

Introduction to Elastic MapReduce

Csci E63 Big Data Analytics

Zoran B. Djordjević

Getting Started with Hadoop

- Hadoop is not all that difficult to start working with.
- For example, you can download Hadoop from hadoop.apache.org
- To install locally, unzip and set JAVA_HOME
- Details: hadoop.apache.org/core/docs/current/quickstart.html
- Several ways to write jobs:
 - Java API
 - Hadoop Streaming (for Python, Perl, R, Ruby, etc.)
 - Pipes API (C++)
- If you want to do very sophisticated work and create special map/reduce procedures you have few options. You have learn one of Hadoop's API-s

Elastic MapReduce

- Amazon Elastic MapReduce is a web service that utilizes a hosted Hadoop framework running on the web-scale infrastructure of Amazon Elastic Compute Cloud (Amazon EC2) and Amazon Simple Storage Service (Amazon S3).
- Using Amazon Elastic MapReduce, you can instantly provision as much or as little capacity as you like to perform data-intensive tasks for applications such as web indexing, data mining, log file analysis, machine learning, financial analysis, scientific simulation, and bioinformatics research.

Benefits of Elastic MapReduce

- Amazon Elastic MapReduce lets you focus on crunching or analyzing your data without having to worry about time-consuming set-up, management or tuning of Hadoop clusters or the hardware capacity upon which they sit.
- Amazon Elastic MapReduce automatically sub-divides the data in a job flow into smaller chunks so that data can be processed (the “map” function) in parallel, and eventually recombining the processed data into the final solution (the “reduce” function).
- Amazon S3 serves as the source for the data being analyzed, and as the output destination for the end results.

Elastic MapReduce Functionality

- Develop your data processing application.
- Amazon Elastic MapReduce enables job flows to be developed in SQL-like languages, such as Hive and Pig.
- If desired, more sophisticated applications can be run in: Java, Ruby, Perl, Python, PHP, R, or C++.
- Upload your data and your processing application into Amazon S3.
- Log in to the AWS Management Console to start an Amazon Elastic MapReduce “job flow.” Alternatively you can start a job flow by specifying the same information mentioned above via our Command Line Tools or APIs.
- Monitor the progress of your job flow(s) directly from the AWS Management Console, Command Line Tools or APIs.

Service Highlights

- Amazon Elastic MapReduce enables you to use as many or as few compute instances running Hadoop as you want. You can commission one, hundreds, or even thousands of instances.
- You don't need to worry about setting up, running, or tuning the performance of Hadoop clusters.
- Amazon Elastic MapReduce is built on Amazon's highly reliable infrastructure, and has tuned Hadoop's performance specifically for Amazon's infrastructure environment.
- Amazon Elastic MapReduce is designed to integrate easily with other AWS services such as Amazon S3 and EC2.
- Secure and inexpensive.

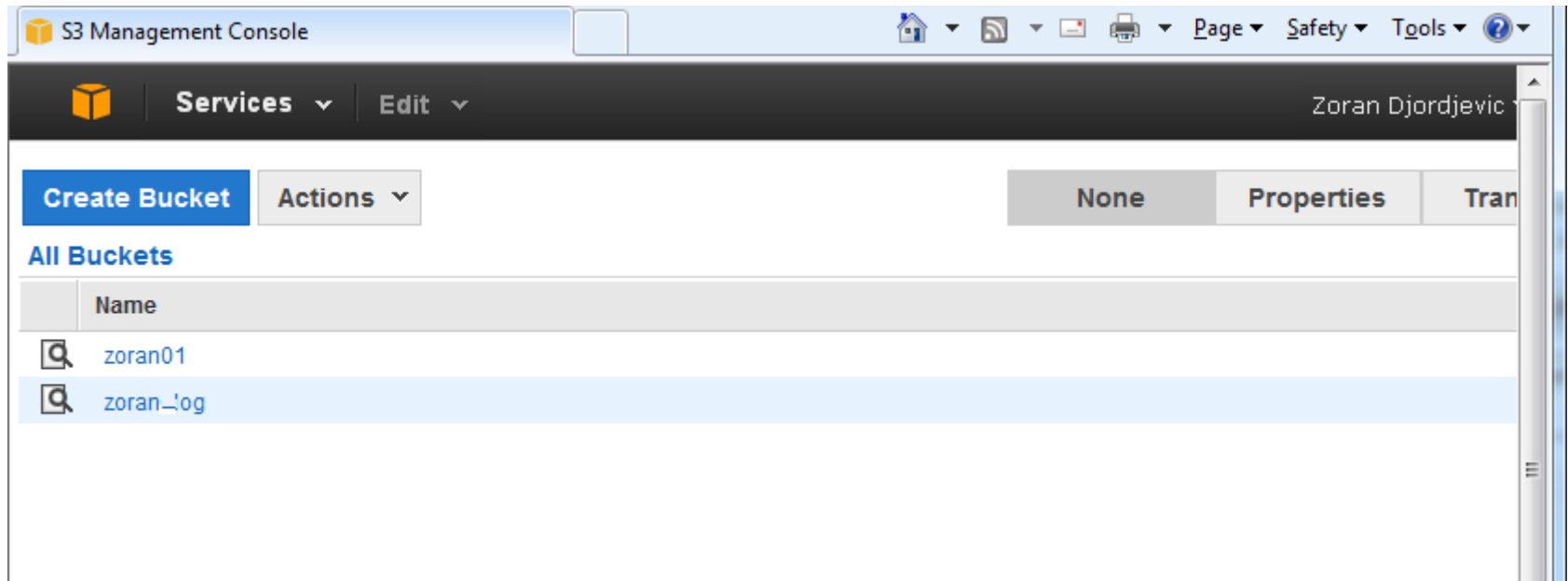
Pricing is coming down

Region: <input type="text" value="US East (N. Virginia)"/>		
Standard On-Demand Instances	Amazon EC2 Price	Amazon Elastic MapReduce Price
Small (Default)	\$0.065 per hour	\$0.015 per hour
Large	\$0.26 per hour	\$0.06 per hour
Extra Large	\$0.52 per hour	\$0.12 per hour
Hi-Memory On-Demand Instances		
Extra Large	\$0.45 per hour	\$0.09 per hour
Double Extra Large	\$0.90 per hour	\$0.21 per hour
Quadruple Extra Large	\$1.80 per hour	\$0.42 per hour
Hi-CPU On-Demand Instances		
Medium	\$0.165 per hour	\$0.03 per hour
Extra Large	\$0.66 per hour	\$0.12 per hour
Cluster Compute On-Demand Instances		
Quadruple Extra Large	\$1.30 per hour	\$0.27 per hour
Cluster Compute Eight Extra Large	\$2.40 per hour	\$0.50 per hour
Cluster GPU On-Demand Instances		
Quadruple Extra Large	\$2.10 per hour	\$0.42 per hour

- EMR prices are paid atop of instance prices. These prices are approximate.
- Amazon EC2, Amazon S3 and Amazon SimpleDB charges are billed separately.

Before you start, Create two S3 Buckets

- We will need one bucket for normal results of our Map Reduce programs and one bucket for log data
- bucket name must contain only lowercase letters, numbers, periods (.), and dashes (-)




AWS Services, Select Elastic Map Reduce

The screenshot shows the AWS Management Console interface. The top navigation bar includes the AWS logo, a search bar, and a user profile dropdown for 'Zoran Djordjevic'. The main content area is divided into several sections:

- Welcome:** A message stating '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.' It includes links for 'Getting started guides', 'Reference architectures', and 'Free Usage Tier'.
- Set Start Page:** A dropdown menu currently set to 'Console Home'.
- AWS Marketplace:** A section for finding and buying software, with the text 'Find & buy software, launch with 1-Click and pay by the hour.'
- Amazon Web Services:** A grid of service categories and individual services:
 - Compute & Networking:** Direct Connect, EC2, Elastic MapReduce (highlighted), Route 53, VPC.
 - Storage & Content Delivery:** CloudFront, Glacier, S3, Storage Gateway.
 - Database:** DynamoDB.
 - Deployment & Management:** CloudFormation, CloudWatch, Data Pipeline, Elastic Beanstalk, IAM, OpsWorks.
 - App Services:** CloudSearch, Elastic Transcoder, SES, SNS, SQS.
- Announcements:** A section for news, including 'Announcing New AWS CloudFormation Deployment Enhancements', 'Announcing AWS OpsWorks', and 'Amazon Redshift Now Available to All Customers'.
- Service Health:** A section for monitoring service status, with a link to 'Service Health Dashboard'.

@Zoran B. Djordjevic

Select Create cluster

 **Services** ▾ **Edit** ▾

Welcome to Amazon Elastic MapReduce

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

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

[Create cluster](#)

How Elastic MapReduce Works

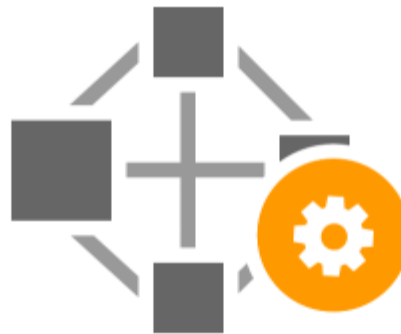
Upload



Upload your data and processing application to S3.

[Learn more](#)

Create



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

[Learn more](#)

Monitor



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

[Learn more](#)

Cluster Configuration, Hadoop version

Elastic MapReduce ▾ Create Cluster

Cluster Configuration

[Configure sample application](#)

Cluster name

Termination protection ☒ Yes

☐ No

Prevents accidental termination of the cluster: to shut down the cluster, you must turn off termination protection. [Learn more](#)

Logging ☒ Enabled

Copy the cluster's log files automatically to S3. [Learn more](#)

Log folder S3 location

s3://<bucket-name>/<folder>/

Tags

Debugging ☒ Enabled

Index logs to enable console debugging functionality (requires logging). [Learn more](#)

i 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

Use Amazon's Hadoop distribution. [Learn more](#)

AMI version

Determines the base configuration of the instances in your cluster, including the Hadoop version. [Learn more](#)

☐ MapR

Use MapR's Hadoop distribution. [Learn more](#)

Applications to be installed	Version			
Hive	0.11.0.1			
Pig	0.11.1.1			

Configure Hardware, Add EC2 key pair

Hardware Configuration

i Specify the [networking](#) and [hardware](#) configuration for your cluster. If you need more than 20 EC2 instances, [complete this form](#).
[Request Spot instances](#) (unused EC2 capacity) to save money.

Network	<div>Launch into EC2-Classic ▼</div> <div>i To create a cluster in a VPC, you must first create a VPC. For more information, click here.</div>	Use a Virtual Private Cloud (VPC) to process sensitive data or connect to a private network. Create a VPC																
EC2 availability zone	<div>No preference ▼</div>	Launch the cluster in a specific EC2 Availability Zone.																
	<table><thead><tr><th></th><th>EC2 instance type</th><th>Count</th><th>Request spot</th></tr></thead><tbody><tr><td>Master</td><td><div>m1.small ▼</div></td><td><div>1</div></td><td><input type="checkbox"/></td></tr><tr><td>Core</td><td><div>m1.small ▼</div></td><td><div>2</div></td><td><input type="checkbox"/></td></tr><tr><td>Task</td><td><div>m1.small ▼</div></td><td><div>0</div></td><td><input type="checkbox"/></td></tr></tbody></table>		EC2 instance type	Count	Request spot	Master	<div>m1.small ▼</div>	<div>1</div>	<input type="checkbox"/>	Core	<div>m1.small ▼</div>	<div>2</div>	<input type="checkbox"/>	Task	<div>m1.small ▼</div>	<div>0</div>	<input type="checkbox"/>	<p>The Master instance assigns Hadoop tasks to core and task nodes, and monitors their status.</p> <p>Core instances run Hadoop tasks and store data using the Hadoop Distributed File System (HDFS).</p> <p>Task instances run Hadoop tasks.</p>
	EC2 instance type	Count	Request spot															
Master	<div>m1.small ▼</div>	<div>1</div>	<input type="checkbox"/>															
Core	<div>m1.small ▼</div>	<div>2</div>	<input type="checkbox"/>															
Task	<div>m1.small ▼</div>	<div>0</div>	<input type="checkbox"/>															

Security and Access

EC2 key pair	<div>e63 ▼</div>	Use an existing key pair to SSH into the master node of the Amazon EC2 cluster as the user "hadoop". Learn more
IAM user access	<div><input type="radio"/> All other IAM users</div> <div><input checked="" type="radio"/> No other IAM users</div>	Control the visibility of this cluster to other IAM users. Learn more

Bootstrap Actions	IAM role	<div>Proceed without role ▼</div>	Control permissions for applications on the cluster. Learn more
--------------------------	-----------------	-----------------------------------	---

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	<div>Select a bootstrap action ▼</div>				

@Zoran B. Djordjevic

Setup Bootstrap Actions

- Bootstrap actions allow you to pass a reference to a script stored in Amazon S3. This script can contain configuration settings and arguments related to Hadoop or Elastic MapReduce.
- Bootstrap actions are run before Hadoop starts and before the node begins processing data. Actions are like:
 - Install software on the node,
 - Modify the default Hadoop site configuration,
 - Change the way Java parameters use Hadoop daemons
- You can specify up to 16 bootstrap actions per job flow by providing multiple *--bootstrap-action* parameters from the CLI or API.

Steps, Select Sample App, a Pig program

- We will use a demo application to illustrate processing.
- Select a Pig program from Add step dropdown

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 S3 location	Arguments		
------	-------------------	-----------------	-----------	--	--

Add step

Configure and add

Auto-terminate ☐ Yes

☒ No

Automatically terminate cluster after the last step is completed.

Keep cluster running until you terminate it.

Cancel

Create cluster

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

Feedback

- Hit Configure and add button

Select script, S3 bucket for, input, output

The screenshot shows the 'Add Step' dialog box with the following configuration:

- Step type:** Pig program
- Name:** Pig program
- Script S3 location*:** s3://elasticmapreduce/samples/pig-apache/do-reports2
s3://<bucket-name>/<path-to-file>
- Input S3 location:** s3://elasticmapreduce/samples/pig-apache/input
s3://<bucket-name>/<folder>/
- Output S3 location:** s3://zoran01/pig-apache
s3://<bucket-name>/<folder>/
- Arguments:** (Empty text box)
- Action on failure:** Terminate cluster

Buttons at the bottom: Cancel, Add

- The script we are running is called `do-reports2.pig` and is written in a special scripting language written specially for Hadoop.
- We will look at that language in fine detail during one of subsequent classes.
- Once the step is configured hit `Create cluster` button on the main screen

Cluster is Starting

Elastic MapReduce ▾

Cluster List > Cluster Details

EMR Help

Add step

Resize

Clone

Terminate

↻

Cluster: MyCluster01

Starting

Master public DNS: --

Tags: name = pig_example [View All / Edit](#)

Summary	Configuration Details	Security/Network
<div>ID: j-EPR2LBMSORDJ</div> <div>Creation date: 2014-02-21 11:52 (UTC-5)</div> <div>Elapsed time: 10 seconds</div> <div>Auto- No</div> <div>terminate: terminate</div> <div>Termination On Change</div> <div>protection: protection</div>	<div>AMI version: 2.4.2</div> <div>Hadoop Amazon 1.0.3</div> <div>distribution: distribution</div> <div>Applications: Hive 0.11.0.1, Pig 0.11.1.1</div> <div>Log URI: s3n://zoran-log/pig-apache/ </div>	<div>Availability --</div> <div>zone: zone</div> <div>Subnet ID: --</div> <div>Key name: e63</div> <div>IAM role: --</div> <div>Visible to all None</div> <div>users: users</div>

Hardware

Master: Provisioning 1 m1.small

Core: Provisioning 2 m1.small

Task: --

► Monitoring

► Steps

► Bootstrap Actions

Cluster is Running

- After a while cluster will change its state to Running and eventually to Waiting.

Elastic MapReduce ▾Cluster List > Cluster DetailsEMR Help

Add stepResizeCloneTerminate

Cluster: MyCluster01RunningRunning step

Master public DNS:ec2-50-17-121-86.compute-1.amazonaws.com

Tags: name = pig_exampleView All / Edit

Summary

ID: j-EPR2LBMSORDJ

Creation date: 2014-02-21 11:52 (UTC-5)

Elapsed time: 4 minutes

Auto-terminate: No

Termination protection: OnChange

Hardware

Master: Running 1 m1.small

Core: Running 2 m1.small

Task: --

Monitoring

Steps

Bootstrap Actions

Configuration Details

AMI version: 2.4.2

Hadoop distribution: Amazon 1.0.3

Applications: Hive 0.11.0.1, Pig 0.11.1.1

Log URI: s3n://zoran-log/pig-apache/

Security/Network

Availability zone: us-east-1d

Subnet ID: --

Key name: e63

IAM role: --

Visible to all users: None

Elastic MapReduce ▾Cluster List

Create clusterView detailsCloneTerminate

Filter: All clusters ▾Filter clusters ...2 clusters (all loaded)

	Name	ID	Status
<input type="checkbox"/>	MyCluster01	j-EPR2LBMSORDJ	Waiting

There are 3 EC2 Instance, a Master and 2 Slaves

Launch Instance Connect Actions

Filter: Running instances All instance types Search Instances 1 to 3 of 3 Instances

Name	Instance ID	Instance Type	Availability Zone	Instance State	Status Checks	Alarm Status	Public DNS
<input checked="" type="checkbox"/>	i-e486ccc5	m1.small	us-east-1d	running	2/2 checks pas...	None	ec2-50-17-121-86
<input type="checkbox"/>	i-4781cb66	m1.small	us-east-1d	running	2/2 checks pas...	None	ec2-54-226-74-94
<input type="checkbox"/>	i-5181cb70	m1.small	us-east-1d	running	2/2 checks pas...	None	ec2-54-81-34-143

Instance: i-e486ccc5 Public DNS: ec2-50-17-121-86.compute-1.amazonaws.com

Description Status Checks Monitoring Tags

Instance ID	i-e486ccc5	Public DNS	ec2-50-17-121-86.compute-1.amazonaws.com
Instance state	running	Public IP	50.17.121.86
Instance type	m1.small	Elastic IP	-
Private DNS	ip-10-167-7-161.ec2.internal	Availability zone	us-east-1d
Private IPs	10.167.7.161	Security groups	ElasticMapReduce-master. view rules
Secondary private IPs	-	Scheduled events	No scheduled events
VPC ID	-	AMI ID	debian-amd64.manifest.xml (ami-9f2e78f6)
Subnet ID	-	Platform	-
Network interfaces	-	IAM role	-
Source/dest. check	False	Key pair name	e63
		Owner	951414139794
EBS-optimized	False	Launch time	2014-02-21T16:52:31.000Z (less than one hour)

Job is Waiting, Steps are Completed

Elastic MapReduce ▾ [Cluster List](#) > Cluster Details EM

[Add step](#) [Resize](#) [Clone](#) [Terminate](#)

Cluster: MyCluster01 **Waiting** Waiting after step completed C

Master public DNS: ec2-50-17-121-86.compute-1.amazonaws.com
Tags: name = pig_example [View All](#) / [Edit](#)

Summary	Configuration Details	Security/Network	Hardware
ID: j-EPR2LBMSORDJ Creation date: 2014-02-21 11:52 (UTC-5) Elapsed time: 46 minutes Auto-terminate: No Termination protection: On Change	AMI version: 2.4.2 Hadoop distribution: Amazon 1.0.3 Applications: Hive 0.11.0.1, Pig 0.11.1.1 Log URI: s3n://zoran-log/pig-apache/	Availability zone: us-east-1d Subnet ID: -- Key name: e63 IAM role: -- Visible to all users: None	Master: Running 1 m1.small Core: Running 2 m1.small Task: --

► Monitoring

▼ Steps C

[Add step](#)

Steps View all interactive jobs | View all jobs


Filter: All steps ▾ 4 steps (all loaded)

	ID	Name	Status	Start time (UTC-5) ▾	Elapsed time	Log files	Actions
►	s-3L2Y2QOA2LTH5	Pig program	Completed	2014-02-21 11:59	14 minutes	View logs	View jobs
►	s-3JZ9STK2USE6Q	Setup pig	Completed	2014-02-21 11:58	1 minute	View logs	View jobs
►	s-2AZSWWJRY8S1Y	Setup hive	Completed	2014-02-21 11:56	1 minute	View logs	View jobs
►	s-9TB08QCNP5N	Setup hadoop debugging	Completed	2014-02-21 11:56	36 seconds	View logs	View jobs

Results in S3 Buckets





Upload Create Folder Actions ▾

All Buckets / zoran-log

Name
 pig-apache



Upload Create Folder Actions ▾

All Buckets / zoran01 / pig-apache

Name	Storage Class	Size
 top_50_external_referrers	--	--
 top_50_ips	--	--
 top_50_search_terms_from_bing_google	--	--
 total_requests_bytes_per_hour	--	--

Upload Create Folder Actions ▾

All Buckets / zoran01 / pig-apache / top_50_search_terms_from_bing_google

Name
 _SUCCESS
 part-r-00000

Top 50 search terms from file `part-r-00000`
in folder `top_50_search_terms_from_bing_google`

```
C:\Users\zdjrdje\Downloads\part-r-00000 - Notepad++
File Edit Search View Encoding Language Se
credential-file.bt x listener.ora x init.ora.332013
1 value 625
2 views 426
3 login 224
4 search 195
5 items 112
6 bigtable 68
7 google+bigtable 59
8 %23%21%2Fusr%2Fbin%2Fperl+-w
9 philmont+pictures 45
10 %23%21%2Fusr%2Fbin%2Fperl 44
11 philmont 37
12 google+quick+links 33
13 pvc+instrument 33
14 pig 30
15 walla 28
16 vegas 27
17 about+me+website 26
18 pig+the+pc+nerd 26
19 homemade 24
20 fishing 23
21 comments 23
22 travis 21
23 hadoop+0.20 21
24 google+big+table 21
25 miscellaneous 19
26 seattle 19
27 m0n0wall+ipv6 19
28 pebble 19
```

We can control Job Flow thru EMR Command Line

- Download and install Ruby 1.8.7

- http://rubyforge.org/frs/?group_id=167&release_id=28426
- Select `rubyinstaller-1.8.7-p398-rc2.exe`, perhaps.
- On Linux, do: `$ sudo apt-get install ruby`

- Download `elastic-mapreduce-client.zip` from

- <http://aws.amazon.com/developertools/2264>

- Unzip into `c:\AWS\elastic-mapreduce-ruby`

- Add `C:\AWS\elastic-mapreduce-ruby;C:\Ruby187\bin;` to your PATH.

- In the directory `c:\AWS\elastic-mapreduce-ruby` create file `credentials.json` and add:

```
{  
    "access_id":    "<insert your aws Access Key Id here>",  
    "private_key": "<insert your aws Secret Access Key here>",  
    "keypair":      "<insert path of your amazon ec2 Key Pair file>",  
    "log_uri":      "s3://name of a bucket in s3 to place logs from job"  
}
```

credentials.json

- Be careful with the content of this file. It must be right.

```
{  
  "access_id": "AKGGGGHJT7WWWWFDTHJQ",  
  "private_key": "gUlaTrEwIrQBsyqh3w6253422cK+FlUeRtBWE",  
  "keypair": "e63",  
  "key-pair-file": "C:\\AWS\\hu\\e63.pem",  
  "log_uri": "s3n://zoran-log/",  
  "region": "us-east-1"  
}
```

- When running EMR commands from the command prompt, you might have to go to the directory where this file resides.

Examples of Command Line Usage

Listing Active Job Flows (MapReduce jobs)

```
ruby elastic-mapreduce --list
ruby elastic-mapreduce --list --active
ruby elastic-mapreduce --list --all
# create a job flow that requires manual termination
ruby elastic-mapreduce --create --alive
```

To create a job flow that will run a mapper written in python, all one line

```
ruby elastic-mapreduce --create --stream --mapper \
s3://elasticmapreduce/samples/wordcount/wordSplitter.py \
--input s3n://elasticmapreduce/samples/wordcount/input
--output s3://zoran01/python --reducer aggregate
Created job flow j-QUHVJI5FUZFE
```

Bucket needs to be there but the output folder should not exist before the command is run. If the folder is present, you will get an error.

To terminate all active job flows

```
ruby elastic-mapreduce --list --active --terminate
# terminate a running job flow
ruby elastic-mapreduce --terminate --jobflow j-2WSXRVDHH08T1
```

Hadoop Streaming

- What we just did relied on a special feature/utility called Hadoop Streaming.
- Rather than writing scripts for Hadoop in a special language or writing jobs in Java, which is Hadoop's native language, you can write your `Map` and `Reduce` routines in almost any language and use a utility called Hadoop Streaming to run them.
- We demonstrated Hadoop Streaming using a provided example in Python.
- That same example could be run from the AWS EMR Console

Add S3 Permissions to all Authenticated Users

- Right click on your bucket and grant permissions to Authenticated users, Everyone and **Log Delivery**.
- You are a bit more generous than necessary.

The screenshot shows the AWS IAM console interface for configuring permissions on a bucket. At the top, there are tabs for 'Services' and 'Edit'. Below these, there are buttons for 'Create Bucket' and 'Actions'. A 'Buckets' list on the left shows four buckets: 'zdjordje001', 'zdjordje002' (selected), 'zoran002', and 'zoran1116log'. The main panel displays the 'Permissions' section for the selected bucket. It shows three permission entries:

- Grantee:** 'zdjordj' with checkboxes for 'List', 'Upload/Delete', and 'View'. There is also an 'Edit Permissions' checkbox.
- Grantee:** 'Authenticated Users' with checkboxes for 'List' and 'Upload/Delete'. Below this entry are 'View Permissions' and 'Edit Permissions' links.
- Grantee:** 'Everyone' with checkboxes for 'List' and 'Upload/Delete'. Below this entry are 'View Permissions' and 'Edit Permissions' links.
- Grantee:** 'Log Delivery' with checkboxes for 'List' and 'Upload/Delete'. Below this entry are 'View Permissions' and 'Edit Permissions' links.

At the bottom, there are three buttons: 'Add more permissions' (with a plus icon), 'Add bucket policy' (with a document icon), and 'Add CORS' (with a document icon).

Create new Cluster, Run Sample App, Word Count

- Next we go to the Elastic Map Reduce service and create a new cluster
- **Select** Configure sample application

Elastic MapReduce ▾

Create Cluster

EMR Help

Cluster Configuration

Configure sample application

Cluster name

My cluster

Termination protection

☒ Yes

☐ No

Prevents accidental termination of the cluster: to shut down the cluster, you must turn off termination protection. [Learn more](#)

Logging

☒ Enabled

Copy the cluster's log files automatically to S3. [Learn more](#)

Log folder S3 location

s3://zoran01/python/

s3://<bucket-name>/<folder>/

Debugging

☒ Enabled

Index logs to enable console debugging functionality (requires logging). [Learn more](#)

Tags

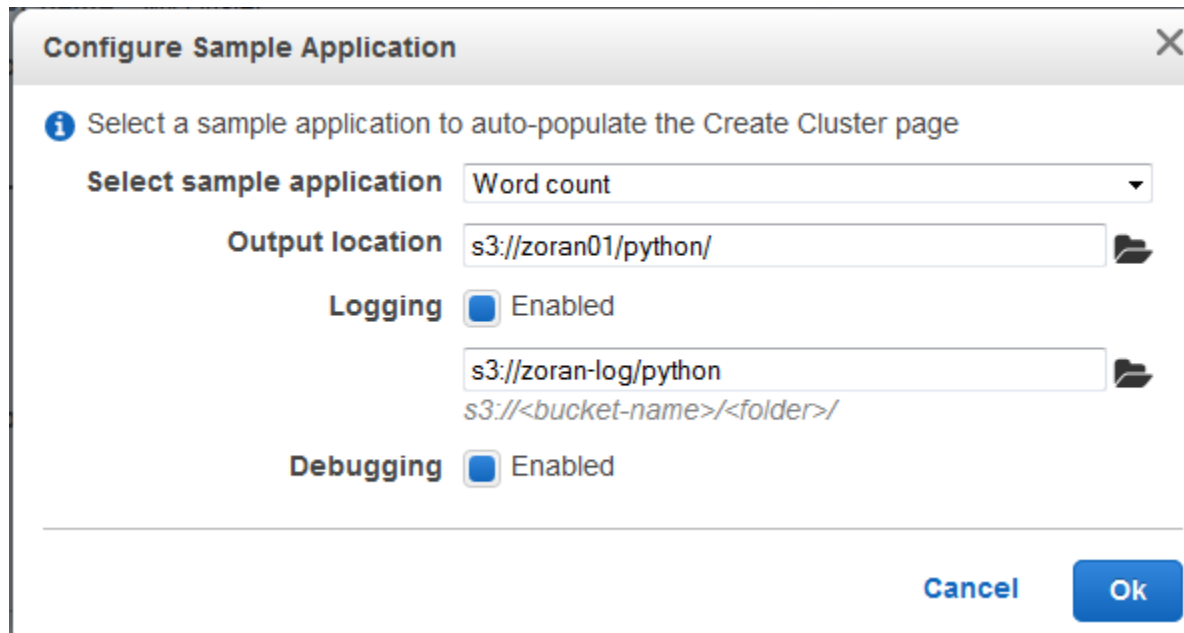
Optional: Add up to 10 tags to your EMR cluster. A tag consists of a case-sensitive key-value pair. Tags on EMR clusters are propagated to the underlying EC2 instances. [Learn more](#) about tagging your Amazon EMR clusters.

@Zoran B. Djordjevic

26

Configure sample app, select output, log Bucket

- We change the Output Location to our bucket. Just replace the bucket name, leave folders unchanged. Hit Continue



The screenshot shows a dialog box titled "Configure Sample Application" with a close button (X) in the top right corner. Inside the dialog, there is an information icon (i) followed by the text "Select a sample application to auto-populate the Create Cluster page". Below this, there are four configuration sections:


- Select sample application:** A dropdown menu currently showing "Word count".
- Output location:** A text input field containing "s3://zoran01/python/" with a folder icon to its right.
- Logging:** A checkbox labeled "Enabled" which is checked. Below it is another text input field containing "s3://zoran-log/python" with a folder icon to its right, and a placeholder text "s3://<bucket-name>/<folder>/" below that.
- Debugging:** A checkbox labeled "Enabled" which is checked.



At the bottom right of the dialog, there are two buttons: "Cancel" and "Ok".

Configure Instances, provide key pair

- Examine Step and hit Create cluster

Steps

 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](#)

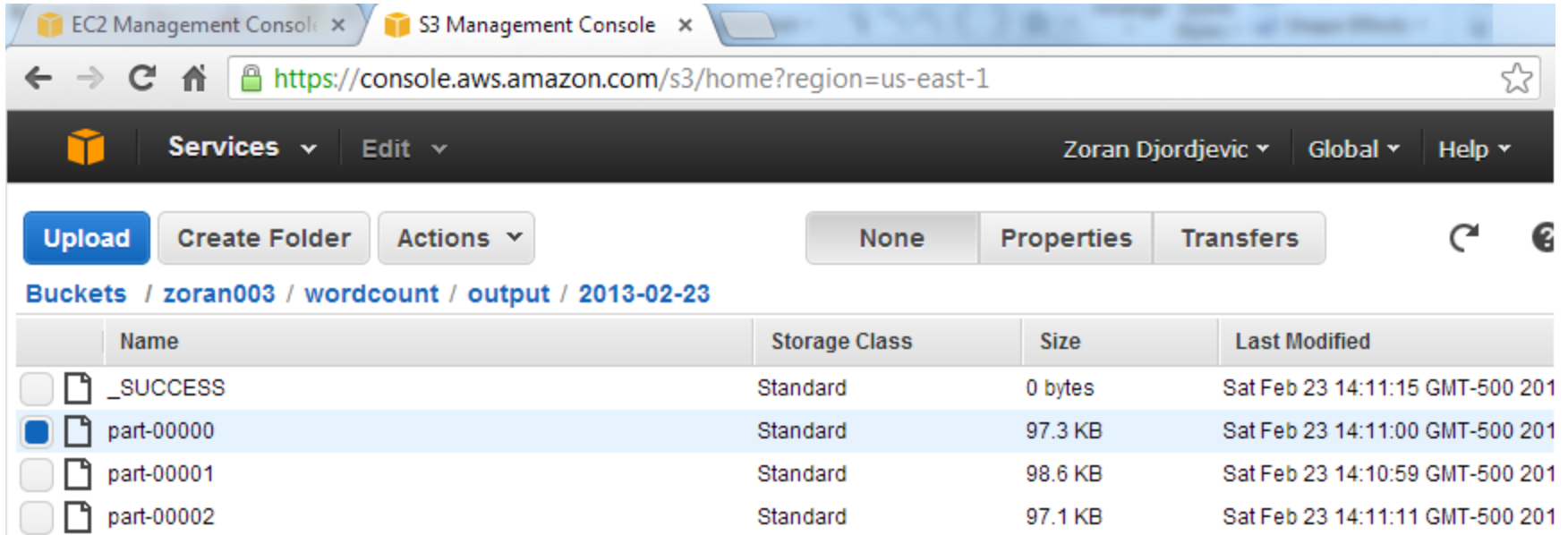
Name	Action on failure	JAR S3 location	Arguments	
Word count	Terminate cluster	/home/hadoop/contrib /streaming/hadoop- streaming.jar	-mapper s3n://elasticmapreduce /samples/wordcount /wordSplitter.py -reducer aggregate -input s3n://elasticmapreduce /samples/wordcount/input -output s3://zoran01/python/	 

Add step

Select a step ▼

Once Job Flow is Waiting you are Done

- Once the job flow moves to Waiting state you can go to your S3 buckets and examine the results

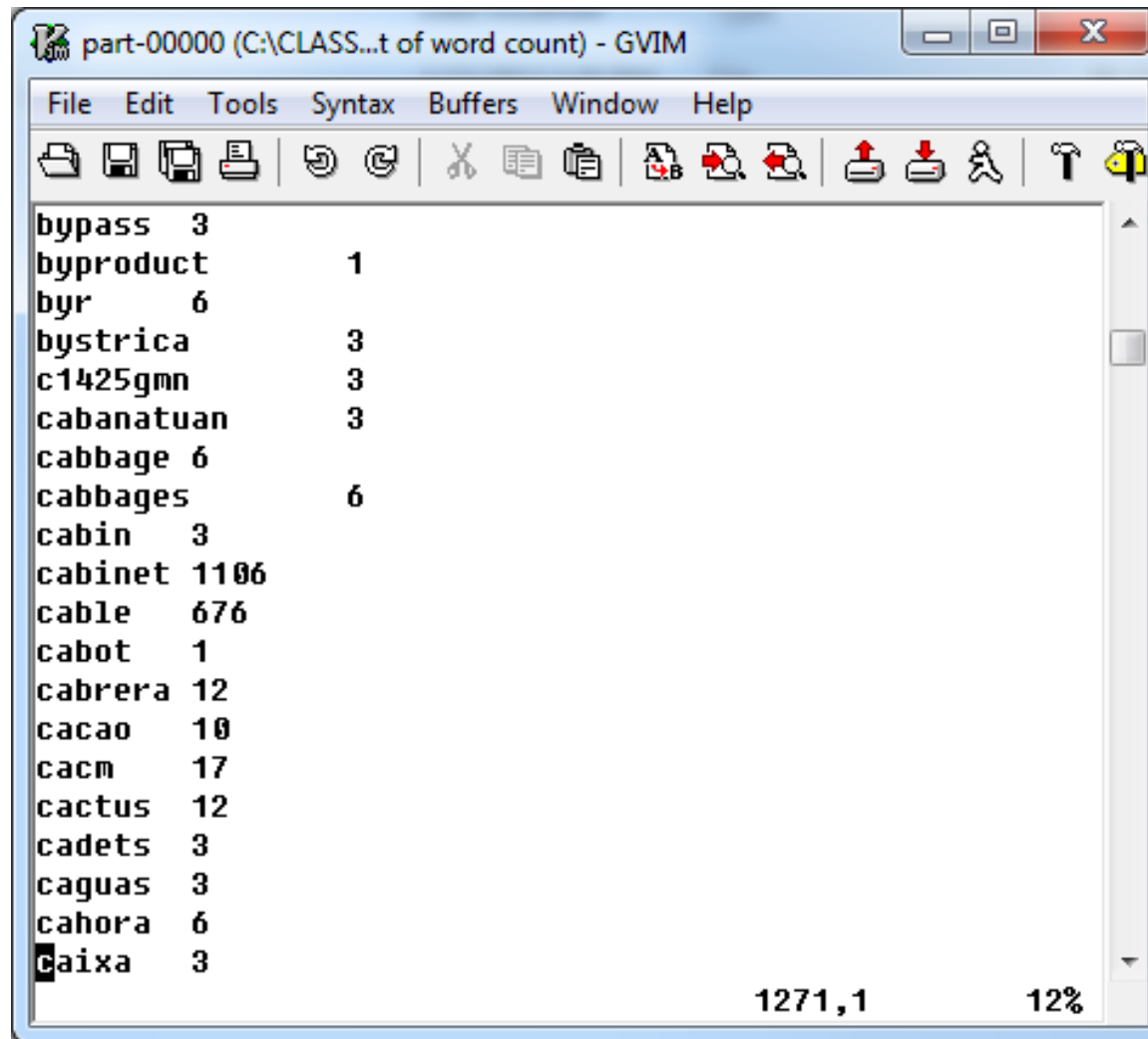


The screenshot shows the AWS S3 Management Console interface. The browser address bar displays <https://console.aws.amazon.com/s3/home?region=us-east-1>. The console header includes the AWS logo, 'Services' and 'Edit' dropdowns, and the user 'Zoran Djordjevic' with 'Global' and 'Help' links. The main content area shows the 'Buckets' section with a breadcrumb path: 'zoran003 / wordcount / output / 2013-02-23'. Above the file list are buttons for 'Upload', 'Create Folder', and 'Actions', along with tabs for 'None', 'Properties', and 'Transfers'. The file list table has columns for 'Name', 'Storage Class', 'Size', and 'Last Modified'. The files listed are '_SUCCESS' (0 bytes), 'part-00000' (97.3 KB), 'part-00001' (98.6 KB), and 'part-00002' (97.1 KB). The 'part-00000' file is selected.

	Name	Storage Class	Size	Last Modified
<input type="checkbox"/>	_SUCCESS	Standard	0 bytes	Sat Feb 23 14:11:15 GMT-500 201
<input checked="" type="checkbox"/>	part-00000	Standard	97.3 KB	Sat Feb 23 14:11:00 GMT-500 201
<input type="checkbox"/>	part-00001	Standard	98.6 KB	Sat Feb 23 14:10:59 GMT-500 201
<input type="checkbox"/>	part-00002	Standard	97.1 KB	Sat Feb 23 14:11:11 GMT-500 201

- You can download any of the files, like part-00000 and read the word count. Result is presented on the

Word Count Output



part-00000 (C:\CLASS...t of word count) - GVIM

bypass	3	
byproduct		1
byr	6	
bystrica		3
c1425gmn		3
cabanatuan		3
cabbage	6	
cabbages		6
cabin	3	
cabinet	1106	
cable	676	
cabot	1	
cabrera	12	
cacao	10	
cacm	17	
cactus	12	
cadets	3	
caguas	3	
cahora	6	
caixa	3	

1271,1 12%

Select EC2 Service and Review State of Your Instance

- Select EC2 Service first and then on the EC2 Dashboard, select Instances

The image shows two overlapping screenshots of the AWS Management Console. The top screenshot displays the 'All AWS Services' grid with 'EC2' selected in the 'History' sidebar. The bottom screenshot shows the 'EC2 Dashboard' with the 'INSTANCES' link highlighted in a red box. The 'Resources' section on the right indicates 0 Running Instances, 0 Elastic IPs, 0 Volumes, and 1 Snapshot.

EC2 Management Console

Services Edit Zoran Djordjevic N. Virginia

History: EC2, S3, Console Home

All AWS Services: Compute & Networking, Storage & Content Delivery, Database, CloudFormation, CloudFront, CloudSearch, CloudWatch, Elastic Beanstalk, Elastic MapReduce, Elastic Transcoder NEW, Glacier

EC2 Dashboard

Events

INSTANCES (highlighted)

Instances Spot Requests

Resources

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

- 0 Running Instances
- 0 Elastic IPs
- 0 Volumes
- 1 Snapshot

@Zoran B. Djordjevic

SSH to the Master Node

- Use user `hadoop`:

```
$ ssh -i e63.pem hadoop@ec2-54-221-152-76.compute-1.amazonaws.com
```

```
The authenticity of host 'ec2-54-221-152-76.compute-1.amazonaws.com  
(54.221.152.76)' can't be established.
```

```
RSA key fingerprint is d2:70:f7:30:ac:ea:b2:bc:19:0e:5e:05:9b:7a:57:cf.
```

```
Are you sure you want to continue connecting (yes/no)? yes
```

```
Warning: Permanently added 'ec2-54-221-152-76.compute-  
1.amazonaws.com,54.221.152.76' (RSA) to the list of known hosts.
```

```
Linux (none) 3.2.30-49.59.amzn1.x86_64 #1 SMP Wed Oct 3 19:54:33 UTC 2012  
x86_64
```

```
-----  
Welcome to Amazon Elastic MapReduce running Hadoop and Debian/Squeeze.
```

```
Hadoop is installed in /home/hadoop. Log files are in /mnt/var/log/hadoop.  
/mnt/var/log/hadoop/steps for diagnosing step failures.
```

```
The Hadoop UI can be accessed via the following commands:
```

```
JobTracker      lynx http://localhost:9100/
```

```
NameNode        lynx http://localhost:9101/
```

```
hadoop@ip-10-185-27-143:~$ pwd
```

```
/home/hadoop
```

- **Please note. EMR Console might suggest that you should connect as root, what might be wrong.**

On EC2 Dashboard

Launch Instance		Connect	Actions ▾						
Filter: All instances ▾		All instance types ▾		Search Instances			1 to 3 of 3 Instances		
<input type="checkbox"/>	Name	Instance ID ▾	Instance Type ▾	Availability Zone ▾	Instance State ▾	Status Checks ▴	Alarm Status ▾	Public DNS	
<input type="checkbox"/>		i-c9c206b3	m1.small	us-east-1d	● running	⚠ Insufficient...	None	ec2-50-17-18-35.com	
<input type="checkbox"/>		i-c7c206bd	m1.small	us-east-1d	● running	⚠ Insufficient...	None	ec2-184-73-59-219.c	
<input checked="" type="checkbox"/>		i-3cabe358	m1.small	us-east-1d	● running	⚠ Insufficient...	None	ec2-54-221-152-76.c	

Instance: i-3cabe358 Public DNS: ec2-54-221-152-76.compute-1.amazonaws.com

Description	Status Checks	Monitoring	Tags
Instance ID	i-3cabe358	Public DNS	ec2-54-221-152-76.compute-1.amazonaws.com
Instance state	running	Public IP	54.221.152.76
Instance type	m1.small	Elastic IP	-
Private DNS	ip-10-185-27-143.ec2.internal	Availability zone	us-east-1d
Private IPs	10.185.27.143	Security groups	ElasticMapReduce-master. view rules
Secondary private IPs	-	Scheduled events	No scheduled events
VPC ID	-	AMI ID	debian-amd64.manifest.xml (ami-9f2e78f6)
Subnet ID	-	Platform	-
Network interfaces	-	IAM role	-
Source/dest. check	False	Key pair name	hu0906
EBS-optimized	False	Owner	951414139794
Root device type	instance-store	Launch time	2013-11-15T16:07:05.000Z (less than one hour)
Root device	-	Termination protection	True
Block devices	-	Lifecycle	normal
		Monitoring	basic
		Alarm status	None

Security Group: ElasticMapReduce-master

- Let us open ports 9100 and 9101

Create Security GroupDelete

Viewing: EC2 Security Groups Search

1 to 6 of 6 Items

	Group ID	Name	VPC ID	Description
<input type="checkbox"/>	sg-4996e822	launch-wizard-2		launch-wizard-2 created on Thursday, October 24, 2013 5:07:00 PM UTC-4
<input checked="" type="checkbox"/>	sg-51134839	ElasticMapReduce-master		Master group for Elastic MapReduce
<input type="checkbox"/>	sg-79adcd12	db_group		Demo security group for RDS access
<input type="checkbox"/>	sg-edcf2c84	default		default group
<input type="checkbox"/>	sg-03cbb868	launch-wizard-1		launch-wizard-1 created on Friday, October 18, 2013 2:44:56 PM UTC-4
<input type="checkbox"/>	sg-5313483b	ElasticMapReduce-slave		Slave group for Elastic MapReduce

1 Security Group selected

Security Group: ElasticMapReduce-master

DetailsInbound

Create a new rule: Custom TCP rule

Port range: 9100-9101
(e.g., 80 or 49152-65535)


Source: 0.0.0.0/0
(e.g., 192.168.2.0/24, sg-47ad482e, or 1234567890/default)

Add Rule

Apply Rule Changes

ICMP	Port (Service)	Source	Action
	ALL	sg-51134839 (ElasticMapReduce-master)	Delete
	ALL	sg-5313483b (ElasticMapReduce-slave)	Delete
TCP	Port (Service)	Source	Action
	0 - 65535	sg-51134839 (ElasticMapReduce-master)	Delete
	0 - 65535	sg-5313483b (ElasticMapReduce-slave)	Delete
	22 (SSH)	0.0.0.0/0	Delete
	8443 (HTTPS*)	207.171.167.101/32	Delete
	8443 (HTTPS*)	207.171.167.25/32	Delete
	8443 (HTTPS*)	207.171.167.26/32	Delete

<http://ec2-54-221-152-76.compute-1.amazonaws.com:9101/dfshealth.jsp>

 ec2-54-221-152-76.compute-1.amazonaws.com:9101/dfshealth.jsp

NameNode 'ip-10-185-27-143.ec2.internal:9000'

Started: Fri Nov 15 16:09:57 UTC 2013
Version: 1.0.3, r
Compiled: Wed Oct 2 12:17:08 PDT 2013 by Elastic MapReduce
Upgrades: There are no upgrades in progress.

[Browse the filesystem](#)
[Namenode Logs](#)

Cluster Summary

9 files and directories, 1 blocks = 10 total. Heap Size is 25.12 MB / 185.62 MB (13%)

Configured Capacity	:	297.07 GB
DFS Used	:	88 KB
Non DFS Used	:	0 KB
DFS Remaining	:	297.07 GB
DFS Used%	:	0 %
DFS Remaining%	:	100 %
Live Nodes	:	2
Dead Nodes	:	0
Decommissioning Nodes	:	0
Number of Under-Replicated Blocks	:	0

NameNode Storage:

Storage Directory	Type	State
/mnt/var/lib/hadoop/dfs-name	IMAGE_AND_EDITS	Active

Hadoop Environment

- hadoop is also an executable which actually controls your cluster.
You can for example, run the following Linux commands

```
hadoop@domU-12-31-39-00-69-A7:~$ pwd
```

```
/home/hadoop
```

```
hadoop@domU-12-31-39-00-69-A7:~$ which hadoop
```

```
/home/hadoop/bin/hadoop
```

```
hadoop@domU-12-31-39-00-69-A7:~$ which java
```

```
/usr/bin/java
```

```
hadoop@domU-12-31-39-00-69-A7:~$ ls
```

PATCHES.txt	hadoop-core-1.0.3.jar	hadoop-tools.jar
bin	hadoop-core.jar	lib
conf	hadoop-examples-1.0.3.jar	lib64
contrib	hadoop-examples.jar	libexec
etc	hadoop-miniclusterc-1.0.3.jar	native
hadoop-ant-1.0.3.jar	hadoop-test-1.0.3.jar	sbin
. . . .		

Distributed File System, dfs command

- Hadoop has access not only to the local, Linux, file system. It also has its own distributed file system (HDFS Hadoop Distributed File System)
- We access that file system through hadoop file system shell, `dfs`.
Type

```
$ hadoop dfs
```

- and you will get a long list of options. We will present those options on the next slide. Some of those resemble Unix (Linux) commands. Some are different.
- We use those commands to create directories in the HDFS, copy files between HDFS and the local file system, Internet and AWS S3 buckets.
- When you use `dfs`, you always prefix it with `hadoop`.

File system shell dfs

```
hadoop@domU-12-31-39-00-69-A7:~$ hadoop dfs
```

```
Usage: java FsShell
```

```
[-ls <path>]  
[-lsr <path>]  
[-du <path>]  
[-dus <path>]  
[-count[-q] <path>]  
[-mv <src> <dst>]  
[-cp <src> <dst>]  
[-rm [-skipTrash] <path>]  
[-rmr [-skipTrash] <path>]  
[-expunge]  
[-put <localsrc> ... <dst>]  
[-copyFromLocal <localsrc> ... <dst>]  
[-moveFromLocal <localsrc> ... <dst>]  
[-get [-ignoreCrc] [-crc] <src> <localdst>]  
[-getmerge <src> <localdst> [addnl]]  
[-cat <src>]  
[-text <src>]  
[-copyToLocal [-ignoreCrc] [-crc] <src> <localdst>]  
[-moveToLocal [-crc] <src> <localdst>]
```

File system shell dfs

```
[-moveToLocal [-crc] <src> <localdst>]
[-mkdir <path>]
[-setrep [-R] [-w] <rep> <path/file>]
[-touchz <path>]
[-test -[ezd] <path>]
[-stat [format] <path>]
[-tail [-f] <file>]
[-chmod [-R] <MODE[,MODE]... | OCTALMODE> PATH...]
[-chown [-R] [OWNER][:[GROUP]] PATH...]
[-chgrp [-R] GROUP PATH...]
[-help [cmd]]
```

Generic options supported are

-conf <configuration file>	specify an application configuration file
-D <property=value>	use value for given property
-fs <local namenode:port>	specify a namenode
-jt <local jobtracker:port>	specify a job tracker
-files <comma separated list of files>	specify comma separated files to be copied to the map reduce cluster
-libjars <comma separated list of jars>	specify comma separated jar files to include in the classpath.

File system shell `dfs`

`-libjars <comma separated list of jars>` specify comma separated jar files to include in the classpath.

`-archives <comma separated list of archives>` specify comma separated archives to be unarchived on the compute machines.

The general command line syntax is

`bin/hadoop command [genericOptions] [commandOptions]`

- For example, we can use `dfs` to fetch the Python script used in our job flow:

```
$ hadoop dfs -copyToLocal\  
s3://elasticmapreduce/samples/wordcount/wordSplitter.py .
```

- The last dot (.) on the line is significant. This is “this” directory.
- Notice that options following `dfs` shell always start with a dash.
- Examine local directory and see the local copy of `wordSplitter.py`

```
hadoop@domU-12-31-39-00-69-A7:~$ ls -la wordSplitter.py  
-rw-r--r-- 1 hadoop hadoop 294 Feb 23 19:50 wordSplitter.py
```


wordSplitter.py

- We can vi the Python script or transfer it to our local Windows or Mac terminal and discover that it reads like:

```
#!/usr/bin/python
import sys
import re

def main(argv):
    pattern = re.compile("[a-zA-Z][a-zA-Z0-9]*")
    for line in sys.stdin:
        for word in pattern.findall(line):
            print "LongValueSum:" + word.lower() + "\t" + "1"

if __name__ == "__main__":
    main(sys.argv)
```

Input Data

- We could similarly use good services of hadoop's distributed file system shell `dfs` and first examine the input folder and then fetch the input data we used in the job flow.

```
hadoop@domU-12-31-39-00-69-A7:~$ hadoop dfs -ls  
s3://elasticmapreduce/samples/wordcount/input
```

Found 12 items

```
-rwxrwxrwx 1 2392524 2009-04-02 02:55 /samples/wordcount/input/0001  
-rwxrwxrwx 1 2396618 2009-04-02 02:55 /samples/wordcount/input/0002  
-rwxrwxrwx 1 1593915 2009-04-02 02:55 /samples/wordcount/input/0003  
-rwxrwxrwx 1 1720885 2009-04-02 02:55 /samples/wordcount/input/0004  
-rwxrwxrwx 1 2216895 2009-04-02 02:55 /samples/wordcount/input/0005  
-rwxrwxrwx 1 1906322 2009-04-02 02:55 /samples/wordcount/input/0006  
-rwxrwxrwx 1 1930660 2009-04-02 02:55 /samples/wordcount/input/0007  
-rwxrwxrwx 1 1913444 2009-04-02 02:55 /samples/wordcount/input/0008  
-rwxrwxrwx 1 2707527 2009-04-02 02:55 /samples/wordcount/input/0009  
-rwxrwxrwx 1 327050 2009-04-02 02:55 /samples/wordcount/input/0010  
-rwxrwxrwx 1 8 2009-04-02 02:55 /samples/wordcount/input/0011  
-rwxrwxrwx 1 8 2009-04-02 02:55 /samples/wordcount/input/0012
```

Copy input file 0001 to Local File System

```
hadoop@domU-12-31-39-00-69-A7:~$ hadoop dfs -copyToLocal  
s3://elasticmapreduce/samples/wordcount/input/0001 .  
13/02/23 20:06:21 INFO s3native.NativeS3FileSystem: Opening  
's3://elasticmapreduce/samples/wordcount/input/0001' for  
reading  
hadoop@domU-12-31-39-00-69-A7:~$ ls -la 0001  
-rw-r--r-- 1 hadoop hadoop 2392524 Feb 23 20:06 0001  
hadoop@domU-12-31-39-00-69-A7:~$
```

- Unix (Linux) utilities tail and head will tell us what are the lines at the end and beginning of file 001.

tail 0001, head 0001

```
hadoop@domU-12-31-39-00-69-A7:~$ tail 0001
```

males age 16-49: 7,322,965

females age 16-49: 6,859,064 (2008 est.)

Manpower fit for military service:

males age 16-49: 4,886,103

females age 16-49: 5,525,764 (2009 est.)

Manpower reaching militarily significant

age annually:

male: 365,567

female: 352,643 (2009 est.)

Military expenditures:

1.6% of GDP (2006)

```
hadoop@domU-12-31-39-00-69-A7:~$ head 0001
```

CIA -- The World Factbook -- Country Listing

World Factbook Home

The World Factbook

 ;

Country Listing . . .

Terminate all Instances

- Since we had enough for the day, we should terminate all instances, so that we stop incurring additional charges.
- Select all instances, and under Actions select Terminate, and Yes Terminate.

The screenshot shows the AWS Management Console interface for EC2 instances. The 'Launch Instance' button is visible at the top left. Below it, the 'Viewing: Running Instances' filter is applied. A table of instances is shown with columns for Name, Instance ID, and State. Three instances are listed, all with a state of 'running'. The 'Actions' dropdown menu is open, showing a list of actions. The 'Terminate' option is highlighted in orange. Below the 'Actions' menu, the 'Description' tab is selected, showing details for three instances: i-14ce7f67, i-1acf7e69, and i-18cf7e6b.

State	Status Check	Alarm Status	Monitoring
running	2/2 checks passed	none	basic
running	2/2 checks passed	none	basic
running	2/2 checks passed	none	basic

3 EC2 Instances selected

EC2 Instance Details

Description

i-14ce7f67: ec2-...
i-1acf7e69: ec2-...
i-18cf7e6b: ec2-...

Actions

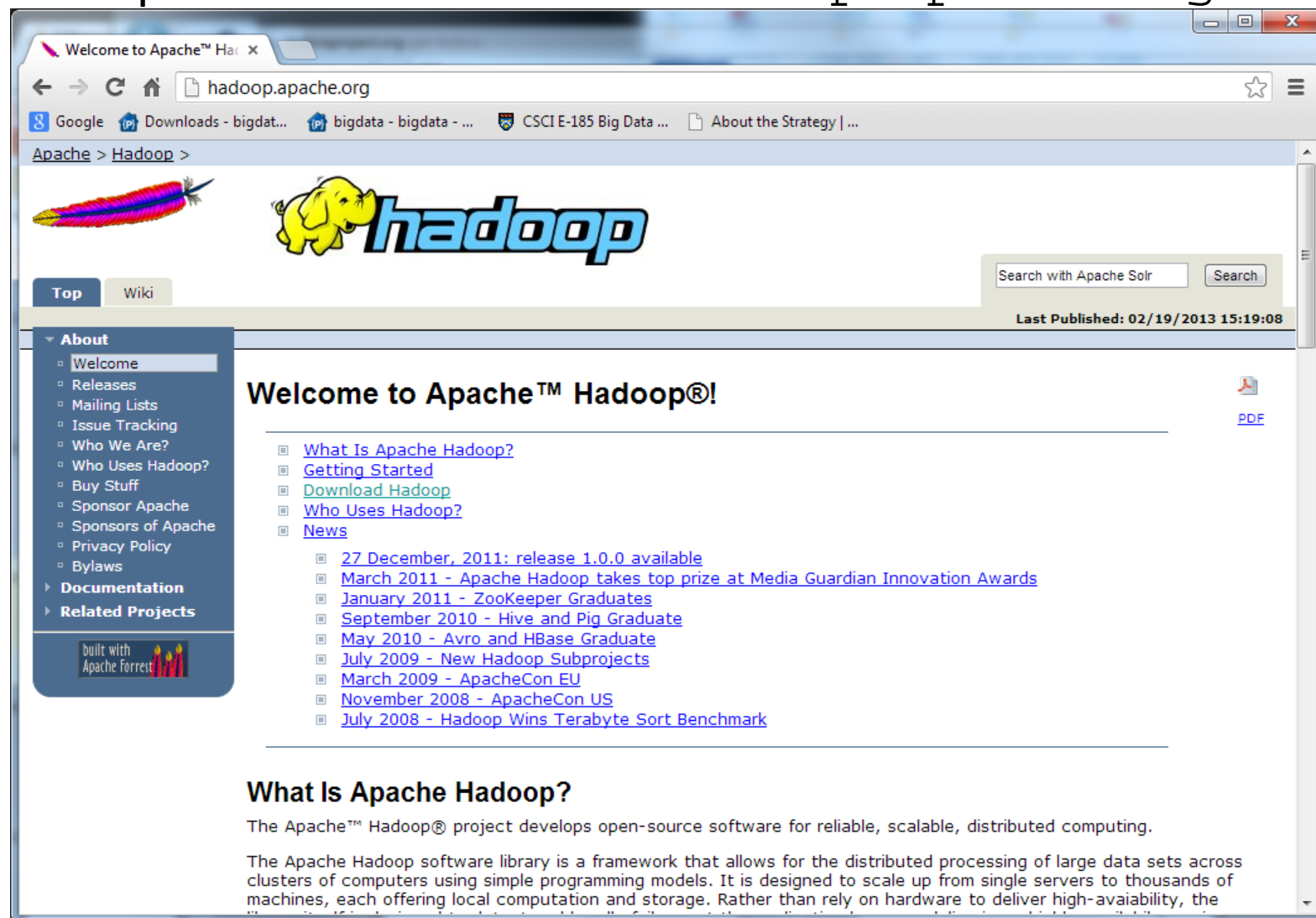
Terminate
Reboot
Stop
Start

Run Hadoop on Your Own Machine or Cluster

- Typically one wants to install a recent release of Hadoop on a recent release of a popular Linux OS.
- \$30 or \$60 for the academic license for a Red Hat OS is not bad. Major manufacturers, like IBM, like Red Hat.
- Fedora is open source, free version of Red Hat.
- Another option is CentOS OS. Free. For example, Cloudera is publishing its downloadable VMs on CentOS.
- Yet another option is Ubuntu. It is free. You might want to contribute and help save the world.
- SUSE as well. A few others.

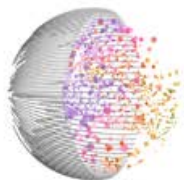
Hadoop

- Hadoop could be fetched from `hadoop.apache.org`



Cloudera

- People who invented Hadoop are mostly at Cloudera.



Ask Bigger Questions

Cloudera develops open-source software for a world dependent on Big Data. With Cloudera, businesses and other organizations can now interact with the world's largest data sets at the speed of thought — and ask bigger questions in the pursuit of discovering something incredible.

@Zoran B. Djordjevic

Cloudera CDH

- Cloudera offers CDH (Cloudera Distribution Hadoop including Apache Hadoop). The latest appears to be CDH 4.5

<https://ccp.cloudera.com/display/SUPPORT/CDH+Downloads>

- We read documentation first. We always do.
- If you have a 32 bit machine, you better do. You might be in trouble.
- Most of Cloudera tools are available for 64 bit machines only.

Cloudera CDH

← → ↺ 🏠 <https://www.cloudera.com/content/support/en/downloads/download-components/download-products.html?prc> ☆

📱 Apps 🏠 E-185 Big Data 📺 Ansambl Svilen Kon... 📖 Ebook library Booko... 🏠 E-90 Cloud 🏠 E-175 🌐 it-books.info 📺 ServiceNow Wiki 📱 Google »

Support Dev Center PARTNERS Sign In Register Contact Us Downloads

Contact Sales: 866-843-7207

cloudera
Ask Bigger Questions

Search 🔍

COMMUNITY | DOCUMENTATION | **DOWNLOADS** | KNOWLEDGEBASE | CASES | SUPPORT INFO

CDH

- Cloudera QuickStart VM
- Previous Cloudera Manager Releases
- Connectors
- Components
- Cloudera Manager Download

CDH

CDH consists of 100% open source Apache Hadoop plus nine other open source projects from the Hadoop ecosystem. CDH is thoroughly tested and certified to integrate with the widest range of operating systems and hardware, databases and data warehouses, and business intelligence and ETL systems.



See CDH Version and Packaging Information for a list of project packages and versions included in this distribution, as well as package version information for previous CDH releases.

Version: CDH 4.5 ▼

CDH 4.5

Last update: 26 Nov 2013

Installation Options

-  [Linux Packages](#)
-  [Quick Start VM](#)

[Release Notes](#)
[CDH4 Documentation](#)

CDH Requirements and Supported Versions

<https://ccp.cloudera.com/display/CDH4DOC/CDH4+Documentation>

Operating Systems

CDH4 provides packages for Red-Hat-compatible, SLES, Ubuntu, and Debian systems as described below.

Operating System	Version	Packages
Red Hat compatible		
Red Hat Enterprise Linux (RHEL)	5.7	64-bit
	6.2	64-bit, 32-bit
CentOS	5.7	64-bit
	6.2	64-bit, 32-bit
Oracle Linux with Unbreakable Enterprise Kernel	5.6	64-bit
SLES		
SLES Linux Enterprise Server (SLES)	11 with Service Pack 1 or later	64-bit
Ubuntu/Debian		
Ubuntu	Lucid (10.04) - Long-Term Support (LTS)	64-bit
	Precise (12.04) - Long-Term Support (LTS)	64-bit
Debian	Squeeze (6.03)	64-bit



Notes

- For production environments, 64-bit packages are recommended. Except as noted above, CDH4 provides **only** 64-bit packages.

32 vs 64 bit

- If your laptop is 32 bit and you want to run a local Hadoop installation with a 32 bit Hadoop on a 32 bit OS, the only choice is a Red Hat 6.2 or CentOS 6.2.
- If you have a new machine, it is most probably 64 bit and you are free to work with any OS that supports CHD4.5.
- Fedora is a free version of OS that is ahead of Red Hat in releases and serves as a development platform for what will eventually be packaged as Red Hat.
- CentOS 6.2 is a “repackaged” Red Hat 6.2 for non-commercial use and you are free to go with it.

Download VM for now

- Cloudera VM is a 64-bit VM, and requires a 64-bit host OS and a virtualization product that can support a 64-bit guest OS.
- This VM uses 4 GB of total RAM. The total system memory required varies depending on the size of your data set and on the other processes that are running.
- The demo VM file is in a [7-zip](#) format and is approximately 2 GB. Feel free to mirror internally or externally to minimize bandwidth usage.
- To use the VMware VM, you must use a player compatible with WorkStation 8.x or higher: Player 4.x or higher, ESXi 5.x or higher, or Fusion 4.x or higher. Older versions of WorkStation can be used to create a new VM using the same virtual disk (VMDK file), but some features in VMware Tools won't be available.