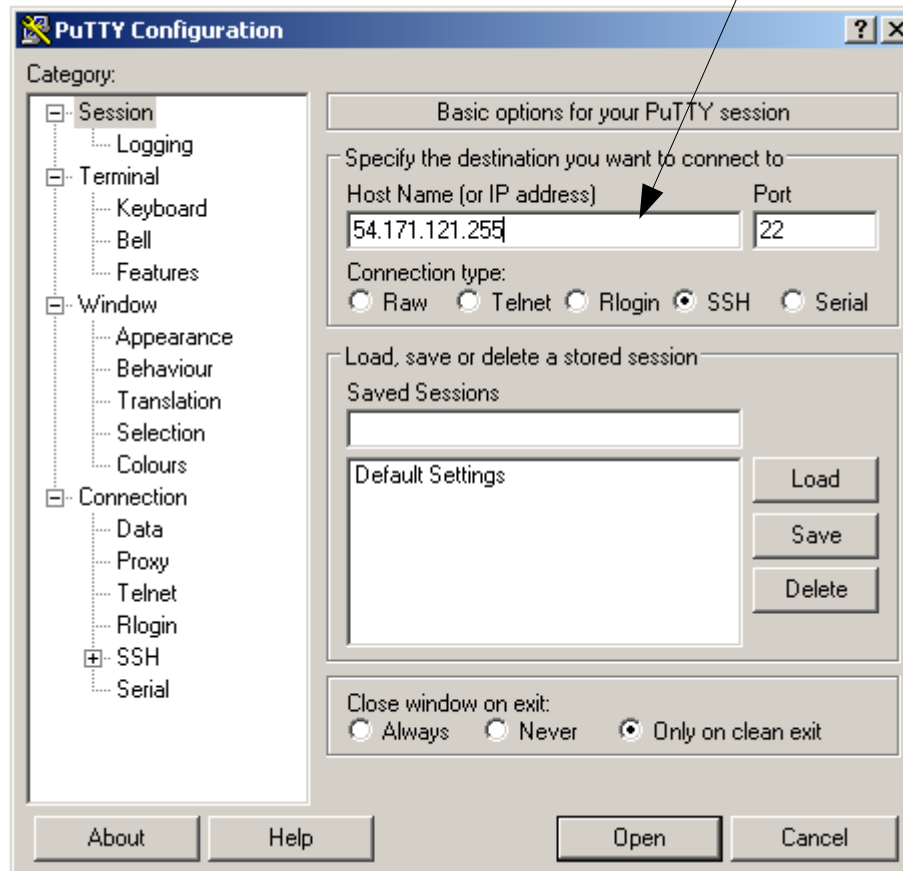
Name	Instance ID	Instance Type	Availability Zone	Instance State	Status Checks	Alarm Status	Public DNS	Public IP	Key Name	Monitoring
	i-57d2fcb2	t2.micro	eu-west-1a	running	2/2 checks...	None		54.171.121.255	mykeypair	disabled

Instance: i-57d2fcb2    Public IP: 54.171.121.255

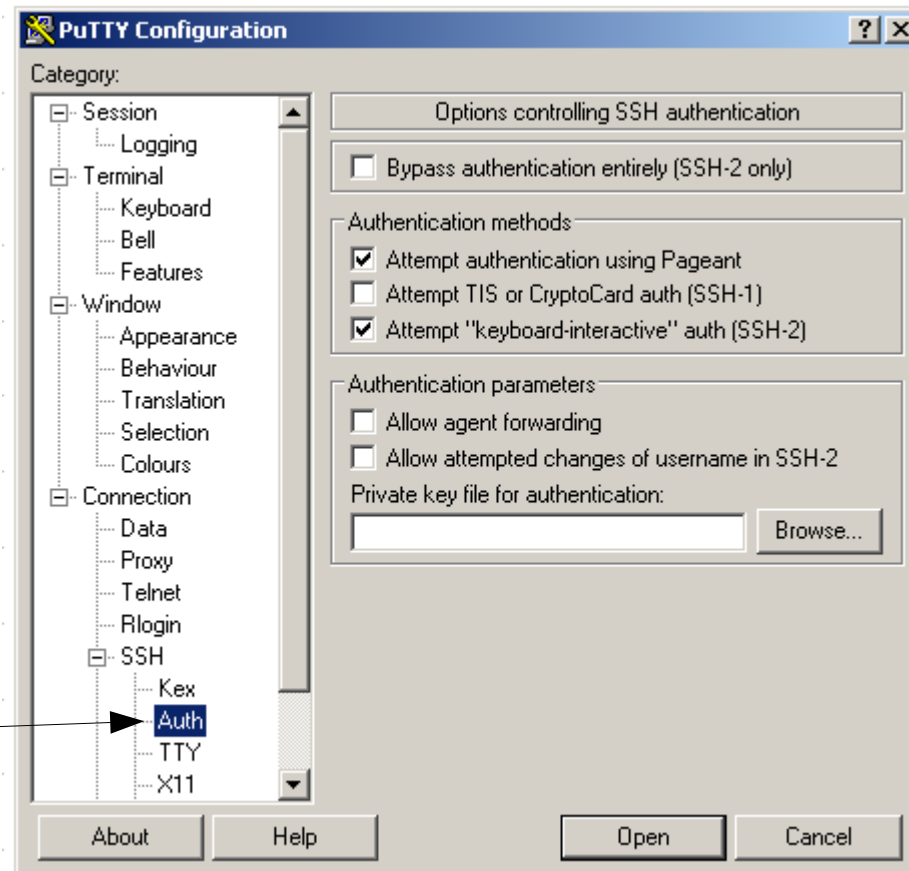
Description	Status Checks	Monitoring	Tags
Instance ID	i-57d2fcb2		
Instance state	running		
Public DNS	-		
Public IP	54.171.121.255		

- This is the address of the instance that we'll be using to connect to.

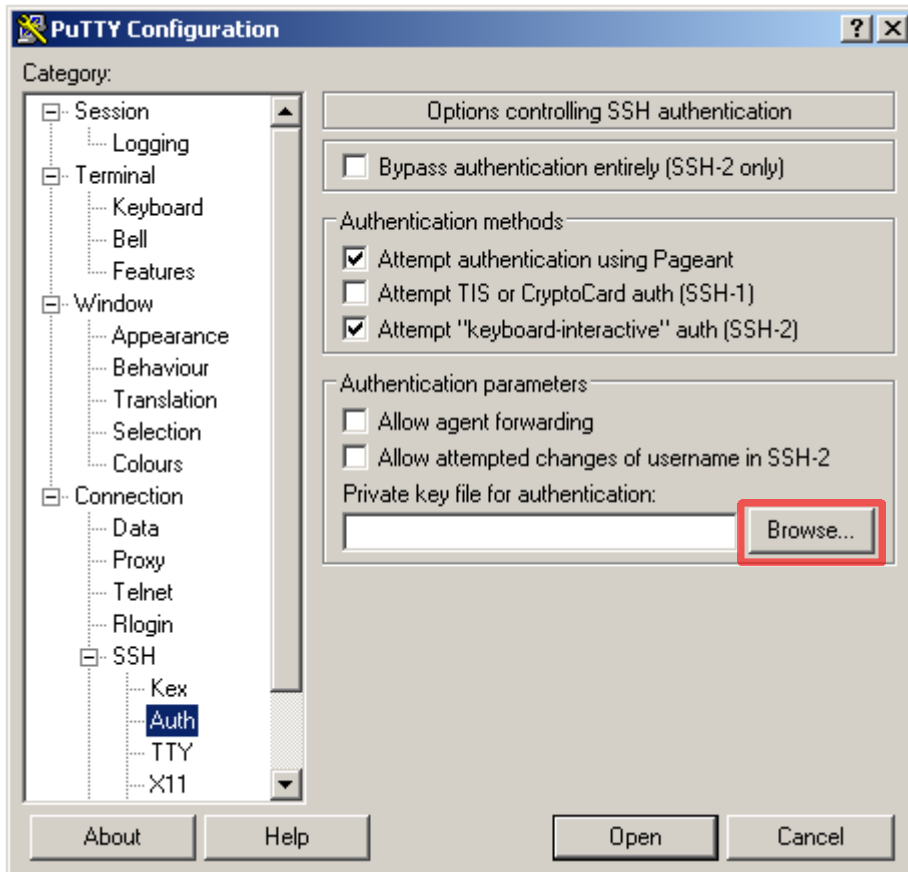
Paste the address here



Scroll down and click on **Auth**



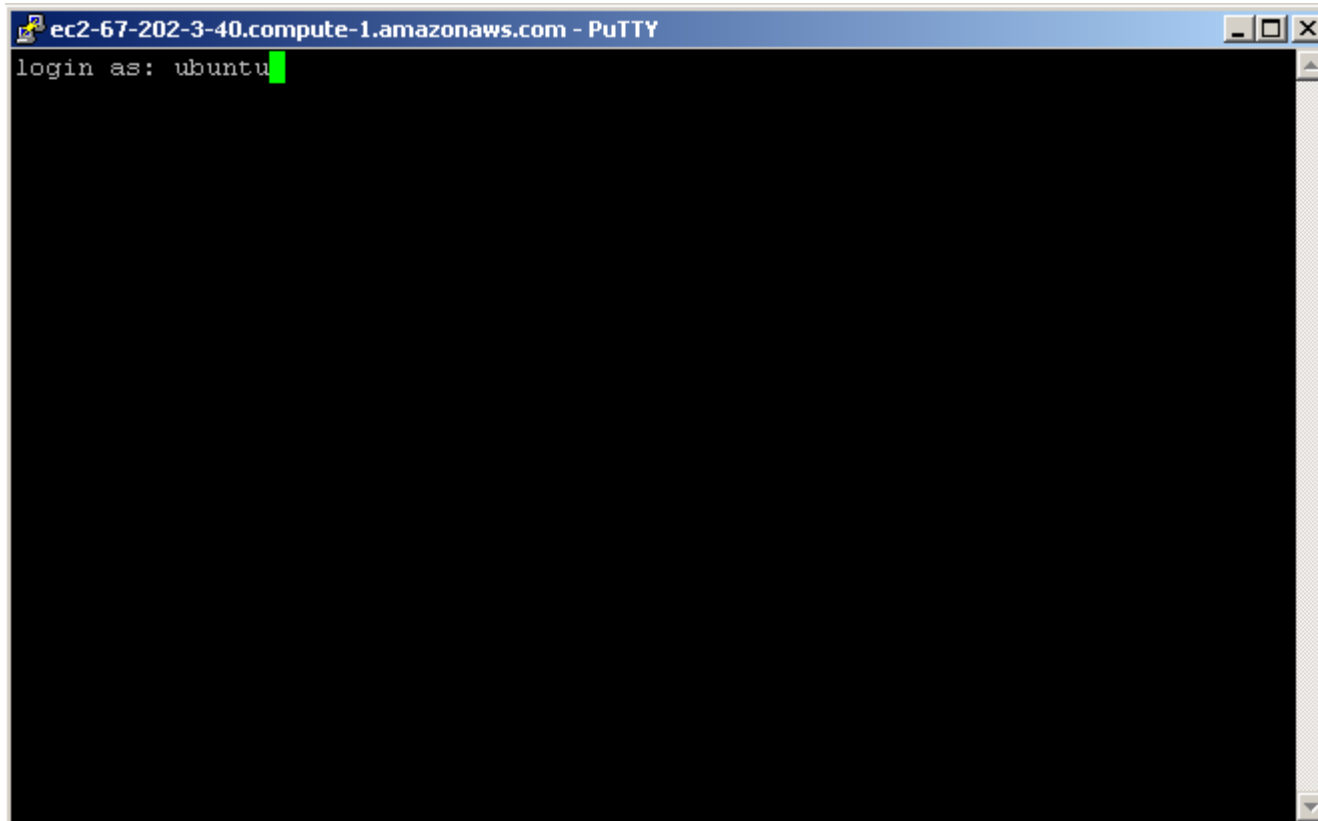
- Now click on **Browse** and navigate to the key you just saved (ends with **'.ppk'** extension).



- Now click on **Open**.
- Click on **yes** when the security alert appears.

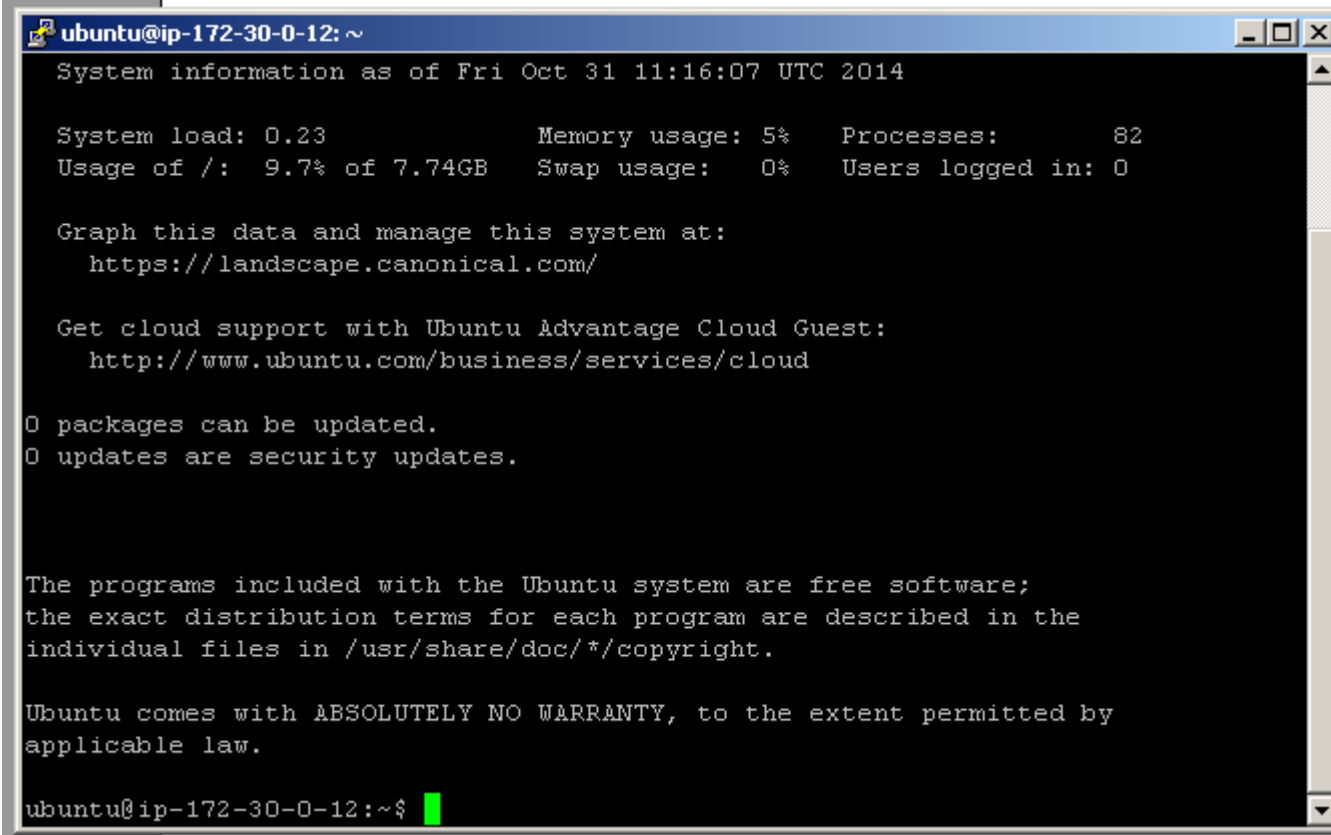


- Type **ubuntu** as the login name and press **Enter** key



- We don't need a password as our key will be sent across to the instance.

- Success! We're now logged in to our **Ubuntu** instance

A terminal window titled 'ubuntu@ip-172-30-0-12: ~' with standard window controls. The terminal displays system information for Ubuntu as of Friday, October 31, 2014, at 11:16:07 UTC. It shows system load, memory usage, processes, disk usage, swap usage, and users logged in. It also provides links for system management and cloud support, and mentions that 0 packages can be updated, all of which are security updates. A green cursor is visible at the bottom prompt.

```
ubuntu@ip-172-30-0-12: ~  
System information as of Fri Oct 31 11:16:07 UTC 2014  
  
System load: 0.23          Memory usage: 5%    Processes:      82  
Usage of /:  9.7% of 7.74GB Swap usage:   0%    Users logged in: 0  
  
Graph this data and manage this system at:  
  https://landscape.canonical.com/  
  
Get cloud support with Ubuntu Advantage Cloud Guest:  
  http://www.ubuntu.com/business/services/cloud  
  
0 packages can be updated.  
0 updates are security updates.  
  
The programs included with the Ubuntu system are free software;  
the exact distribution terms for each program are described in the  
individual files in /usr/share/doc/*/copyright.  
  
Ubuntu comes with ABSOLUTELY NO WARRANTY, to the extent permitted by  
applicable law.  
  
ubuntu@ip-172-30-0-12:~$ █
```

## Installing Java:

```
$ sudo apt-get update
```

```
$ sudo apt-get install openjdk-6-jre
```

## Installing Hadoop:

- **Get the file from external site:**

```
$ wget https://archive.apache.org/dist/hadoop/core/hadoop-0.22.0/hadoop-0.22.0.tar.gz
```

- **Unpack it:**

```
$ tar xzf hadoop-0.22.0.tar.gz
```

- **Copy it to somewhere more sensible like our local user directory.**

```
$ sudo cp -r hadoop-*/ /usr/local
```



There's a space here

**Note:** You can copy the below and press **SHIFT + Ins** to paste in to your terminal window.



- Did you get this error?

sudo: unable to resolve host ip-172-30-0-12

\$ sudo nano /etc/hosts

```
127.0.0.1 localhost
127.0.1.1 ip-172-30-0-12
```

The following lines are desirable for IPv6 capable hosts

```
::1 ip6-localhost ip6-loopback
fe00::0 ip6-localnet
ff00::0 ip6-mcastprefix
ff02::1 ip6-allnodes
ff02::2 ip6-allrouters
ff02::3 ip6-allhosts
#
```

- Save the file (**ctrl-x** then type **y** for yes).

- **Edit the terminal script**

```
$ nano ~/.bash
```

- **Add these lines at the bottom:**

```
export JAVA_HOME=usr/  
export HADOOP_HOME=usr/local/hadoop-0.22.0
```

- **Save the file (ctrl-x and type 'y')**

- **Add it to the terminal environment**

```
$ source ~/.bash
```

- **Now when Hadoop needs Java the terminal will point it in the right direction**



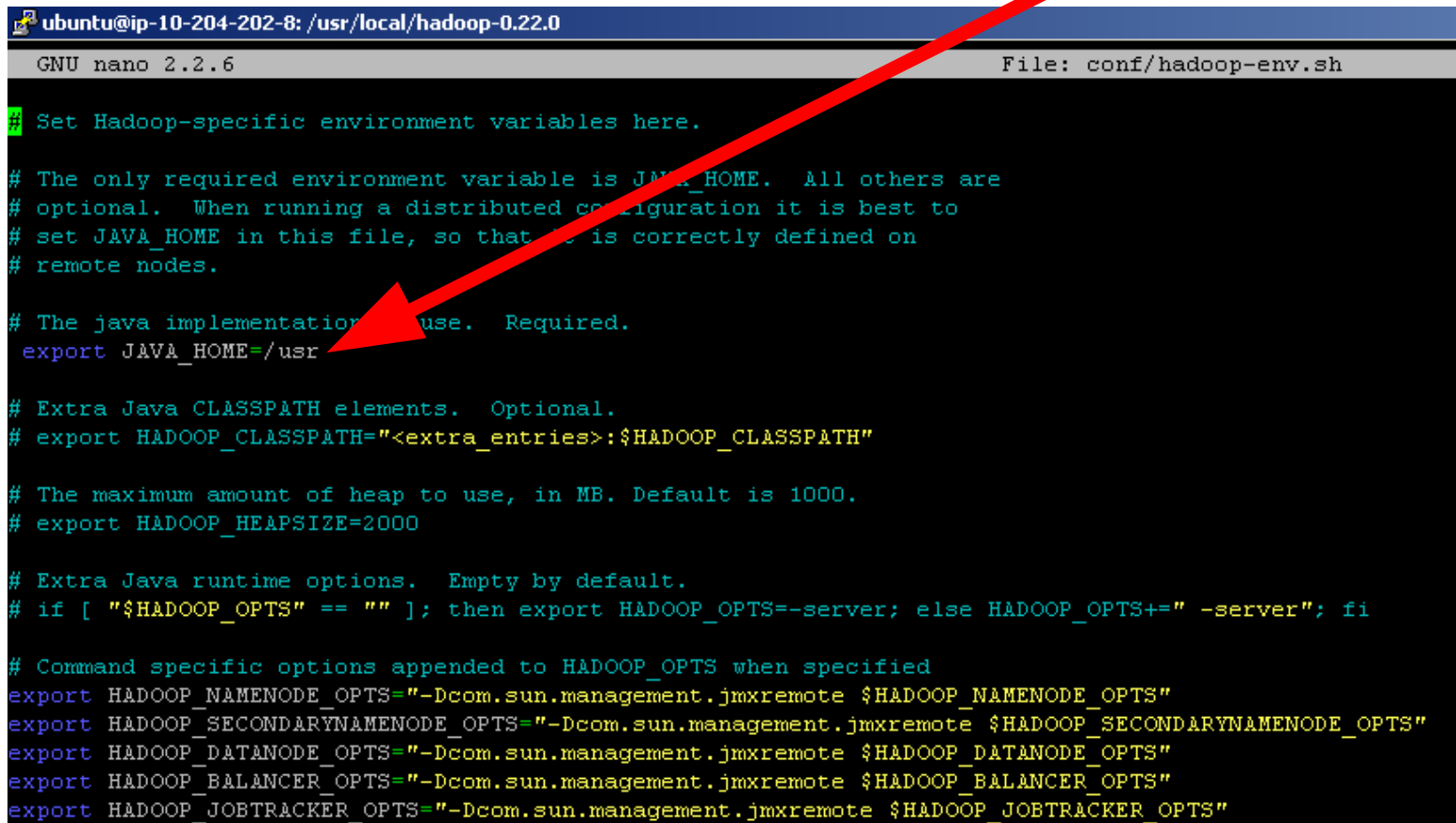
- Let's move in to the main directory of the application

```
$ cd /usr/local/hadoop-*
```

- Now edit Hadoop's set up script

```
$ sudo nano conf/hadoop-env.sh
```

```
export JAVA_HOME=/usr
```



```
ubuntu@ip-10-204-202-8: /usr/local/hadoop-0.22.0
GNU nano 2.2.6 File: conf/hadoop-env.sh
# Set Hadoop-specific environment variables here.

# The only required environment variable is JAVA_HOME. All others are
# optional. When running a distributed configuration it is best to
# set JAVA_HOME in this file, so that it is correctly defined on
# remote nodes.

# The java implementation to use. Required.
export JAVA_HOME=/usr

# Extra Java CLASSPATH elements. Optional.
# export HADOOP_CLASSPATH="<extra_entries>:$HADOOP_CLASSPATH"

# The maximum amount of heap to use, in MB. Default is 1000.
# export HADOOP_HEAPSIZE=2000

# Extra Java runtime options. Empty by default.
# if [ "$HADOOP_OPTS" == "" ]; then export HADOOP_OPTS=-server; else HADOOP_OPTS+=" -server"; fi

# Command specific options appended to HADOOP_OPTS when specified
export HADOOP_NAMENODE_OPTS="-Dcom.sun.management.jmxremote $HADOOP_NAMENODE_OPTS"
export HADOOP_SECONDARYNAMENODE_OPTS="-Dcom.sun.management.jmxremote $HADOOP_SECONDARYNAMENODE_OPTS"
export HADOOP_DATANODE_OPTS="-Dcom.sun.management.jmxremote $HADOOP_DATANODE_OPTS"
export HADOOP_BALANCER_OPTS="-Dcom.sun.management.jmxremote $HADOOP_BALANCER_OPTS"
export HADOOP_JOBTRACKER_OPTS="-Dcom.sun.management.jmxremote $HADOOP_JOBTRACKER_OPTS"
```

- Save (ctrl-x, then type 'y')

- **Add the configuration file to the terminals scope:**

```
$ source conf/hadoop-env.sh
```

- **Running an example using Single node mode:**

- **Calculating PI:**

```
$ sudo bin/hadoop jar hadoop-mapred-examples-*.jar pi 10 10000000
```

## Another example, using some actual data

- Create a directory to put our data in

```
$ sudo mkdir input
```

- Copy the very interesting README.txt file to our new input folder

```
$ sudo cp README.txt LICENSE.txt input
```


- Now we count up the total words and what they are  
(Hadoop will create the output folder for us)

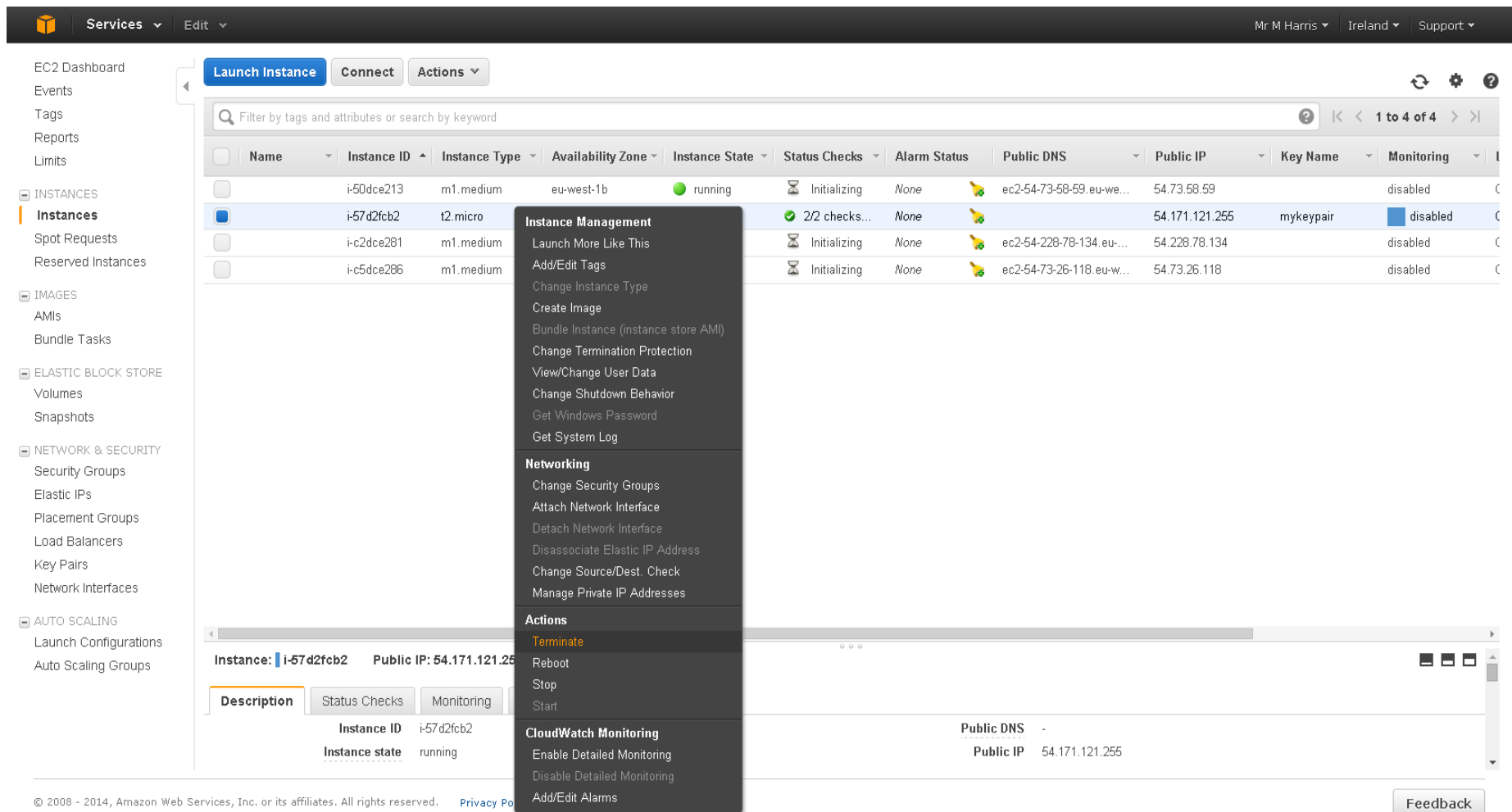
```
$ sudo bin/hadoop jar hadoop-mapred-examples-*.jar wordcount input output
```

- Have a look at the final output

```
$ nano output/part-r-00000
```

# Shutting down your instance

- Amazon charges by the hour, so make sure you close your instance after each session.
- Select the instance that is running through **EC2** option in the **AWS console**
- Right-click and select **Stop** to halt the instance, or **Terminate** to remove and delete everything. 



The screenshot displays the AWS Management Console interface for the EC2 service. The left-hand navigation pane shows various services, with 'INSTANCES' selected. The main content area shows a table of EC2 instances. The instance 'i-57d2fcb2' is highlighted, and a context menu is open over it, showing options such as 'Instance Management', 'Networking', 'Actions', and 'CloudWatch Monitoring'. The 'Stop' option is selected in the 'Actions' section. Below the table, the details for the selected instance are visible, including its ID, state, and public IP address.

Name	Instance ID	Instance Type	Availability Zone	Instance State	Status Checks	Alarm Status	Public DNS	Public IP	Key Name	Monitoring
	i-50dce213	m1.medium	eu-west-1b	running	Initializing	None	ec2-54-73-58-59.eu-we...	54.73.58.59		disabled
	i-57d2fcb2	t2.micro		running	2/2 checks...	None	ec2-54-171-121-255.eu-w...	54.171.121.255	mykeypair	disabled
	i-c2dce281	m1.medium		Initializing	Initializing	None	ec2-54-228-78-134.eu-...	54.228.78.134		disabled
	i-c5dce286	m1.medium		Initializing	Initializing	None	ec2-54-73-26-118.eu-w...	54.73.26.118		disabled

Instance: **i-57d2fcb2** Public IP: **54.171.121.255**

**Description** | Status Checks | Monitoring

Instance ID: **i-57d2fcb2**  
Instance state: **running**

Public DNS: **-**  
Public IP: **54.171.121.255**

© 2008 - 2014, Amazon Web Services, Inc. or its affiliates. All rights reserved. [Privacy Policy](#) [Feedback](#)

# Hadoop in the AWS Cloud



One last example, this time using AWS to create the Hadoop cluster for us.

First we need a place to put the data after it has been produced...

**Amazon S3** (Simple Storage Service):

An online storage web service providing storage through web services interfaces (REST, SOAP, and BitTorrent)

# Setting up the storage

- Select **S3** from the console

The screenshot shows the AWS Management Console interface. At the top, there is a navigation bar with the AWS logo, 'Services' dropdown, and 'Edit' dropdown. Below this, the console is divided into several sections:

- Welcome:** A introductory message about the AWS Management Console.
- Getting started guides:** Links for 'Getting started guides', 'Reference architectures', and 'Free Usage Tier'.
- Set Start Page:** A dropdown menu currently set to 'Console Home'.
- Amazon Web Services:** A grid of service categories:
  - Compute & Networking:** Direct Connect, EC2, Elastic MapReduce, Route 53, VPC.
  - Storage & Content Delivery:** CloudFront, Glacier, **S3** (highlighted with a mouse cursor), Storage Gateway.
  - Database:** DynamoDB, ElastiCache, RDS.
  - Deployment & Management:** CloudFormation, CloudWatch, Data Pipeline, Elastic Beanstalk, IAM.
  - App Services:** CloudSearch, Elastic Transcoder, SES, SNS, SQS, SWF.
- AWS Marketplace:** A promotional box for finding and buying software.





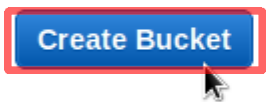
# Welcome to Amazon Simple Storage Service

Amazon S3 is storage for the Internet. It is designed to make web-scale computing easier for developers.

Amazon S3 provides a simple web services interface that can be used to store and retrieve any amount of data, at any time, from anywhere on the web. It gives any developer access to the same highly scalable, reliable, secure, fast, inexpensive infrastructure that Amazon uses to run its own global network of web sites. The service aims to maximize benefits of scale and to pass those benefits on to developers.

You can read, write, and delete objects ranging in size from 1 byte to 5 terabytes each. The number of objects you can store is unlimited. Each object is stored in a bucket with a unique key that you assign.

Get started by simply creating a bucket and uploading a test object, for example a photo or .txt file.



## Additional Information

[Getting Started Guide](#)

[Documentation](#)

[All S3 Resources](#)

[Forums](#)

## S3 at a glance

### Create



Create a bucket in one of several Regions. You can choose a Region to optimize for latency, minimize costs, or address regulatory environments.

### Add



Upload objects to your bucket. Amazon S3 durably stores your data in multiple facilities and on multiple devices within each facility.

### Manage



Manage your data with Amazon S3's lifecycle management capabilities, including the ability to automatically archive objects to even lower cost storage options.

**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
- Oregon
- Northern California
- Ireland**
- Singapore
- Tokyo
- Sydney
- Sao Paulo

[Set Up Logging >](#) **Create** [Cancel](#)

Give it a name

(not **MyBucket** – something unique, also NO CAPITAL LETTERS)


Choose **Ireland** from the region list

(it's closer, so less latency)

Create Bucket Actions ▾

**Buckets**

	Name
--	------

 lazyeels

← Your new bucket

# Running a MapReduce program in AWS

- Select **Elastic MapReduce** in AWS console

The screenshot shows the AWS Management Console interface. At the top, there is a navigation bar with the AWS logo, 'Services' dropdown, and 'Edit' dropdown. Below this, the 'Welcome' section provides introductory text and links for getting started guides, reference architectures, and the free usage tier. The 'Set Start Page' section includes a 'Console Home' dropdown. The main 'Amazon Web Services' section is organized into categories: Compute & Networking, Storage & Content Delivery, Database, Deployment & Management, and App Services. In the 'Compute & Networking' category, 'Elastic MapReduce' is highlighted with a mouse cursor. The 'Elastic MapReduce' service is described as 'Managed Hadoop Framework'. Other services listed include Direct Connect, EC2, Route 53, VPC, CloudFormation, CloudWatch, Data Pipeline, Elastic Beanstalk, IAM, CloudSearch, Elastic Transcoder, SES, SNS, SQS, and SWF.

**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 Marketplace**  
Find & buy software, launch with 1-Click and pay by the hour.

**Amazon Web Services**

**Compute & Networking**

- Direct Connect**  
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**  
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
- RDS**  
Managed Relational Database Service

**Deployment & Management**

- CloudFormation**  
Templated AWS Resource Creation
- CloudWatch**  
Resource & Application Monitoring
- Data Pipeline** **NEW**  
Orchestration for data-driven workflows
- Elastic Beanstalk**  
AWS Application Container
- IAM**  
Secure AWS Access Control

**App Services**

- CloudSearch**  
Managed Search Service
- Elastic Transcoder** **NEW**  
Easy-to-use scalable media transcoding
- SES**  
Email Sending Service
- SNS**  
Push Notification Service
- SQS**  
Message Queue Service
- SWF**  
Workflow Service for Coordinating Application Components

# Select Create Cluster



## 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.



## How Elastic MapReduce Works

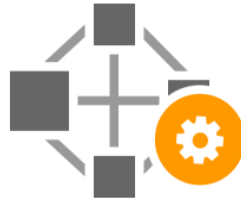
### Upload



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](#)

## Additional Information

More about Elastic MapReduce

- [EMR overview](#)
- [FAQs](#)
- [Pricing](#)

More Help Using Elastic MapReduce

- [Forum](#)
- [Documentation](#)
- [Developer Guide](#)
- [Quick Reference Card](#)
- [API Reference](#)
- [EMR on GitHub](#)
- [Help portal](#)

- Select **Configure sample application**.
- Choose the **Word count** example from the drop down menu.
- Click on the **Output location** folder and select your new **bucket**.

Cluster Configuration

Cluster name: My cluster

Termination protection:  Yes  No

Logging:  Enabled

Log folder S3 location: s3://

Debugging:  Enabled

Tags

Optional: Add up to 10 tags to your EMR cluster. A tag consists of a case-sensitive key-value pair. Tags on EMR clusters are propagated to the underlying EC2 instances. [Learn more](#) about tagging your Amazon EMR clusters.

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

Software Configuration

Hadoop distribution:  Amazon  MapR

AMI version: 3.2.1

Applications to be installed	Version
Hive	0.13.1

Change to your bucket name.

Configure Sample Application

Select a sample application to auto-populate the Create Cluster page

Select sample application: Word count

Output location: s3://<bucket-name>/wordcount/output/2014-10-31/11-5

Logging:  Enabled

Debugging:  Enabled

Cancel OK

s3://<your bucket-name>/logging/

- Click **OK** when done.

Next, specify how many instances you want – just leave it at two for now (the more instances the more £££ it will be to run your job).

### Create a New Job Flow Cancel

---

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:   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:

Instance Type:   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:

Instance Type:   Request Spot Instances

---

[Back](#)   \* Required field

Select your keypair 

## Create a New Job Flow Cancel

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

Here you enter advanced details about your job flow, such as an EC2 key pair, to use VPC, and your job flow debugging options.

Amazon 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:

To run this job flow in a Virtual Private Cloud (VPC), select a subnet. See [Create a VPC](#).

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

Amazon S3 Log Path:

Optional: To copy log files from the job flow to Amazon S3, specify an Amazon S3 bucket.

Enable Debugging:  Yes  No

Yes means EMR will store an index of your logs (requires an Amazon S3 Log Path).

Set advanced job flow options.

Keep Alive  Yes  No

Yes means the job flow will keep running after processing is complete.

Termination Protection  Yes  No

Yes prevents your nodes from shutting down due to accident or error.

Visible To All IAM Users  Yes  No

Yes means the job flow will be visible to all IAM users under your account.

[Back](#)

\* Required field



- Scroll to the bottom of the page.

## Bootstrap Actions

**i** Bootstrap actions are scripts that are executed during setup before Hadoop starts on every cluster node. You can use them to install additional software and customize your applications. [Learn more](#)

Bootstrap action type	Name	S3 location	Optional arguments
Add bootstrap action			
	Select a bootstrap action		
<a href="#">Configure and add</a>			

## Steps

**i** A step is a unit of work you submit to the cluster. A step might contain one or more Hadoop jobs, or contain instructions to install or configure an application. You can submit up to 256 steps to a cluster. [Learn more](#)

Name	Action on failure	JAR location	Arguments
Word count	Terminate cluster	/home/hadoop/contrib/streaming/hadoop-streaming.jar	-files s3://eu-west-1.elasticmapreduce/samples/wordcount/wordSplitter.py -mapper wordSplitter.py -reducer aggregate -input s3://eu-west-1.elasticmapreduce/samples/wordcount/input -output s3://lazyeels/

Add step Select a step

[Configure and add](#)

**Auto-terminate**  Yes Automatically terminate cluster after the last step is completed.

No Keep cluster running until you terminate it.

**i** No EC2 key pair has been selected, SSH access will not be enabled for this cluster. [Learn how to create an EC2 Key Pair.](#)

Cancel

Create cluster

# Setting up your own job (for coursework)

This is the place to configure your Hadoop job by uploading your code and data to your **S3** bucket.

**Add Step** [X]

**Step type** Streaming program

**Name\*** Word count

**Mapper\*** s3://eu-west-1.elasticmapreduce/samples/wordcount/w S3 location of the map function or the name of the Hadoop streaming command to run.

**Reducer\*** aggregate S3 location of the reduce function or the name of the Hadoop streaming command to run.

**Input S3 location\*** s3://eu-west-1.elasticmapreduce/samples/wordcount/in  
*s3://<bucket-name>/<folder>/*

**Output S3 location\*** s3://lazyeels/output/  
*s3://<bucket-name>/<folder>/*

**Arguments**

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

Cancel **Save**

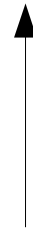
**Input data:**

eu-west-1.elasticmapreduce/samples/wordcount/input

**Output data:**

This is going to be stored on our **S3** bucket...

s3n://**lazyeels**/wordcount/output/2013-11-01



Todays date

## • Click on **Create cluster**.

### Bootstrap Actions

**i** Bootstrap actions are scripts that are executed during setup before Hadoop starts on every cluster node. You can use them to install additional software and customize your applications. [Learn more](#)

Bootstrap action type	Name	S3 location	Optional arguments		
-----------------------	------	-------------	--------------------	--	--

Add bootstrap action

Configure and add

### Steps

**i** A step is a unit of work you submit to the cluster. A step might contain one or more Hadoop jobs, or contain instructions to install or configure an application. You can submit up to 256 steps to a cluster. [Learn more](#)

Name	Action on failure	JAR location	Arguments		
------	-------------------	--------------	-----------	--	--

Word count

Terminate cluster

/home/hadoop/contrib/streaming/hadoop-streaming.jar

-files s3://eu-west-1.elasticmapreduce/samples/wordcount/wordSplitter.py -mapper wordSplitter.py -reducer aggregate -input s3://eu-west-1.elasticmapreduce/samples/wordcount/input -output s3://lazeels/



Add step

Configure and add

Auto-terminate  Yes

Automatically terminate cluster after the last step is completed.

No

Keep cluster running until you terminate it.

**i** No EC2 key pair has been selected, SSH access will not be enabled for this cluster. [Learn how to create an EC2 Key Pair.](#)

Cancel

Create cluster

- Your MapReduce job is now running.

Services | Edit | Mr M Harris | Ireland | Support

Elastic MapReduce | Cluster List > Cluster Details | EMR Help

Add step | Resize | Clone | Terminate

Cluster: Word count **Starting** Provisioning Amazon EC2 capacity

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

Summary	Configuration Details	Security/Network	Hardware
ID: j-2M1YZHITBO8KV Creation date: 2014-10-31 12:27 (UTC+0) Elapsed time: 54 seconds Auto-terminate: Yes Termination protection: <a href="#">Change</a>	AMI version: 3.2.1 Hadoop distribution: Amazon 2.4.0 Applications: -- Log URI: s3://lazyeels/logging/ EMRFS consistent view: Disabled	Availability zone: eu-west-1b Subnet ID: -- Key name: -- EC2 instance profile: -- EMR role: -- Visible to all users: All <a href="#">Change</a>	Master: Provisioning 1 m1.medium Core: Provisioning 2 m1.medium Task: --

Monitoring

Steps

Add step | Clone step

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


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

	ID	Name	Status	Start time (UTC+0)	Elapsed time	Log files	Actions
<input type="radio"/>	s-2POCER36NP36G	Setup hadoop debugging	Pending			<a href="#">View logs</a>	<a href="#">View jobs</a>
<input type="radio"/>	s-3LYN6QUZ1J98U	Word count	Pending			<a href="#">View logs</a>	<a href="#">View jobs</a>

Bootstrap Actions

- Go to your **S3** bucket via the **AWS** console.
- The results have been written to the output folder in parts in HDFS format

	Name	Storage Class	Size	Last Modified
<input type="checkbox"/>	_SUCCESS	Standard	0 bytes	Fri Oct 31 12:35:48 GMT+000 2014
<input type="checkbox"/>	part-00000	Standard	97.3 KB	Fri Oct 31 12:35:35 GMT+000 2014
<input type="checkbox"/>	part-00001	Standard	98.6 KB	Fri Oct 31 12:35:47 GMT+000 2014
<input type="checkbox"/>	part-00002	Standard	97.1 KB	Fri Oct 31 12:35:48 GMT+000 2014



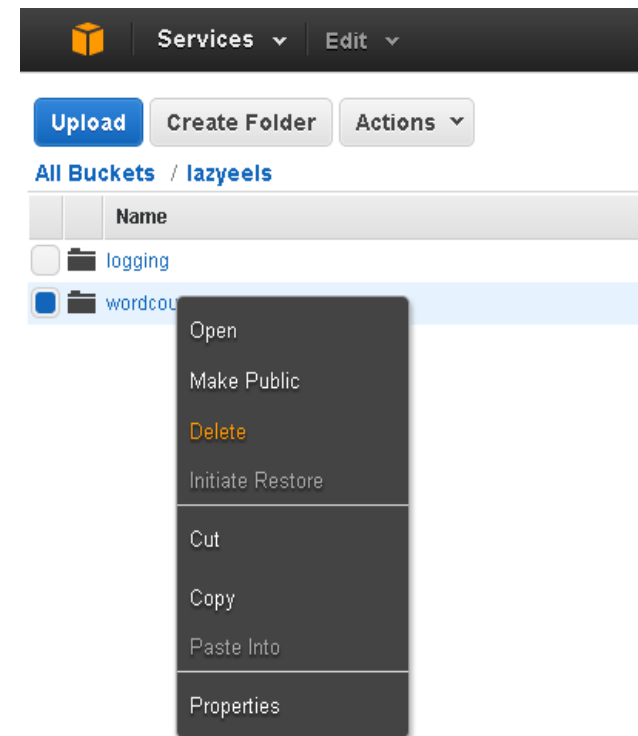
You can delete the results by right-clicking on the folder and selecting **delete**.

Amazon charges for storage so this is worth doing if you no longer need it.

In addition, Hadoop will fail if it finds a folder with the same name when it writes the output.

**Note:** The S3 bucket is where you would upload your **.jar** or **.py** files representing your code, as well as any data. It is worth creating a separate folder for each of your runs.

Click on the upload button to upload them from your local machine.



## Some tips:

**Hadoop** is not designed to run on Windows. Consider using **Cygwin** or **Virtualbox** (<https://www.virtualbox.org>), or installing **Linux Mint** (<http://www.linuxmint.com/>) alongside your Windows install (at home).

Stick to earlier versions of Hadoop such as **0.22.0** (they keep moving things around, especially the class files that you'll need to compile your code to **.jar**)

Most books and tutorials are based on earlier versions of Hadoop.

**Single-node** mode is fine for testing your map-reduce code before deploying it.

There are example programs in the folder at:

**Hadoop-0.22.0/mapreduce/src/examples/org/apache/hadoop/examples/**



**Get in the habit of ~~stopping~~ your instances when you're finished!**

**Hadoop in Action** is your friend! Consider getting a copy:

### **Chapter 2**

Shows you how to set everything up from scratch.

### **Chapter 3**

Provides some good templates to base your code on.

### **Chapter 4**

Discusses issues you may encounter with the different API versions

### **Chapter 9**

Tells you how to launch your MapReduce programs from the command line and AWS console, as well as using S3 buckets for data storage and how to access it.

## Some useful links

### Installing and usage:

<http://www.higherpass.com/linux/Tutorials/Installing-And-Using-Hadoop/>

### Running a job using the AWS Jobflow (Elastic Map Reduce):

<http://cloud.dzone.com/articles/how-run-elastic-mapreduce-job>

### Theory:

<http://developer.yahoo.com/hadoop/tutorial/module1.html>

<http://www.cs.washington.edu/education/courses/cse490h/08au/readings/communications200801-dl.pdf> (Page 108)

### Accessing AWS and Hadoop through the terminal (for Linux users):

<http://rodrigodsousa.blogspot.co.uk/2012/03/hadoop-amazon-ec2-updated-tutorial.html>