Up and Running with Hadoop

By :
Brahmbhatt Spandan H

Ways to install :

1. Brew (MacOS only).
2. Manual Install.
3. Using Cloudera CM (GUI Based)
4. Using Apache Ambari (GUI Based)
5. Using Amazon Elastic mapreduce.

1. Using brew (PseudoDistributed mode)

Run the following command :

1. Install Brew :

ruby -e "$(curl -fsSL https://raw.githubusercontent.com/Homebrew/install/master/install)"

2. Install Hadoop :

brew install hadoop

3. Hadoop is install in

/usr/local/Cellar/hadoop

4. Edit hadoop-env.sh

Replace :

export HADOOP_OPTS="$HADOOP_OPTS -Djava.net.preferIPv4Stack=true"

with

export HADOOP_OPTS="$HADOOP_OPTS -Djava.net.preferIPv4Stack=true -Djava.security.krb5.realm= -Djava.security.krb5.kdc="

5.Edit Core-site.xml

```
<configuration>
  <property>
    <name>hadoop.tmp.dir</name>
    <value>/usr/local/Cellar/hadoop/hdfs/tmp</value>
    <description>A base for other temporary directories.</description>
  </property>
  <property>
    <name>fs.default.name</name>
    <value>hdfs://localhost:9000</value>
  </property>
</configuration>
```

6. Edit mapred-site.xml

Rename mapred-site.xml.template to mapred-site.xml and then add

```
<configuration>
    <property>
      <name>mapred.job.tracker</name>
      <value>localhost:9010</value>
    </property>
</configuration>
```

7.Edit hdfs-site.xml

```
<configuration>
    <property>
      <name>dfs.replication</name>
      <value>1</value>
    </property>
  </configuration>
```

8. Edit bash_profile file

~$ cd ~ //Run this command to go to home.

~$ vi .bash_profile

Add following lines after pressing “i” (i.e go in insert mode)

alias hstart="/usr/local/Cellar/hadoop/2.6.0/sbin/start-dfs.sh;/usr/local/Cellar/hadoop/2.6.0/sbin/start-yarn.sh"

alias hstop="/usr/local/Cellar/hadoop/2.6.0/sbin/stop-yarn.sh;/usr/local/Cellar/hadoop/2.6.0/sbin/stop-dfs.sh"

Now run following command :

~$ source ~/.bash_profile

~$ hadoop namenode -format

9. Before you run hadoop :

~ $ ssh-keygen -t rsa //Just press enter when asked to specify location to save file. Do not type anything.

// if it ask you to overwrite the file then do type no and skip step 9.

Enable Remote Login

“System Preferences” -> “Sharing”. Check “Remote Login”

~ $ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys

10. Run hadoop :

~$ hstart

~ $ jps // use this command to see hadoop services are properly running.

~ $ // Use this command to stop Hadoop.

Your Hadoop installation is at :

/usr/local/Cellar/hadoop

Reference : http://amodernstory.com/2014/09/23/installing-hadoop-on-mac-osx-yosemite/

2. Manual Install

Steps to install Hadoop 2.5.0 on AWS :

1. Run your instances from AWS EC2 console.
   a. Log into EC2 console. Click on EC2 → Launch Instance
   b. Select **Ubuntu Server 14.04 LTS (HVM), SSD Volume Type** - ami-9a562df2 (Free tier) (General Purpose)
   c. Change number of instance to 4 (1 to be Namenode.1-SNN and 2 will be Datanode).
   d. Create a new key pair and save it.
2. Connect to each instance using ssh (in MacOS) or use WinSCP and Putty (in Windows)

   ~ $ ssh -i keypair.pem ubuntu@publicDNS
3. Change hostname to be public DNS with command - "~ $ sudo hostname publicDNS".
4. Change host file with following command :

   ~ $sudo vi /etc/hosts

   Make entry in this format: ipaddress publicDNS
5. Repeat step 2 and 3 for all instances.
6. Repeat following commands on each instance :

```
--------------------------------------------------------------------------------
~ $sudo apt-get update
~ $sudo add-apt-repository ppa:webupd8team/java
~ $sudo apt-get update && sudo apt-get install oracle-jdk7-installer
~ $java -version
~ $wget http://apache.petsads.us/hadoop/common/hadoop-2.5.0/hadoop-2.5.0.tar.gz
~ $tar -xzvf hadoop-2.5.0.tar.gz
~ $mv hadoop-2.5.0 hadoop
~ $rm hadoop-2.5.0.tar.gz
~ $ vi .bashrc
    Add following lines use i key to go in insert mode and esc and :wq to quit:
              export HADOOP_CONF=/home/ubuntu/hadoop/conf
              export HADOOP_PREFIX=/home/ubuntu/hadoop
              #Set JAVA_HOME
              export JAVA_HOME=/usr/lib/jvm/java-7-oracle
              # Add Hadoop bin/ directory to path
              export PATH=$PATH:$HADOOP_PREFIX/bin
```

Run the following commands now

```
~ $source ~/.bashrc
~ $echo $HADOOP_PREFIX
~ $chmod 644 .ssh/authorized_keys
~ $chmod 400 keypair.pem
~ $eval `ssh-agent`
~ $ssh-add keypair.pem
```

Check connection to each using :

```
~ $ssh ubuntu@publicDNS
~ $exit
```

Now on master node run following command :

```
~ $cd hadoop/etc/hadoop
~ $vi hadoop-env.sh
     export JAVA_HOME=/usr/lib/jvm/java-7-oracle       //Add following line


~ $cd ~
~ $mkdir hdfstmp
~ $cd hadoop/etc/hadoop
```

7. Run on Master instance

Edit following files :

vi core-site.xml

```
----------
<property>
<name>fs.default.name</name>
<value>hdfs://ec2-54-173-155-242.compute-1.amazonaws.com:8020</value>
</property>
<property>
<name>hadoop.tmp.dir</name>
<value>/home/ubuntu/hdfstmp</value>
</property>
-----------
```

vi hdfs.site.xml

```
-----------
<property>
<name>dfs.replication</name>
<value>2</value>
</property>
<property>
<name>dfs.permissions</name>
<value>false</value>
</property>
------------
```

```
~ $ mv mapred-site.xml.template mapred.site.xml
vi mapred-site.xml

            ------------
            <property>
            <name>mapreduce.framework.name</name>
            <value>yarn</value>
            </property>
            -------------

vi yarn-site.xml
            ------------------
            <property>
            <name>yarn.nodemanager.aux-services</name>
            <value>mapreduce_shuffle</value>
            </property>
            <property>
            <name>yarn.nodemanager.aux-services.mapreduce.shuffle.class</name>
            <value>org.apache.hadoop.mapred.ShuffleHandler</value>
            </property>
            <property>
            <name>yarn.resourcemanager.resource-tracker.address</name>
            <value>ec2-54-173-155-242.compute-1.amazonaws.com:8025</value>
            </property>
            <property>
            <name>yarn.resourcemanager.scheduler.address</name>
            <value>ec2-54-173-155-242.compute-1.amazonaws.com:8030</value>
            </property>
            <property>
            <name>yarn.resourcemanager.address</name>
            <value>ec2-54-173-155-242.compute-1.amazonaws.com:8040</value>
            </property>
            ------------------
```

8. Run command for all slaves to transfer the config files.

```
~ $scp hadoop-env.sh core-site.xml hdfs-site.xml yarn-site.xml mapred-site.xml
ubuntu@publicDNS:/home/ubuntu/hadoop/etc/hadoop
```

9. On master update file to have slaves publicDNS:

```
~ $ vi slaves
     -------------
     publicDNS
     -------------
```

10. Start your cluster using :

```
~ $ bin/hadoop namenode -format
~ $sbin/hadoop-daemon.sh start namenode
~ $ sbin/hadoop-daemons.sh start datanode
~ $sbin/yarn-daemon.sh start resourcemanager
~ $sbin/yarn-daemons.sh start nodemanager
```

11. To check if HDFS is working:

```
vi sample.txt
hdfs dfs -mkdir -p /usr/spandan
hfs pdfs -copyFromLocal sample.txt /usr/spandan
```

12. Hadoop Web UI can be accessed from you system browser using

http://masterpublicDNS:50070

3. Using Cloudera CM5 **(Not free)**

Using Cloudera CM5, Hadoop cluster can be easily created. Although using CM5 is free, if you run it on Amazon AWS the system requirements for instances do not come in free tier and hence will incure a small cost. However if you run it on your own cluster of machines, it will be completely free.

Steps :

Create 4 instances from Amazon EC2 console. Select

Ubuntu Server 14.04 LTS (HVM), SSD Volume Type - ami-9a562df2 (Free tier) -- > Select m3.large which has 2 vCPU and 7.5 Gb Memory.

Save the keypair.

ssh into master node (anyone that you design as your master):

```
$ ssh -i your-key.pem ubuntu@ec2-xx-xx-xx-xx.compute-1.amazonaws.com
```

Run following commands :

```
$ wget http://archive.cloudera.com/cm4/installer/latest/cloudera-manager-installer.bin
$ chmod +x cloudera-manager-installer.bin
$ sudo ./cloudera-manager-installer.bin
```

You will be presented with a GUI in shell. Accept everything and let it install.

After you are complete with it. Open your browser and go to following link

http://ec2-xx-xx-xx-xx.compute-1.amazonaws.com:7180

Username and password is: admin

EC2 Management Console | Cloudera Manager Walkth | Amazon EC2 Instance Con | Global Knowledge Training | Welcome to Cloudera. Wh

ec2-54-212-156-230.us-west-2.compute.amazonaws.com:7180/cmf/license/wizard?returnUrl=%2Fcmf%2Fcloud-express-wizard%2Fwelcome

cloudera manager

Support admin

Welcome to Cloudera. Which edition do you want to deploy?

Upgrading to **Cloudera Enterprise** provides important features that help you manage and monitor your Hadoop clusters in mission-critical environments.

	Cloudera Standard	Cloudera Enterprise Trial ✔	Cloudera Enterprise
License	Free	60 Days Post trial period, the product will continue to function as **Cloudera Standard**. Your cluster and your data will remain **unaffected**.	Annual Subscription(s) Upload License
Node Limit	Unlimited	Unlimited	Unlimited
CDH	✔	✔	✔
Core Cloudera Manager Features	✔	✔	✔
Advanced Cloudera Manager Features (click link below for details)		✔	✔
Backup & Disaster Recovery †		✔	✔
Cloudera Navigator †		✔	✔
Cloudera Support			✔

† Purchased as separate products.

For full list of features available in **Cloudera Standard** and **Cloudera Enterprise**, click here.

Continue

EC2 Management Console | Cloudera Manager Walkth | Amazon EC2 Instance Con | Global Knowledge Trainin | Specify hosts for your CDH

ec2-54-212-156-230.us-west-2.compute.amazonaws.com:7180/cmf/express-wizard/hosts

cloudera manager

Support admin

Specify hosts for your CDH cluster installation.

Cloudera recommends including Cloudera Manager server's host because it is often used for the Cloudera Management Service, and because this will enable health monitoring for that host.

Hint: Search for hostnames and/or IP addresses using patterns.

SSH Port: 22 Search

Enter your dns here. One per line.

EC2 Management Console | Cloudera Manager Walkth | Amazon EC2 Instance Con | Global Knowledge Trainin | Specify hosts for your CDH

ec2-54-212-156-230.us-west-2.compute.amazonaws.com:7180/cmf/express-wizard/hosts

cloudera manager

Support admin

Specify hosts for your CDH cluster installation.

Cloudera recommends including Cloudera Manager server's host because it is often used for the Cloudera Management Service, and because this will enable health monitoring for that host.

Hint: Search for hostnames and/or IP addresses using patterns.

Every machine will come here.

11 hosts scanned, 11 running SSH. New Search

Expanded Query	Hostname (FQDN)	IP Address	Currently Managed	Result
10.224.10.204	ip-10-224-10-204.us-west-2.compute.internal	10.224.10.204	No	Host ready: 1,376 ms response time.
10.224.21.158	ip-10-224-21-158.us-west-2.compute.internal	10.224.21.158	No	Host ready: 916 ms response time.
10.224.35.42	ip-10-224-35-42.us-west-2.compute.internal	10.224.35.42	No	Host ready: 753 ms response time.
10.224.50.174	ip-10-224-50-174.us-west-2.compute.internal	10.224.50.174	No	Host ready: 1,495 ms response time.
10.224.50.211	ip-10-224-50-211.us-west-2.compute.internal	10.224.50.211	No	Host ready: 1,136 ms response time.
10.225.142.36	ip-10-225-142-36.us-west-2.compute.internal	10.225.142.36	No	Host ready: 0 ms response time.
10.226.158.189	ip-10-226-158-189.us-west-2.compute.internal	10.226.158.189	No	Host ready: 1,256 ms response time.
10.249.37.238	ip-10-249-37-238.us-west-2.compute.internal	10.249.37.238	No	Host ready: 2,478 ms response time.
10.249.78.137	ip-10-249-78-137.us-west-2.compute.internal	10.249.78.137	No	Host ready: 1,615 ms response time.
10.250.11.75	ip-10-250-11-75.us-west-2.compute.internal	10.250.11.75	No	Host ready: 1,900 ms response time.

EC2 Management Console | Cloudera Manager Walkth | Amazon EC2 Instance Con | Global Knowledge Trainin | Cluster Installation - Clou

ec2-54-212-156-230.us-west-2.compute.amazonaws.com:7180/cmf/express-wizard/wizard

cloudera manager

Support admin

Cluster Installation

Select Repository

Cloudera Manager Parcels are the easiest way for Cloudera Manager to manage the software on your cluster, by automating the deployment and upgrade of service binaries. Electing not to use parcels will require you to manually upgrade packages on all hosts in your cluster when software updates are available, and will prevent you from using Cloudera Manager's rolling upgrade capabilities.

Choose Method:

- ○ Use Packages
- ◉ Use Parcels (Recommended)
  - > More Options

Use parcels

SOLR
- ◉ SOLR-1.0.0-1.cdh4.3.0.p0.4
- ○ None

Note: Solr is supported only on CDH 4.3 or later deployments.

IMPALA
- ◉ IMPALA-1.1.1-1.p0.17
- ○ None

Note: Impala is supported only on CDH 4.1 or later deployments.

CDH
- ◉ CDH-4.4.0-1.cdh4.4.0.p0.39

EC2 Management Console | Cloudera Manager Walkth | Amazon EC2 Instance Con | Global Knowledge Trainin | Cluster Installation - Clou

ec2-54-212-156-230.us-west-2.compute.amazonaws.com:7180/cmf/express-wizard/wizard#step=hostCredentialsStep

Cluster Installation

Provide SSH login credentials.

Root access to your hosts is required to install the Cloudera packages. This installer will connect to your hosts via SSH and log in either directly as root or as another user with password-less sudo/pbrun privileges to become root.

Login to all hosts as: ○ root
◉ Another User:
ubuntu (password-less sudo/pbrun privileges)

Use ubuntu as Login host

You may connect via password or public-key authentication for the user selected above.

Authentication Method: ○ All hosts accept same password
◉ All hosts accept same private key

Private Key File: Choose File hadoop.pem

Use your keypair file.

Enter Passphrase:

Confirm Passphrase:

SSH Port: 22

Number of simultaneous installations: 20 (Running a large number of installations at once can consume large amounts of network bandwidth and other system resources)

EC2 Management Console | Cloudera Manager Walkth | Amazon EC2 Instance Con | Global Knowledge Trainin | Cluster Installation - Clou

ec2-54-212-156-230.us-west-2.compute.amazonaws.com:7180/cmf/express-wizard/wizard#step=installStep

10 of 11 host(s) completed successfully.

Installation failed on 1 host(s). Uninstall Failed Hosts Retry Failed Hosts

Hostname	IP Address	Progress	Status	
ip-10-224-10-204.us-west-2.compute.internal	10.224.10.204		Installation completed successfully.	Details
ip-10-224-21-158.us-west-2.compute.internal	10.224.21.158		Installation completed successfully.	Details
ip-10-224-35-42.us-west-2.compute.internal	10.224.35.42		Installation completed successfully.	Details
ip-10-224-50-174.us-west-2.compute.internal	10.224.50.174		Installation completed successfully.	Details
ip-10-224-50-211.us-west-2.compute.internal	10.224.50.211		Installation completed successfully.	Details
ip-10-225-142-36.us-west-2.compute.internal	10.225.142.36		Installation completed successfully.	Details
ip-10-226-158-189.us-west-2.compute.internal	10.226.158.189		Installation completed successfully.	Details
ip-10-249-37-238.us-west-2.compute.internal	10.249.37.238		Installation completed successfully.	Details
ip-10-249-78-137.us-west-2.compute.internal	10.249.78.137		Installation completed successfully.	Details
ip-10-250-11-75.us-west-2.compute.internal	10.250.11.75		Installation completed successfully.	Details

EC2 Management Console | Cloudera Manager Walkth | Amazon EC2 Instance Con | Global Knowledge Trainin | Choose the CDH4 services

ec2-54-212-156-230.us-west-2.compute.amazonaws.com:7180/cmf/express-wizard/services?clusterId=1

cloudera manager

Support admin

Choose the CDH4 services that you want to install on your cluster.

Choose a combination of services to install.

- (•) **Core Hadoop**
  HDFS, MapReduce, ZooKeeper, Oozie, Hive, and Hue
- ( ) **Core with Real-Time Delivery**
  HDFS, MapReduce, ZooKeeper, HBase, Oozie, Hive, and Hue
- ( ) **Core with Real-Time Query**
  HDFS, MapReduce, ZooKeeper, Impala, Oozie, Hive, and Hue
- ( ) **All Services**
  HDFS, MapReduce, ZooKeeper, HBase, Impala, Oozie, Hive, Hue and Sqoop.
- ( ) **Custom Services**
  Choose your own services. Services required by chosen services must also be selected. Note that Flume, Solr and Keystore Indexer services can be added after your initial cluster has been set up.

This wizard will also install the **Cloudera Management Services**. These are a set of components that enable monitoring, reporting, events, and alerts; these components require databases to store information, which will be configured on the next page.

[ ] Include Cloudera Navigator

EC2 Management Console | Cloudera Manager Walkth | Amazon EC2 Instance Con | Global Knowledge Trainin | Database Setup - Clouder

ec2-54-212-156-230.us-west-2.compute.amazonaws.com:7180/cmf/express-wizard/autoconfig

cloudera manager

Support admin

Database Setup

On this page you configure and test database connections. If using custom databases, create the databases first according to the **Installing and Configuring an External Database** section of the Installation Guide.

When using the Embedded Database, passwords are auto generated. Please copy them down.

- (•) **Use Embedded Database**
- ( ) **Use Custom Databases**

Hive

Database Host Name:	Database Type:	Database Name :	Username:	Password:
ip-10-225-142-36.us-west-2.compute.intern	PostgreSQL	hive	hive	IITS5VpTqx

Service Monitor

Currently assigned to run on **ip-10-225-142-36.us-west-2.compute.internal**.

Database Host Name:	Database Type:	Database Name :	Username:	Password:
ip-10-225-142-36.us-west-2.compute.intern	PostgreSQL	smon	smon	Bd6gmjDv45

Activity Monitor

Test Connection

EC2 Management Console | Cloudera Manager Walkthr | Amazon EC2 Instance Con | Global Knowledge Traini | Starting your cluster servi

ec2-54-212-156-230.us-west-2.compute.amazonaws.com:7180/cmf/express-wizard/firstrunprogress?id=19

Completed 3 of 17 steps.

- ✓ Waiting for ZooKeeper Service to initialize
  Finished waiting
- ✓ Starting ZooKeeper Service
  Service started successfully.
- ✓ Checking if the name directories of the NameNode are empty. Formatting HDFS only if empty.
  Sucessfully formatted NameNode.
- **Starting HDFS Service**
- Creating HDFS /tmp directory
- Starting MapReduce Service
- Creating Hive Metastore Database
- Creating Hive Metastore Database Tables
- Creating Hive user directory
- Creating Hive warehouse directory
- Starting Hive Service
- Creating Oozie database
- Installing Oozie ShareLib in HDFS

EC2 Management Console | Cloudera Manager Walkth | Amazon EC2 Instance Con | Global Knowledge Training | Congratulations! - Cloude

ec2-54-212-156-230.us-west-2.compute.amazonaws.com:7180/cmf/express-wizard/finish

cloudera manager

Support admin

Congratulations!

The Hadoop services are installed, configured, and running on your cluster.

EC2 Management Console | Cloudera Manager Walkth | Amazon EC2 Instance Con | Global Knowledge Trainin | Home - Cloudera Manager

ec2-54-212-156-230.us-west-2.compute.amazonaws.com:7180/cmf/home

cloudera manager

Search by Service, Support admin

Home Services Hosts Activities Diagnose Audits Charts Reports Administration

Status All Health Issues 1 All Configuration Issues 2 All Recent Commands

September 26 2013, 4:45:29 PM UTC

Status

Cluster 1 - CDH4

- Hosts
- hdfs1
- hive1
- hue1
- mapreduce1
- oozie1
- zookeeper1

Cloudera Management Services

- mgmt1

Charts

30m 1h 2h 6h 12h 1d

Cluster 1 - CDH4

Cluster CPU
percent
100
50
04:30 04:45

Cluster Disk IO
bytes / second
2b/s
1.5b/s
1b/s
0.50b/s
04:30 04:45

Cluster Network IO
bytes / second
1b/s
0.50b/s

HDFS IO
bytes / second
2b/s
1.5b/s
1b/s
0.50b/s

Voila…!!!! You are all set.

5. Using Amazon Elastic Mapreduce

Amazon EMR is a paid service with very easy setup for MapReduce.

Create a JAR file with your mapper and reducer and upload it to S3 bucket.

Amazon Web Services

Compute

EC2
Virtual Servers in the Cloud

Lambda PREVIEW
Run Code in Response to Events

Storage & Content Delivery

S3
Scalable Storage in the Cloud

Storage Gateway
Integrates On-Premises IT Environments with Cloud Storage

Glacier
Archive Storage in the Cloud

CloudFront
Global Content Delivery Network

Database

RDS
MySQL, Postgres, Oracle, SQL Server, and Amazon Aurora

DynamoDB
Predictable and Scalable NoSQL Data Store

ElastiCache
In-Memory Cache

Redshift
Managed Petabyte-Scale Data Warehouse

Administration & Security

Directory Service
Managed Directories in the Cloud

Identity & Access Management
Access Control and Key Management

Trusted Advisor
AWS Cloud Optimization Expert

CloudTrail
User Activity and Change Tracking

Config PREVIEW
Resource Configurations and Inventory

CloudWatch
Resource and Application Monitoring

Deployment & Management

Elastic Beanstalk
AWS Application Container

OpsWorks
DevOps Application Management Service

CloudFormation
Templated AWS Resource Creation

CodeDeploy
Automated Deployments

Analytics

EMR
Managed Hadoop Framework

Application Services

SQS
Message Queue Service

SWF
Workflow Service for Coordinating Application Components

AppStream
Low Latency Application Streaming

Elastic Transcoder
Easy-to-use Scalable Media Transcoding

SES
Email Sending Service

CloudSearch
Managed Search Service

Mobile Services

Cognito
User Identity and App Data Synchronization

Mobile Analytics
Understand App Usage Data at Scale

SNS
Push Notification Service

Enterprise Applications

WorkSpaces
Desktops in the Cloud

WorkDocs

AWS | Services | Edit | Spandan Brahmbhatt | N. Virginia | Support

Elastic MapReduce | Cluster List | EMR Help

Create cluster | Clone | Terminate

Filter: All clusters | Filter clusters ... | 5 clusters (all loaded)

			Name	ID	Status	Creation time (UTC-5)	Elapsed time	Normalized instance hours

Elastic MapReduce ⌄ Create Cluster

Configure sam

Cluster Configuration

Cluster name My cluster

Termination protection Yes
No

Prevents accidental termination of the clust
down the cluster, you must turn off terminat
protection. Learn more

Any Name you want

Logging Enabled

Copy the cluster's log files automatically to
more

Log folder S3 location

s3://aws-logs-235070480323-us-east-1/elasticmapreduce/

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

Debugging Enabled

Index logs to enable console debugging fur
(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
to the underlying EC2 instances. Learn more about tagging your Amazon EMR clusters.

Key	Value (optional)
Add a key to create a tag	

Software Configuration

Hadoop distribution Amazon

Use Amazon's Hadoop distribution. Learn more

AMI version

3.3.1

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	0.13.1
Pig	0.12.0
Hue	3.6.0

Additional applications Select an application

Configure and add

You can add additional Application from here

File System Configuration

The EMR File System (EMRFS) and the Hadoop Distributed File System (HDFS) are both installed on your EMR cluster. HDFS stores data on an EMR cluster, while EMRFS allows EMR clusters to store data on S3. You can enable server-side encryption and consistent view for EMRFS below, or use a bootstrap action to configure additional settings for EMRFS.

Hardware Configuration

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

Network vpc-35325b50 (172.31.0.0/16) (default)

Use a Virtual Private Cloud (VPC) to process sensitive dat
or connect to a private network. Create a VPC

EC2 Subnet No preference (random subnet)

Create a Subnet

Type	Name	EC2 instance type	Count	Request spot	Bid price
Master	Master instance group - 1	m3.xlarge	1		
Core	Core instance group - 2	m3.xlarge	2		
Task	Task instance group - 3	m3.xlarge	0		

Add task instance group

Specify Number of Instances

Security and Access

EC2 key pair Proceed without an EC2 key pair

Use an existing EC2 key pair to SSH into the master node
of the Amazon EMR cluster. 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 Select a bootstrap action

Configure and add

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 location	Arguments

Add step Select a step

Configure and add

Select a step to add.

Auto-terminate Yes

Automatically terminate cluster after the last step is completed.

No

Keep cluster running until you terminate it.

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 location	Arguments

Add step

- ✓ Select a step
- Streaming program
- Hive program
- Pig program
- Impala program
- Custom JAR

Auto-terminate

- No

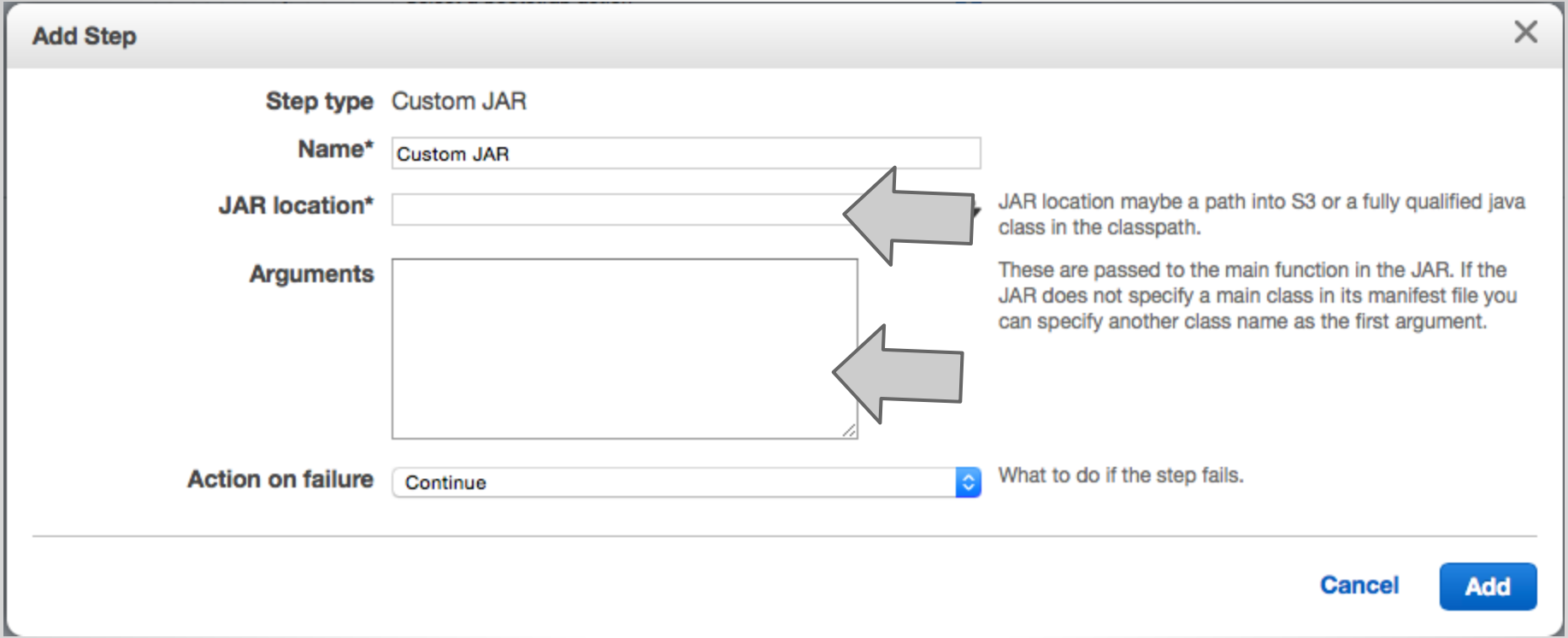
Automatically terminate cluster after the last step is completed.

Keep cluster running until you terminate it.

No EC2 key pair has been selected, so you will not be able t SSH to this cluster or connect to HUE (unless you are using a VPN). Learn how to create an EC2 Key Pair.

Cancel

Create cluste



Click on Create Cluster and sit back and enjoy your Coffee!!!