

Introduction to Elastic MapReduce

Csci E63 Big Data Analytics
Zoran B. Djordjević

@Zoran B. Djordjevic

1

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

@Zoran B. Djordjevic

2

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.

@Zoran B. Djordjevic

3

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.

@Zoran B. Djordjevic

4

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.

@Zoran B. Djordjevic

5

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.

@Zoran B. Djordjevic

6

Pricing is coming down

Region: 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

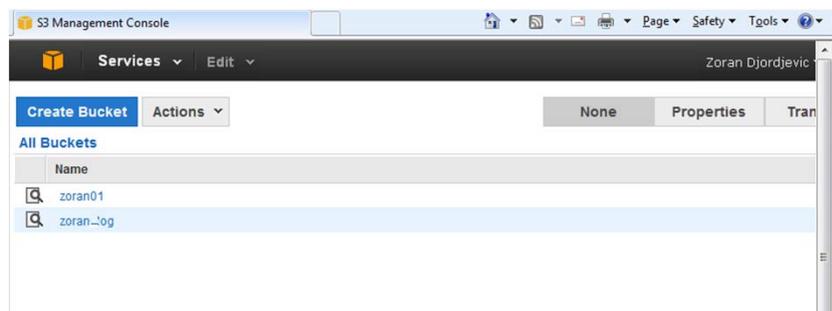
- EMR prices are paid atop of instance prices. These prices are approximate.
- Amazon EC2, Amazon S3 and Amazon SimpleDB charges are billed separately.

@Zoran B. Djordjevic

7

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 (-)



@Zoran B. Djordjevic

8

AWS Services, Select Elastic Map Reduce

The screenshot shows the AWS Management Console interface. The main heading is "AWS Services, Select Elastic Map Reduce". The console is organized into three main sections: "Welcome", "Amazon Web Services", and "Announcements". The "Amazon Web Services" section is categorized into "Compute & Networking", "Deployment & Management", "Storage & Content Delivery", and "Database". Key services listed include Direct Connect, EC2, Elastic MapReduce, Route 53, VPC, CloudFront, Glacier, S3, Storage Gateway, Database, and DynamoDB. The "Announcements" section features news about new AWS services and updates. A search bar is located at the top right of the console interface.

9

Select Create cluster

The screenshot shows the "Welcome to Amazon Elastic MapReduce" page. The main heading is "Welcome to Amazon Elastic MapReduce". Below the heading, there is a brief description of Amazon EMR and a "Create cluster" button. The page is divided into three sections: "Upload", "Create", and "Monitor", each with an icon and a brief description of the step. The "Upload" section shows a cloud icon with an upward arrow, the "Create" section shows a cluster icon with a gear, and the "Monitor" section shows a monitor icon with a downward arrow.

@Zoran B. Djordjevic

10

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

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

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	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
Pig	0.11.1.1	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>

@Zoran B. Djordjevic

11

Configure Hardware, Add EC2 key pair

Hardware Configuration

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 Use a Virtual Private Cloud (VPC) to process sensitive data or connect to a private network. [Create a VPC](#)

To create a cluster in a VPC, you must first create a VPC. For more information, [click here](#).

EC2 availability zone Launch the cluster in a specific EC2 Availability Zone.

	EC2 instance type	Count	Request spot	
Master	<input type="text" value="m1.small"/>	<input type="text" value="1"/>	<input type="checkbox"/>	The Master instance assigns Hadoop tasks to core and task nodes, and monitors their status.
Core	<input type="text" value="m1.small"/>	<input type="text" value="2"/>	<input type="checkbox"/>	Core instances run Hadoop tasks and store data using the Hadoop Distributed File System (HDFS).
Task	<input type="text" value="m1.small"/>	<input type="text" value="0"/>	<input type="checkbox"/>	Task instances run Hadoop tasks.

Security and Access

EC2 key pair 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 All other IAM users No other IAM users Control the visibility of this cluster to other IAM users. [Learn more](#)

Bootstrap Actions

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	<input type="text" value="Select a bootstrap action"/>		

@Zoran B. Djordjevic

12

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.

@Zoran B. Djordjevic

13

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 <input type="text" value="Pig program"/>			
<input type="button" value="Configure and add"/>			
Auto-terminate <input type="radio"/> Yes		Automatically terminate cluster after the last step is completed.	
<input checked="" type="radio"/> No		Keep cluster running until you terminate it.	
<input type="button" value="Cancel"/>			<input type="button" value="Create cluster"/>

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

- Hit **Configure and add** button

@Zoran B. Djordjevic

14

Select script, S3 bucket for, input, output

Add Step

Step type: Pig program

Name: Pig program

Script S3 location*: s3://elasticmapreduce/samples/pig-apache/do-reports2
s3://<bucket-name>/<path-to-file> S3 location of your Pig script.

Input S3 location: s3://elasticmapreduce/samples/pig-apache/input
s3://<bucket-name>/<folder>/ S3 location of your Pig input files.

Output S3 location: s3://zoran01/pig-apache
s3://<bucket-name>/<folder>/ S3 location of your Pig output files.

Arguments: Specify optional arguments for your script.

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

Buttons: 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

@Zoran B. Djordjevic

15

Cluster is Starting

Elastic MapReduce Cluster List Cluster Details EMR Help

Buttons: Add step, Resize, Clone, Terminate

Cluster: MyCluster01 **Starting**

Master public DNS: --

Tags: name = pig_example View All / Edit

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

Hardware

Master: Provisioning 1 m1.small
 Core: Provisioning 2 m1.small
 Task: --

Monitoring

Steps

Bootstrap Actions

@Zoran B. Djordjevic

16

Cluster is Running

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

The screenshot shows the AWS EMR console interface. At the top, there's a navigation bar with 'Elastic MapReduce', 'Cluster List', and 'Cluster Details'. Below this, there are buttons for 'Add step', 'Resize', 'Clone', and 'Terminate'. The main content area shows the cluster 'MyCluster01' in a 'Running' state. It provides various details including Master public DNS, Tags, Summary (ID, Creation date, Elapsed time), Configuration Details (AMI version, Hadoop distribution, Applications, Log URI), and Security/Network (Availability zone, Subnet ID, Key name, IAM role, Visible to all users). A 'Hardware' section lists Master and Core instances. At the bottom, there's a 'Monitoring' section and a 'Steps' section. A table at the bottom right shows a list of clusters with columns for Name, ID, and Status, with 'MyCluster01' listed as 'Waiting'.

There are 3 EC2 Instance, a Master and 2 Slaves

The screenshot shows the AWS EC2 console. At the top, there are buttons for 'Launch Instance', 'Connect', and 'Actions'. Below this, there's a filter for 'Running instances' and a search bar. A table lists three instances: i-e486ccc5, i-4781cb66, and i-5181cb70, all with 'm1.small' instance types and 'us-east-1d' availability zones. Below the table, detailed information for instance 'i-e486ccc5' is shown, including its public DNS, status, type, private DNS, private IPs, secondary private IPs, VPC ID, subnet ID, network interfaces, source/dest. check, EBS-optimized status, public DNS, public IP, elastic IP, availability zone, security groups, scheduled events, AMI ID, platform, IAM role, key pair name, owner, and launch time.

@Zoran B. Djordjevic

18

Job is Waiting, Steps are Completed

Elastic MapReduce > Cluster List > Cluster Details EN

Cluster: MyCluster01 Waiting Waiting after step completed

Master public DNS: ec2-50-17-121-66.compute-1.amazonaws.com

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

Summary	Configuration Details	Security/Network	Hardware
ID: j-EPR2LBM5ORDJ Creation date: 2014-02-21 11:52 (UTC-5) Elapsed time: 46 minutes Auto-terminate: No Termination protection: On	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: o63 IAM role: -- Visible to all users: None	Master: Running 1 m1.small Core: Running 2 m1.small Task: --

Monitoring

Steps

Filter: All steps 4 steps (all loaded) [View all interactive jobs](#) [View all jobs](#)

ID	Name	Status	Start time (UTC-6)	Elapsed time	Log files	Actions
s-3L2Y2Q0A2LTH5	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-2AZ5WWJRY8S1Y	Setup hive	Completed	2014-02-21 11:56	1 minute	View logs	View jobs
s-9TB08QCNPD5N	Setup hadoop	Completed	2014-02-21 11:56	36 seconds	View logs	View jobs

@Zoran B. Djordjevic 19

Results in S3 Buckets

All Buckets / zoran-log

pig-apache

None
P

All Buckets / zoran01 / pig-apache

Name	Storage Class	Size
<input type="checkbox"/> top_50_external_referrers	--	--
<input type="checkbox"/> top_50_ips	--	--
<input type="checkbox"/> top_50_search_terms_from_bing_google	--	--
<input type="checkbox"/> total_requests_bytes_per_hour	--	--

All Buckets / zoran01 / pig-apache / top_50_search_terms_from_bing_google

_SUCCESS

part-r-00000

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

```

C:\Users\zjordje\Downloads\part-r-00000 - Notepad
File Edit Search View Encoding Language Set
credential file.txt listener.ora 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 monowall+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"
}
```

@Zoran B. Djordjevic

21

credentials.json

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


```
{
  "access_id": "AKGGGGHJT7WWWFDTHJQ",
  "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.

@Zoran B. Djordjevic

22

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
```

@Zoran B. Djordjevic

23

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

@Zoran B. Djordjevic

24

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 a bucket named 'zordjje002'. The 'Permissions' section is expanded, showing three permission entries:

- Grantee:** zordjje (User) with permissions: List, Upload/Delete, View.
- Grantee:** Authenticated Users with permissions: List, Upload/Delete.
- Grantee:** Everyone with permissions: List, Upload/Delete.
- Grantee:** Log Delivery with permissions: List, Upload/Delete.

Buttons at the bottom include 'Add more permissions', 'Add bucket policy', and 'Add CORS'.

@Zoran B. Djordjevic

25

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`

The screenshot shows the 'Create Cluster' page in the AWS EMR console. The 'Cluster Configuration' section is visible, with the following settings:

- Cluster name:** My cluster
- Termination protection:** Yes (selected)
- Logging:** Enabled
- Log folder S3 location:** s3://zoran01/python/
- Debugging:** Enabled

There are help links for 'Learn more' and 'Copy the cluster's log files automatically to S3'.

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

@Zoran B. Djordjevic

27

Configure Instances, provide key pair

- Examine Step and hit Create cluster

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
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/

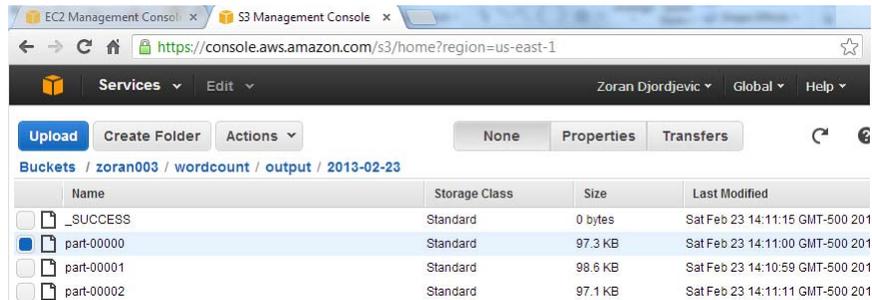
Add step

@Zoran B. Djordjevic

28

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

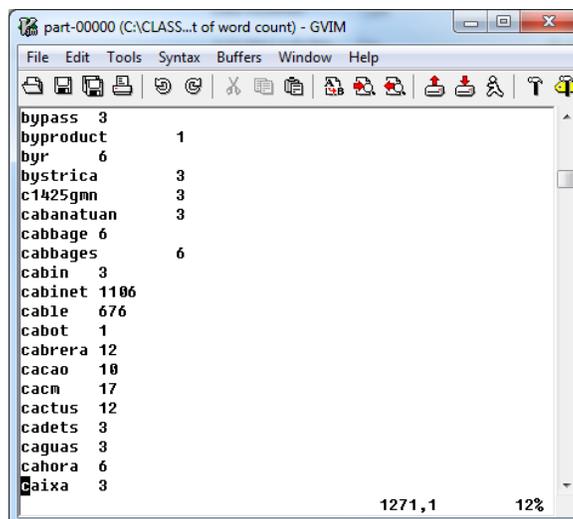


- You can download any of the files, like part-00000 and read the word count. Result is presented on the

@Zoran B. Djordjevic

29

Word Count Output



@Zoran B. Djordjevic

30

Select EC2 Service and Review State of Your Instance

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

The screenshot shows the AWS Management Console interface. At the top, the 'Services' menu is open, and 'EC2' is selected under 'All AWS Services'. Below this, the 'EC2 Dashboard' is displayed. On the left sidebar, the 'INSTANCES' section is expanded, and the 'Instances' link is highlighted with a red rectangular box. The main content area shows the 'Resources' section, which states: 'You are using the following Amazon EC2 resources in the US East (N. Virginia) region:'. Below this, it lists: '0 Running Instances', '0 Elastic IPs', '0 Volumes', and '1 Snapshot'.

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.**

@Zoran B. Djordjevic

32

On EC2 Dashboard

The screenshot shows the AWS Management Console interface for an EC2 instance. At the top, there are buttons for 'Launch Instance', 'Connect', and 'Actions'. Below that is a filter section with 'All Instance types' selected and a search bar. A table lists three instances, with the third one, 'i-3cabe358', selected. Below the table, the 'Instance Details' page for 'i-3cabe358' is displayed, showing various configuration parameters like Instance ID, Instance state, Instance type, Private DNS, Private IPs, Secondary private IPs, VPC ID, Subnet ID, Network interfaces, Source/dest. check, EBS-optimized, Root device type, Root device, Block devices, Public DNS, Public IP, Elastic IP, Availability zone, Security groups, Scheduled events, AMI ID, Platform, IAM role, Key pair name, Owner, Launch time, Termination protection, Lifecycle, Monitoring, and Alarm status.

@Zoran B. Djordjevic 33

Security Group: ElasticMapReduce-master

- Let us open ports 9100 and 9101

The screenshot shows the 'Security Group: ElasticMapReduce-master' page in the AWS Management Console. It includes a 'Create Security Group' and 'Delete' button at the top. Below is a table of security groups, with 'sg-51134839' selected. The 'Details' section shows the 'Inbound' rules for this security group. A 'Custom TCP rule' is being created for ports 9100-9101. The table below lists the existing inbound rules:

Protocol	Port (Service)	Source	Action
ICMP			
TCP	9100-9101	0.0.0.0	Allow
TCP	0 - 65535	sg-51134839 (ElasticMapReduce-master)	Delete
TCP	0 - 65535	sg-5313483b (ElasticMapReduce-slave)	Delete
TCP	22 (SSH)	0.0.0.0/0	Delete
TCP	8443 (HTTPS*)	207.171.167.101/32	Delete
TCP	8443 (HTTPS*)	207.171.167.25/32	Delete
TCP	8443 (HTTPS*)	207.171.167.26/32	Delete

@Zoran B. Djordjevic 34

<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

@Zoran B. Djordjevic

35

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-miniclust-1.0.3.jar native
hadoop-ant-1.0.3.jar  hadoop-test-1.0.3.jar    sbin
. . . . .
```

@Zoran B. Djordjevic

36

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`.

@Zoran B. Djordjevic

37

## 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>]
```

@Zoran B. Djordjevic

38

## 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.
```

@Zoran B. Djordjevic

39

## 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
```

@Zoran B. Djordjevic

40

## 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)
```

@Zoran B. Djordjevic

41

## 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
```

@Zoran B. Djordjevic

42

## 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 0001.

@Zoran B. Djordjevic

43

## 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

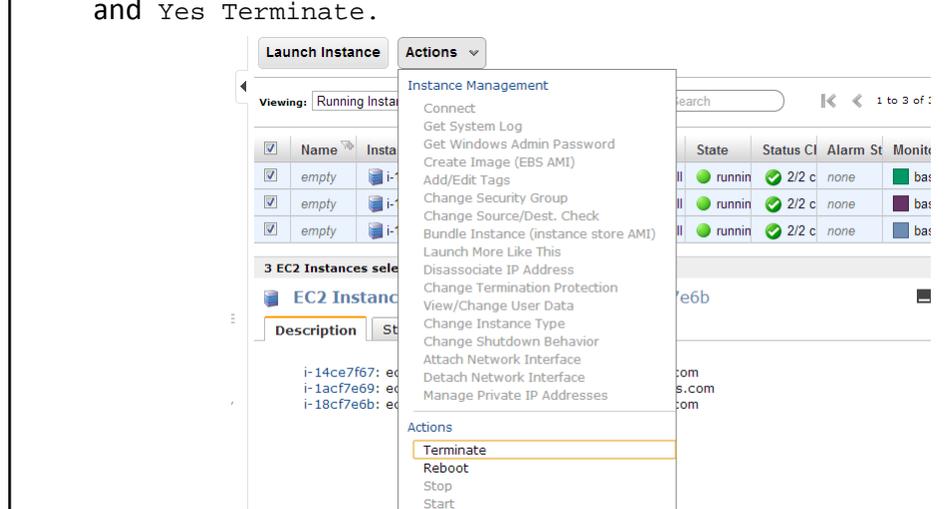
 Country Listing . . .
```

@Zoran B. Djordjevic

44

## 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.

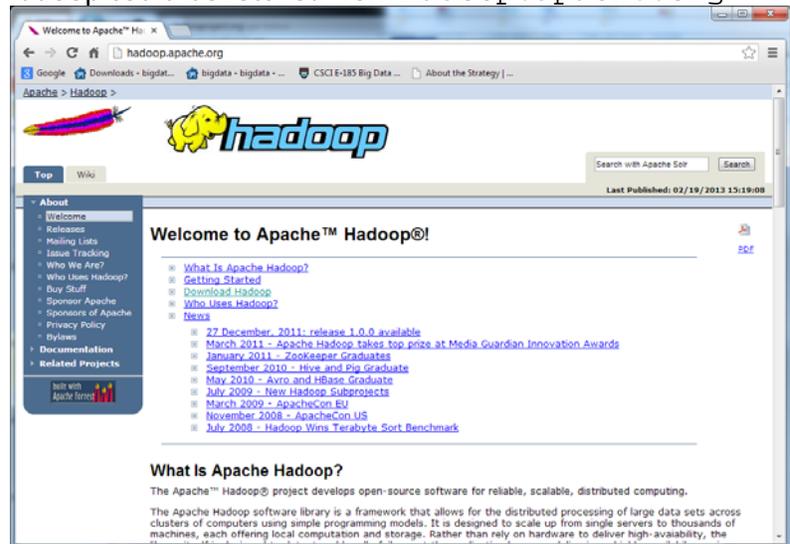


## Run Hadoop on Your Own Machine or Cluster

- 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.



@Zoran B. Djordjevic

48

## 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.

@Zoran B. Djordjevic

49

## Cloudera CDH

The screenshot shows the Cloudera website's CDH download page. The browser address bar displays the URL: <https://www.cloudera.com/content/support/en/downloads/download-components/download-products.html?prc>. The page features a navigation menu with links for Support, Dev Center, PARTNERS, Sign In, Register, Contact Us, and Downloads. The Cloudera logo is prominently displayed, along with a search bar and navigation tabs for COMMUNITY, DOCUMENTATION, DOWNLOADS, KNOWLEDGEBASE, CASES, and SUPPORT INFO. On the left, a sidebar lists various CDH-related links such as Cloudera QuickStart VM, Previous Cloudera Manager Releases, Connectors, Components, and Cloudera Manager Download. The main content area is titled 'CDH' and provides a description: '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.' Below this, it states: '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.' A 'Version:' dropdown menu is set to 'CDH 4.5'. Underneath, 'CDH 4.5' is highlighted, with a 'Last update: 26 Nov 2013' note. Two installation options are provided: 'Linux Packages' (with a Linux logo) and 'Quick Start VM' (with a VM icon). A 'Release Notes' link and 'CDH4 Documentation' link are also visible.

@Zoran B. Djordjevic

50

## 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       |
|-------------------------------------------------|-------------------------------------------|----------------|
| <b>Red Hat compatible</b>                       |                                           |                |
| 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         |
| <b>SLES</b>                                     |                                           |                |
| SLES Linux Enterprise Server (SLES)             | 11 with Service Pack 1 or later           | 64-bit         |
| <b>Ubuntu/Debian</b>                            |                                           |                |
| 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**.

51

## 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.

@Zoran B. Djordjevic

52

## 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.