

# Big Data Analytics Using Hadoop Cluster On Amazon EMR

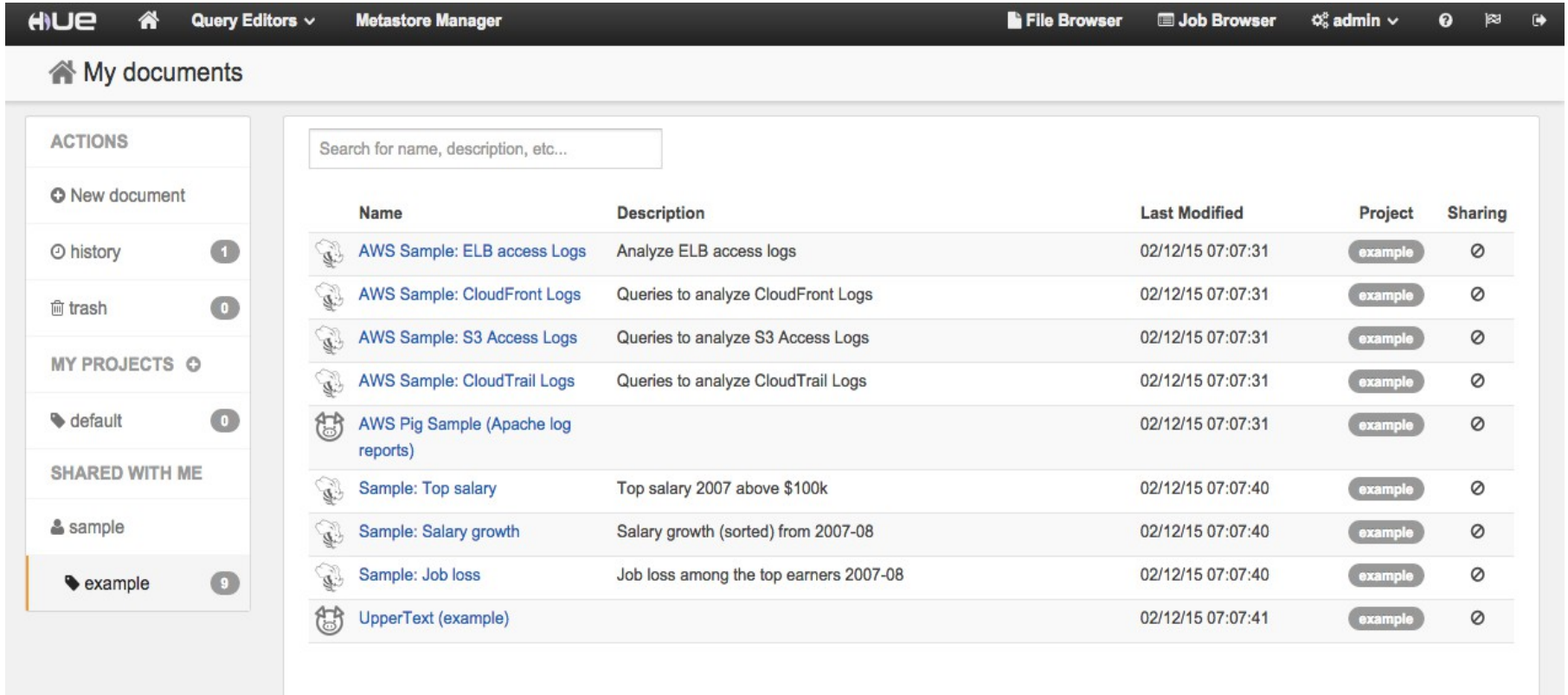
---

**February 2015**

Dr.Thanachart Numnonda  
IMC Institute  
thanachart@imcinstitute.com



















Modifiy from Original Version by Danairat T.  
Certified Java Programmer, TOGAF – Silver  
danairat@gmail.com

# Running this lab using Amazon EMR



**My documents**

Search for name, description, etc...

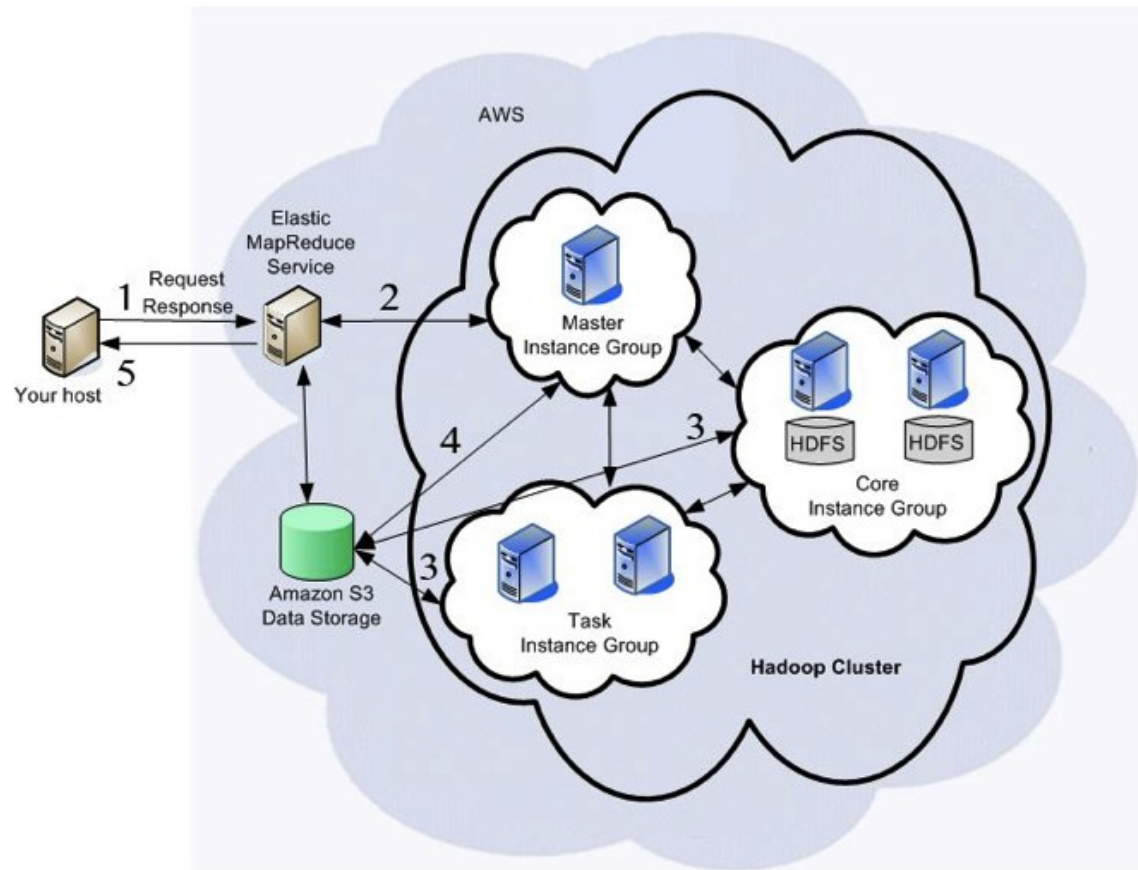
Name	Description	Last Modified	Project	Sharing
 <a href="#">AWS Sample: ELB access Logs</a>	Analyze ELB access logs	02/12/15 07:07:31	example	
 <a href="#">AWS Sample: CloudFront Logs</a>	Queries to analyze CloudFront Logs	02/12/15 07:07:31	example	
 <a href="#">AWS Sample: S3 Access Logs</a>	Queries to analyze S3 Access Logs	02/12/15 07:07:31	example	
 <a href="#">AWS Sample: CloudTrail Logs</a>	Queries to analyze CloudTrail Logs	02/12/15 07:07:31	example	
 <a href="#">AWS Pig Sample (Apache log reports)</a>		02/12/15 07:07:31	example	
 <a href="#">Sample: Top salary</a>	Top salary 2007 above \$100k	02/12/15 07:07:40	example	
 <a href="#">Sample: Salary growth</a>	Salary growth (sorted) from 2007-08	02/12/15 07:07:40	example	
 <a href="#">Sample: Job loss</a>	Job loss among the top earners 2007-08	02/12/15 07:07:40	example	
 <a href="#">UpperText (example)</a>		02/12/15 07:07:41	example	



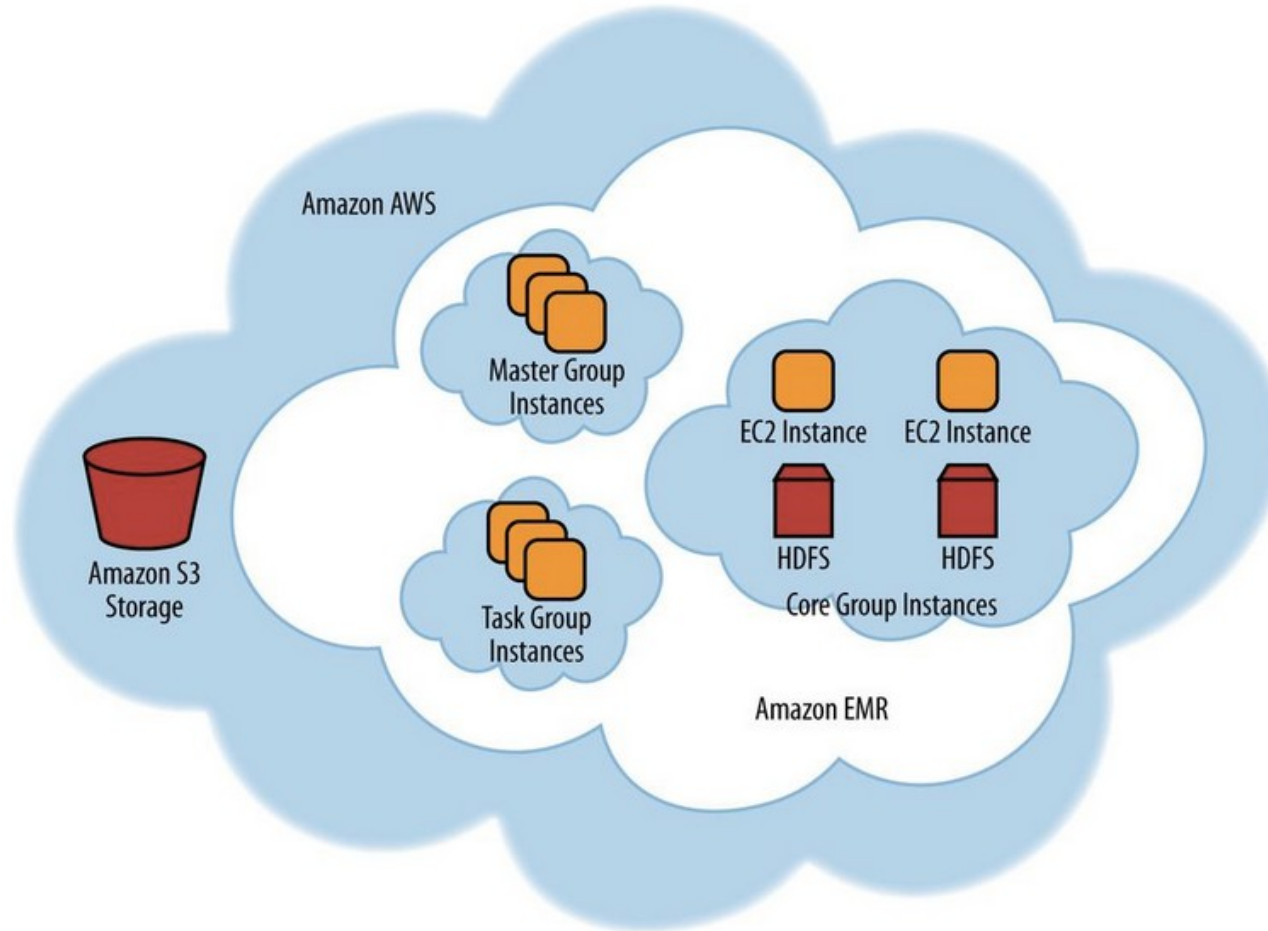
# Hands-On: Create an EMR cluster

---

# Architecture Overview of Amazon EMR



# Amazon EMR Cluster



# Creating an AWS account

[Sign Up](#)[My Account / Console](#) ▾[English](#) ▾[AWS Products & Solutions](#) ▾[Entire Site](#) ▾[Developers](#) ▾[Support](#) ▾

## AWS Free Tier Includes Windows

Receive 750 hours per month of Amazon EC2  
Microsoft Windows Server Micro instance usage.

[Try Windows Server on AWS »](#)[Get Started for Free »](#)

Launch virtual machines and apps in minutes.

# Signing up for the necessary services

- Simple Storage Service (S3)
- Elastic Compute Cloud (EC2)
- Elastic MapReduce (EMR)

Caution! This costs real money!

# Creating Amazon S3 bucket



Services ▾ Edit ▾

## Welcome to Amazon Simple Storage Service

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

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

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

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

[Create Bucket](#)

## Additional Information

[Getting Started Guide](#)

[Documentation](#)

[All S3 Resources](#)

[Forums](#)

## S3 at a glance

Create



Create a bucket in one of several

Add



Upload objects to your bucket. Amazon

Manage



Manage your data with Amazon S3's





# Create access key using Security Credentials in the AWS Management Console

## Access Credentials

There are three types of access credentials used to authenticate your requests to AWS services: (a) access keys, (b) X.509 certificates, and (c) key pairs. Each access credential type is explained below.

 **Access Keys**

 X.509 Certificates

 Key Pairs

Use access keys to make secure REST or Query protocol requests to any AWS service API. We create one for you when your account is created — see your access key below.

### Your Access Keys

Created	Access Key ID	Secret Access Key	Status
July 14, 2013		<a href="#">Show</a>	Active <a href="#">(Make Inactive)</a>

[Create a new Access Key](#)

For your protection, you should never share your secret access keys with anyone. In addition, industry best practice recommends frequent key rotation.

 [Learn more about Access Keys](#)

## Access Credentials

There are three types of access credentials used to authenticate your requests to AWS services: (a) access keys, (b) X.509 certificates, and (c) key pairs. Each access credential type is explained below.

 **Access Keys**
 X.509 Certificates
  Key Pairs

Use access keys to make secure REST or Query protocol requests to any AWS service API. We create one for you when your account is created — see your access key below.

**Your Access Keys**

Created	Access Key ID	Secret Access Key	Status
July 14, 2013		<a href="#">Show</a>	Active <a href="#">(Make Inactive)</a>

[Create a new Access Key](#)

For your protection, you should never share your secret access keys with anyone. In addition, industry best practice recommends frequent key rotation.

 [Learn more about Access Keys](#)


# Select EMR

## 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 Service

### Networking

-  **VPC**  
Isolated Cloud Resources




### 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**  
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
-  **Kinesis**  
Real-time Processing of Streaming Big Data
-  **Data Pipeline**  
Orchestration for Data-driven Workflows

### 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**  
Secure Enterprise Storage and Sharing Service
-  **WorkMail** PREVIEW  
Secure Email and Calendaring Service

# Creating a cluster in EMR



AWS

Services

Edit

IMC Institute

Oregon

Support

Elastic MapReduce

Cluster List

EMR Help

Create cluster

View details

Clone



Terminate

Filter:

All clusters

Filter clusters ...

69 clusters (all loaded)

	Name	ID	Status	Creation time (UTC+7)	Elapsed time	Normalized instance hours
<input type="checkbox"/>	 IMC Cluster	j-VUANGWGU9T2R	Waiting	2015-02-12 22:00 (UTC+7)	52 minutes	24
<input type="checkbox"/>	<a href="#">ChitsuthaCluster2</a>	j-3V8AL7QNS7VLA	Terminated User request	2015-02-12 10:11 (UTC+7)	25 minutes	6
<input type="checkbox"/>	<a href="#">guest7_hadoop</a>	j-1QWGV1MHJUBS8	Terminated All steps completed	2015-02-12 10:11 (UTC+7)	11 minutes	6
<input type="checkbox"/>	<a href="#">Kachapak Hadoop</a>	j-26R59GXKZ0SD6	Terminated User request	2015-02-12 10:11 (UTC+7)	27 minutes	6
<input type="checkbox"/>	<a href="#">guest01_cluster01</a>	j-98FNCDWNXW3L	Terminated All steps completed	2015-02-12 10:10 (UTC+7)	11 minutes	6
<input type="checkbox"/>	 <a href="#">Kachapak Hadoop</a>	j-155QW6L1RIP1T	Terminated with errors Validation error	2015-02-12 10:09 (UTC+7)	20 seconds	0
<input type="checkbox"/>	<a href="#">guest8 cluster</a>	j-1QFKYNQDZSHMS	Terminated All steps completed	2015-02-12 10:06 (UTC+7)	11 minutes	6
<input type="checkbox"/>	<a href="#">guest10 Hadoop</a>	j-2N0RIVGO4XDHE	Terminated All steps completed	2015-02-12 10:05 (UTC+7)	10 minutes	6


© 2008 - 2015, Amazon Web Services, Inc. or its affiliates. All rights reserved.

[Privacy Policy](#)
[Terms of Use](#)

Feedback

# Creating a cluster in EMR (cont.)

Name the cluster and also specify Log folder

 **AWS** ▾ **Services** ▾ **Edit** ▾ **IMC Institute** ▾ **Oregon** ▾ **Support** ▾

**Elastic MapReduce** ▾ **Create Cluster** **EMR Help**

**Cluster Configuration**

**Cluster name**

IMC Cluster

**Termination protection**

☒ Yes  
☐ No

**Logging**

☒ Enabled

**Log folder S3 location**

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

**Debugging**

☒ Enabled


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

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

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

**Configure sample application**

**Tags**

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

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

# Creating a cluster in EMR (cont.)

Leave the Software Configuration as default

## Software Configuration

Hadoop distribution ☒ Amazon

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

AMI version

3.3.2



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



Configure and add



# Creating a cluster in EMR (cont.)

Leave the Hardware Configuration as default

Choose an existing 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** vpc-cd510ca5 (172.31.0.0/16) (default)

Use a Virtual Private Cloud (VPC) to process sensitive data 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	<input type="checkbox"/>				?
Core	Core instance group - 2	m3.xlarge	2	<input type="checkbox"/>				?
Task	Task instance group - 3	m3.xlarge	0	<input type="checkbox"/>				✕ ?

Add task instance group

## Security and Access

**EC2 key pair** imckey

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](#)

# Creating a cluster in EMR (cont.)

Leave the others as default  
Select Create Cluster

Configure and add

## Steps

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

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

Add step

Select a 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






# EMR Cluster Details

**Note on the Master public DNS:**

**To see the details on how to connect to the Master Node using SSH click at SSH**



AWS
Services
Edit
IMC Institute
Oregon
Support

Elastic MapReduce
Cluster List
Cluster Details
EMR Help

Add step
Resize
Clone
Terminate

Cluster: IMC Cluster Waiting Waiting after step completed

**Connections:** [Hue, Resource Manager ... \(View All\)](#)  
**Master public DNS:** [ec2-54-201-231-114.us-west-2.compute.amazonaws.com](#) [SSH](#)  
**Tags:** -- [View All / Edit](#)

Summary	Configuration Details	Network and Hardware	Security and Access
<b>ID:</b> j-VUANGWGU9T2R <b>Creation date:</b> 2015-02-12 22:00 (UTC+7) <b>Elapsed time:</b> 1 hour, 3 minutes <b>Auto-terminate:</b> No <b>Termination protection:</b> On <a href="#">Change</a>	<b>AMI version:</b> 3.3.2 <b>Hadoop distribution:</b> Amazon 2.4.0 <b>Applications:</b> Hive 0.13.1, Pig 0.12.0, Hue <b>Log URI:</b> s3://imcbucket/logs/  <b>EMRFS consistent view:</b> Disabled	<b>Availability zone:</b> us-west-2b <b>Subnet ID:</b> subnet-c0510ca8 <b>Master:</b> <span>Running</span> 1 m3.xlarge <b>Core:</b> <span>Running</span> 2 m3.xlarge <b>Task:</b> --	<b>Key name:</b> imckey <b>EC2 instance profile:</b> -- <b>EMR role:</b> -- <b>Visible to all users:</b> <a href="#">Change</a> <b>Security groups for (ElasticMapReduce-Master):</b> <a href="#">sg-46638029</a> <b>Security groups for (ElasticMapReduce-slave):</b> <a href="#">sg-4563802a</a> <b>Core &amp; Task:</b>

# SSH Instruction

## SSH



### Connect to the Master Node Using SSH

You can connect to the Amazon EMR master node using SSH to run interactive queries, examine log files, submit Linux commands, and so on. [Learn more.](#)

Windows

Mac / Linux

1. Open a terminal window. On Mac OS X, choose Applications > Utilities > Terminal. On other Linux distributions, terminal is typically found at Applications > Accessories > Terminal.
2. To establish a connection to the master node, type the following command. Replace `~/imckey.pem` with the location and filename of the private key file (.pem) used to launch the cluster.

```
ssh hadoop@ec2-54-201-231-114.us-west-2.compute.amazonaws.com -i ~/imckey.pem
```

3. Type yes to dismiss the security warning.

Close

# Connect to the Master Node

```
THANACHARTs-Air:elastic-mapreduce-cli THANACHART$ ssh hadoop@ec2-54-201-231-114.us-west-2.compute.amazonaws.com -i imckey.pem
Last login: Thu Feb 12 16:00:12 2015
```

```

  _ |  _ |  _ |
 _ |  (  _ |  /
 _ | \ _ |  _ |

```

Amazon Linux AMI

```
https://aws.amazon.com/amazon-linux-ami/2014.09-release-notes/
```

```
2 package(s) needed for security, out of 13 available
```

```
Run "sudo yum update" to apply all updates.
```

```
-bash: warning: setlocale: LC_CTYPE: cannot change locale (UTF-8): No such file or directory
```

Welcome to Amazon Elastic MapReduce running Hadoop and Amazon Linux.

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

The Hadoop UI can be accessed via the following commands:

ResourceManager	lynx http://ip-172-31-17-48.us-west-2.compute.internal:9026/
NameNode	lynx http://ip-172-31-17-48.us-west-2.compute.internal:9101/

---

```
[hadoop@ip-172-31-17-48 ~]$ ls
```

# Web Interface Host on EMR Cluster

Name of Interface	URI
Hadoop version 2.x	
Hadoop ResourceManager	<a href="http://master-public-dns-name:9026/">http://master-public-dns-name:9026/</a>
Hadoop HDFS NameNode	<a href="http://master-public-dns-name:9101/">http://master-public-dns-name:9101/</a>
Ganglia Metrics Reports	<a href="http://master-public-dns-name/ganglia/">http://master-public-dns-name/ganglia/</a>
HBase Interface	<a href="http://master-public-dns-name:60010/master-status">http://master-public-dns-name:60010/master-status</a>
Hue Web Application	<a href="http://master-public-dns-name:8888/">http://master-public-dns-name:8888/</a>
Impala Statestore	<a href="http://master-public-dns-name:25000">http://master-public-dns-name:25000</a>
Impalad	<a href="http://master-public-dns-name:25010">http://master-public-dns-name:25010</a>
Impala Catalog	<a href="http://master-public-dns-name:25020">http://master-public-dns-name:25020</a>
Hadoop version 1.x	
Hadoop MapReduce JobTracker	<a href="http://master-public-dns-name:9100/">http://master-public-dns-name:9100/</a>
Hadoop HDFS NameNode	<a href="http://master-public-dns-name:9101/">http://master-public-dns-name:9101/</a>
Ganglia Metrics Reports	<a href="http://master-public-dns-name/ganglia/">http://master-public-dns-name/ganglia/</a>
HBase Interface	<a href="http://master-public-dns-name:60010/master-status">http://master-public-dns-name:60010/master-status</a>

# Launch the Hue Web Interface

- Set Up an SSH Tunnel to the Master Node
  - See instruction at
  - <http://docs.aws.amazon.com/ElasticMapReduce/latest/DeveloperGuide/emr-ssh-tunnel.html>
- Configure Proxy Settings to View Websites
  - See instruction at
  - <http://docs.aws.amazon.com/ElasticMapReduce/latest/DeveloperGuide/emr-connect-master-node-proxy.html>

# Launch the Hue Web Interface (Cont.)

- <http://master-public-dns-name:8888/>
- Create your own Hue account

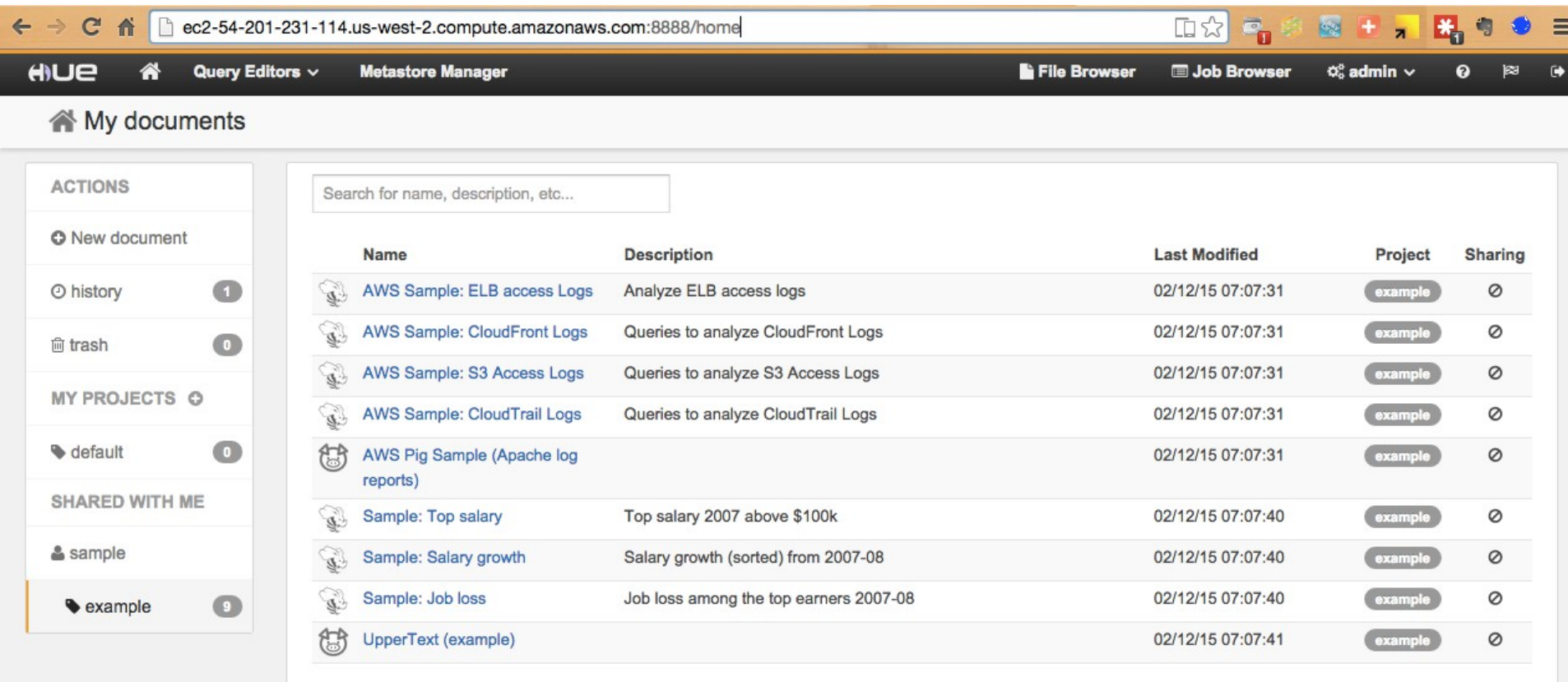


Create your Hue account










Since this is your first time logging in, pick any username and password. Be sure to remember these, as **they will become your Hue superuser credentials**.  
The password must be at least 8 characters long, and must contain both uppercase and lowercase letters, at least one number, and at least one special character.

Create account

# Launch the Hue Web Interface (Cont.)



The screenshot shows the Hue web interface running on an Amazon EC2 instance. The browser address bar displays the URL: `ec2-54-201-231-114.us-west-2.compute.amazonaws.com:8888/home`. The interface includes a top navigation bar with the Hue logo, a home icon, and links to Query Editors, Metastore Manager, File Browser, Job Browser, and an admin dropdown. Below the navigation bar is a sidebar with 'My documents' and a list of actions and projects. The main content area displays a table of documents with columns for Name, Description, Last Modified, Project, and Sharing.

Name	Description	Last Modified	Project	Sharing
 <a href="#">AWS Sample: ELB access Logs</a>	Analyze ELB access logs	02/12/15 07:07:31	<a href="#">example</a>	<a href="#">⊗</a>
 <a href="#">AWS Sample: CloudFront Logs</a>	Queries to analyze CloudFront Logs	02/12/15 07:07:31	<a href="#">example</a>	<a href="#">⊗</a>
 <a href="#">AWS Sample: S3 Access Logs</a>	Queries to analyze S3 Access Logs	02/12/15 07:07:31	<a href="#">example</a>	<a href="#">⊗</a>
 <a href="#">AWS Sample: CloudTrail Logs</a>	Queries to analyze CloudTrail Logs	02/12/15 07:07:31	<a href="#">example</a>	<a href="#">⊗</a>
 <a href="#">AWS Pig Sample (Apache log reports)</a>		02/12/15 07:07:31	<a href="#">example</a>	<a href="#">⊗</a>
 <a href="#">Sample: Top salary</a>	Top salary 2007 above \$100k	02/12/15 07:07:40	<a href="#">example</a>	<a href="#">⊗</a>
 <a href="#">Sample: Salary growth</a>	Salary growth (sorted) from 2007-08	02/12/15 07:07:40	<a href="#">example</a>	<a href="#">⊗</a>
 <a href="#">Sample: Job loss</a>	Job loss among the top earners 2007-08	02/12/15 07:07:40	<a href="#">example</a>	<a href="#">⊗</a>
 <a href="#">UpperText (example)</a>		02/12/15 07:07:41	<a href="#">example</a>	<a href="#">⊗</a>

# Hands-On: Importing/Exporting Data to HDFS

---



# Importing Data to Hadoop

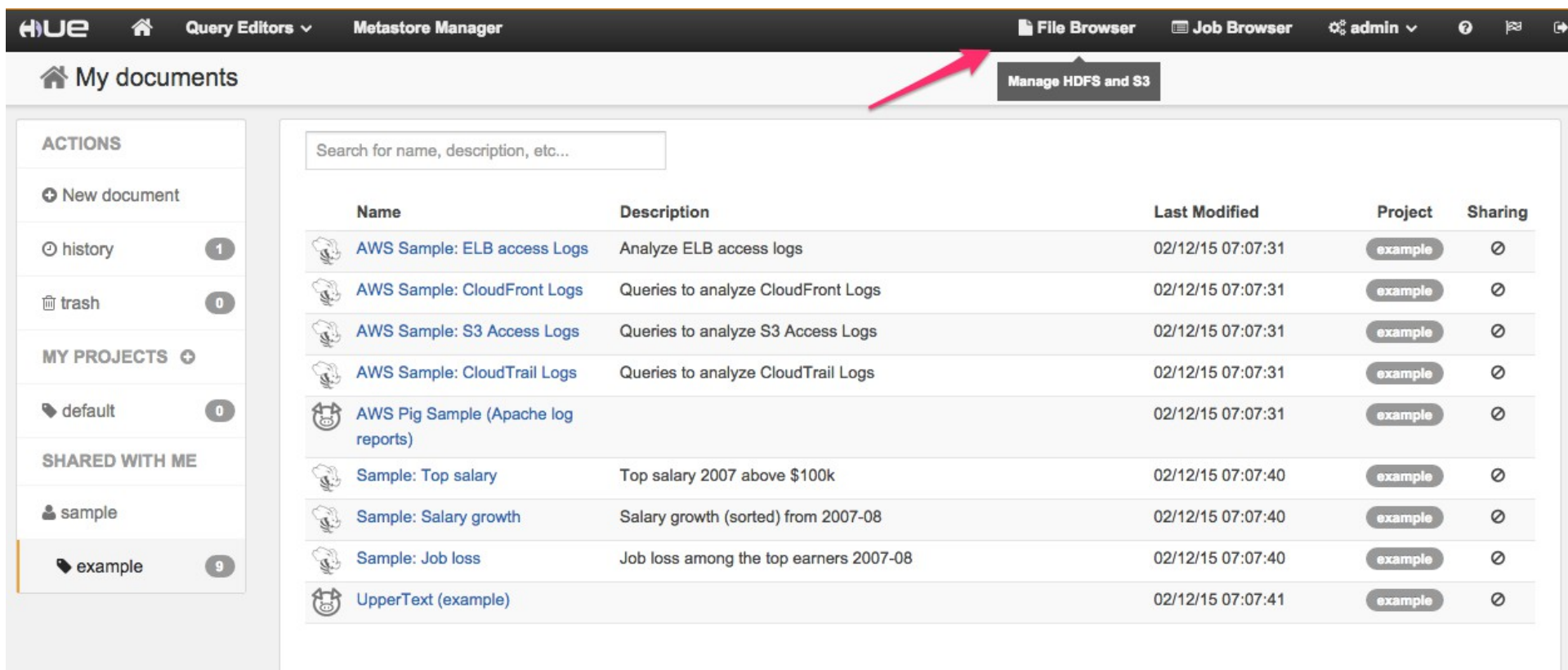
## Download War and Peace Full Text

[www.gutenberg.org/ebooks/2600](http://www.gutenberg.org/ebooks/2600)










The screenshot shows a web browser window with the address bar displaying [www.gutenberg.org/ebooks/2600](http://www.gutenberg.org/ebooks/2600). The page title is "War and Peace by graf Leo Tolstoy". On the left, there is a placeholder for a book cover with the text "No cover available". On the right, there are two tabs: "Download" (selected) and "Bibrec". Below the tabs, the section "Download This eBook" contains a table with download options.

Format ?	Size			
<a href="#">Read this book online: HTML</a>	3.6 MB			
<a href="#">EPUB (no images)</a>	1.3 MB			
<a href="#">Kindle (no images)</a>	5.1 MB			
<a href="#">Plain Text UTF-8</a>	3.1 MB			
<a href="#">More Files...</a>				

# Review file in Hadoop HDFS using File Browse

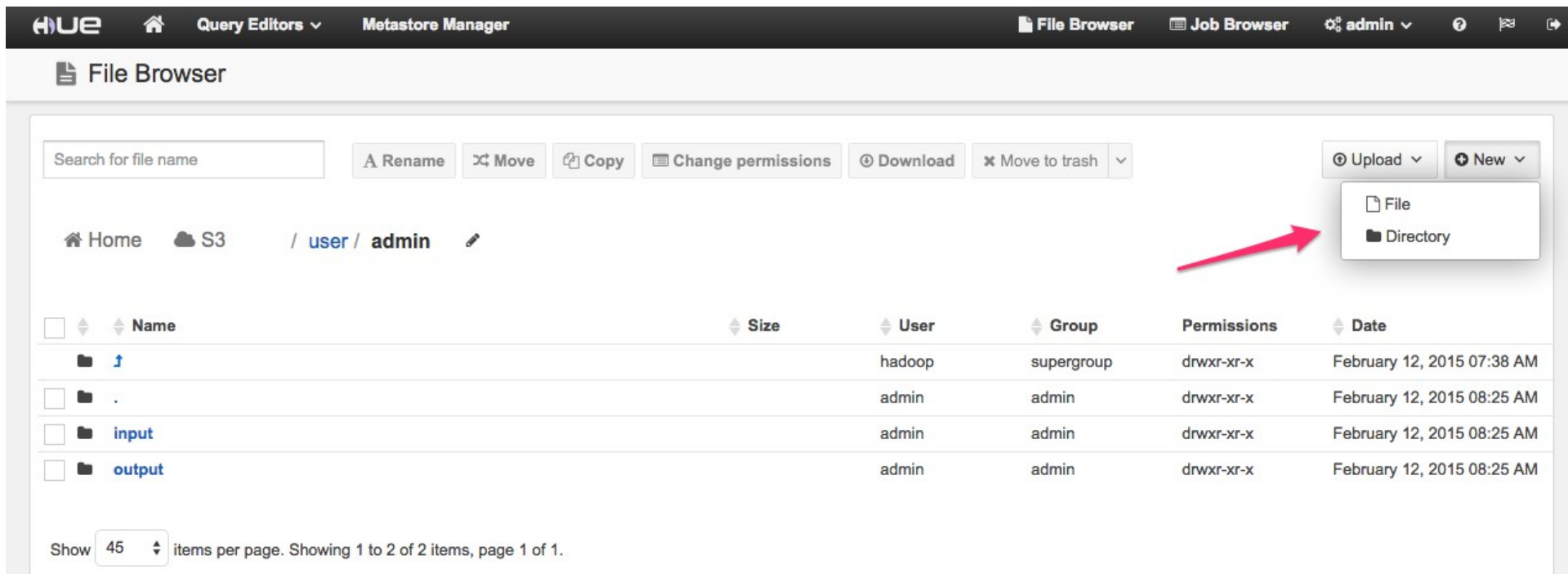


The screenshot shows the Hue web interface. The top navigation bar includes 'HUE', 'Query Editors', 'Metastore Manager', 'File Browser', 'Job Browser', and 'admin'. The 'File Browser' tab is active, and a red arrow points to the 'Manage HDFS and S3' button. The left sidebar shows 'My documents' with sections for 'ACTIONS' (New document, history, trash), 'MY PROJECTS' (default), and 'SHARED WITH ME' (sample, example). The main content area has a search bar and a table of files.

Name	Description	Last Modified	Project	Sharing
 <a href="#">AWS Sample: ELB access Logs</a>	Analyze ELB access logs	02/12/15 07:07:31	example	⊗
 <a href="#">AWS Sample: CloudFront Logs</a>	Queries to analyze CloudFront Logs	02/12/15 07:07:31	example	⊗
 <a href="#">AWS Sample: S3 Access Logs</a>	Queries to analyze S3 Access Logs	02/12/15 07:07:31	example	⊗
 <a href="#">AWS Sample: CloudTrail Logs</a>	Queries to analyze CloudTrail Logs	02/12/15 07:07:31	example	⊗
 <a href="#">AWS Pig Sample (Apache log reports)</a>		02/12/15 07:07:31	example	⊗
 <a href="#">Sample: Top salary</a>	Top salary 2007 above \$100k	02/12/15 07:07:40	example	⊗
 <a href="#">Sample: Salary growth</a>	Salary growth (sorted) from 2007-08	02/12/15 07:07:40	example	⊗
 <a href="#">Sample: Job loss</a>	Job loss among the top earners 2007-08	02/12/15 07:07:40	example	⊗
 <a href="#">UpperText (example)</a>		02/12/15 07:07:41	example	⊗

# Create new directory

Create two new directory name: input and output



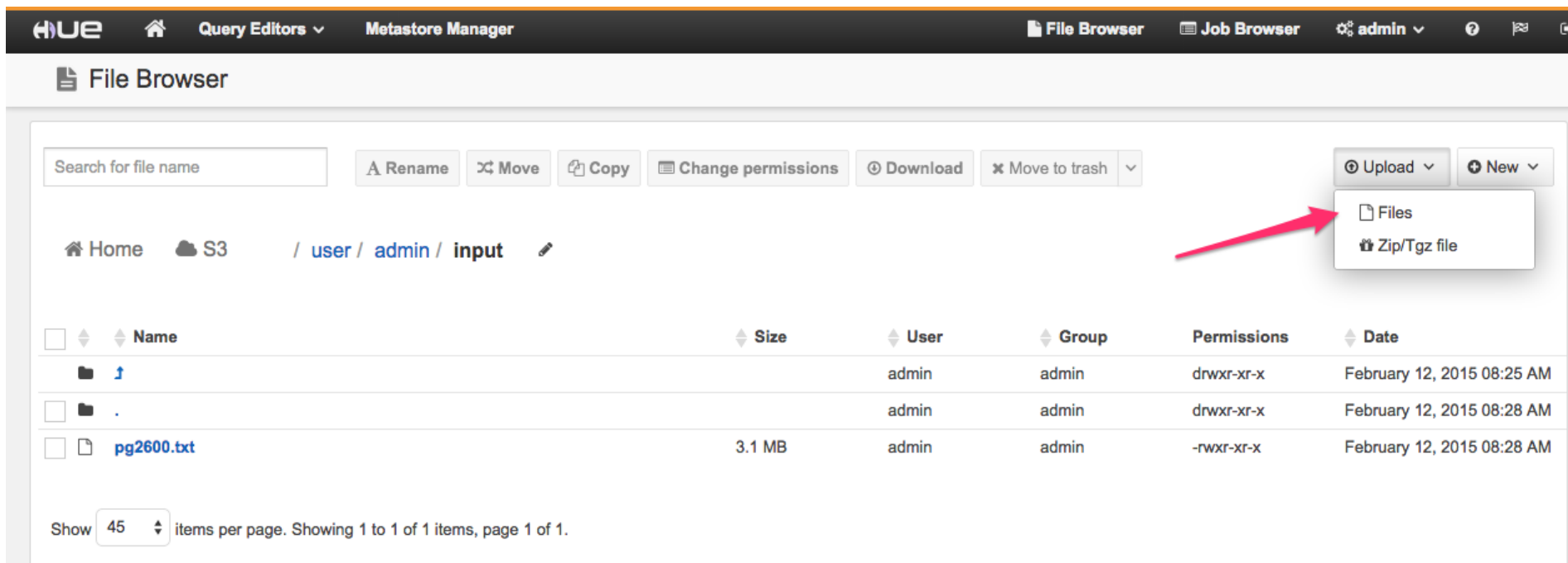
The screenshot shows the HUE File Browser interface. At the top, there's a navigation bar with 'HUE', 'Query Editors', 'Metastore Manager', 'File Browser', 'Job Browser', and a user profile 'admin'. Below this, the 'File Browser' section is active. It includes a search bar and action buttons: 'Rename', 'Move', 'Copy', 'Change permissions', 'Download', and 'Move to trash'. A 'New' button is highlighted with a red arrow, and its dropdown menu is open, showing 'File' and 'Directory' options. The 'Directory' option is selected. Below the menu, a table lists the current directory contents:

	Name	Size	User	Group	Permissions	Date
<input type="checkbox"/>	↑		hadoop	supergroup	drwxr-xr-x	February 12, 2015 07:38 AM
<input type="checkbox"/>	.		admin	admin	drwxr-xr-x	February 12, 2015 08:25 AM
<input type="checkbox"/>	input		admin	admin	drwxr-xr-x	February 12, 2015 08:25 AM
<input type="checkbox"/>	output		admin	admin	drwxr-xr-x	February 12, 2015 08:25 AM

At the bottom, it says 'Show 45 items per page. Showing 1 to 2 of 2 items, page 1 of 1.'

# Upload Files

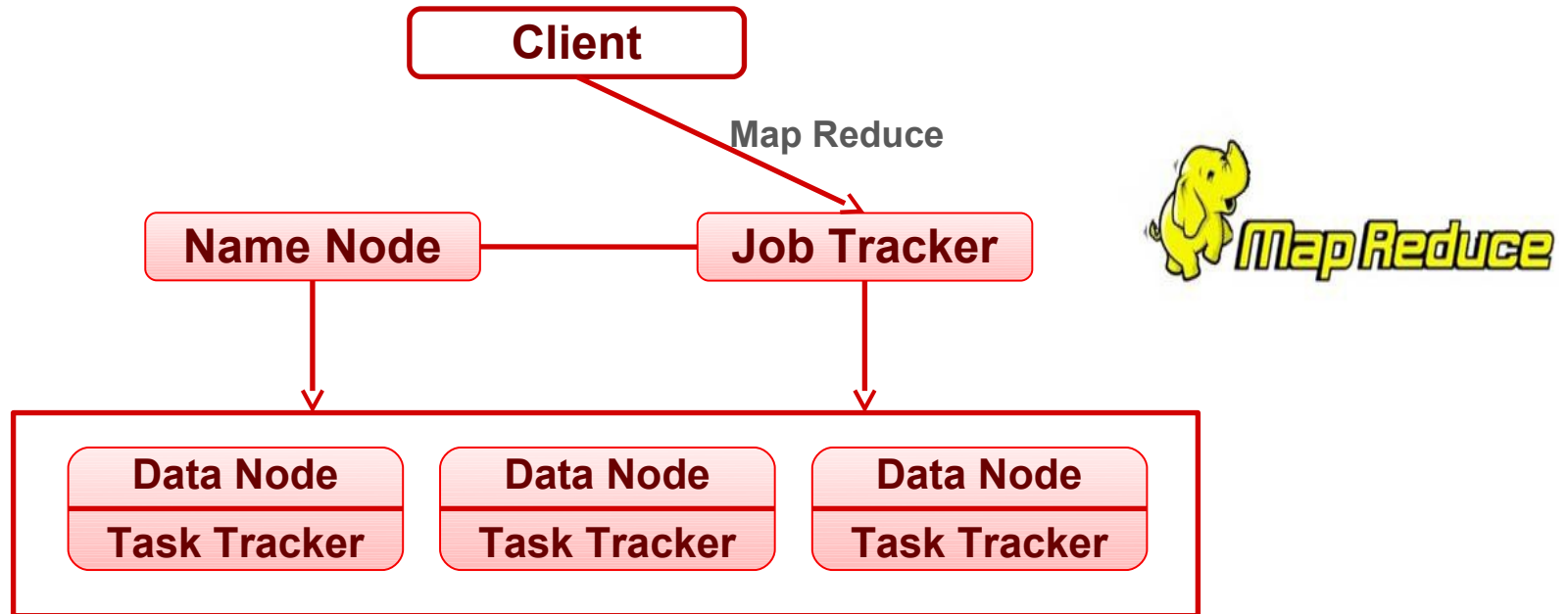
Upload file: **pg2600.txt** into **input** directory



The screenshot shows the HUE File Browser interface. The top navigation bar includes 'HUE', 'Query Editors', 'Metastore Manager', 'File Browser', 'Job Browser', and a user profile 'admin'. The main header is 'File Browser'. Below it, there's a search bar and a row of action buttons: 'Rename', 'Move', 'Copy', 'Change permissions', 'Download', and 'Move to trash'. On the right, there are 'Upload' and 'New' buttons. A red arrow points to the 'Upload' button, which has a dropdown menu open showing 'Files' and 'Zip/Tgz file'. The breadcrumb path is 'Home / S3 / user / admin / input'. Below this is a table listing files in the 'input' directory.

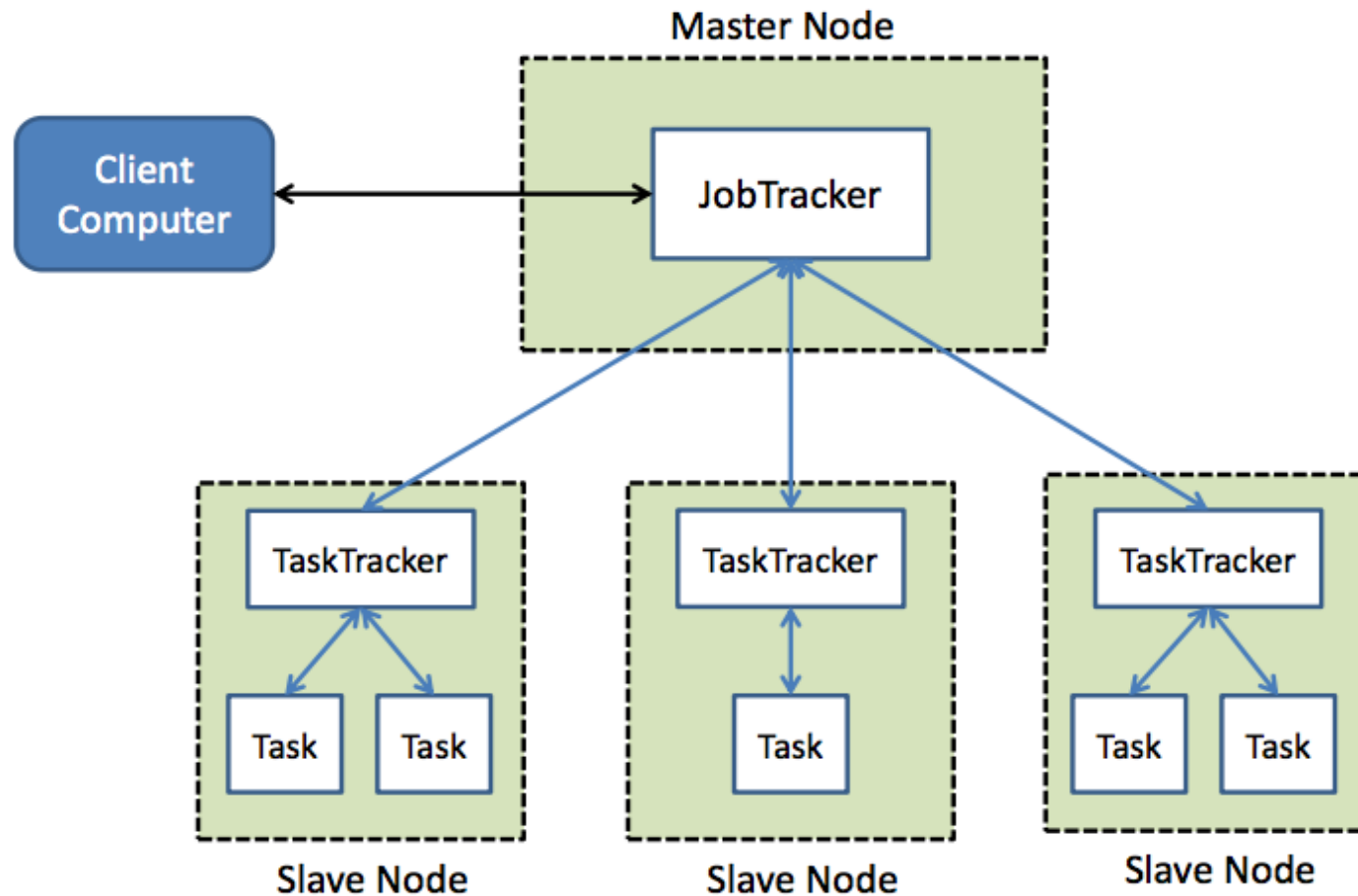
<input type="checkbox"/>	Name	Size	User	Group	Permissions	Date
<input type="checkbox"/>	<a href="#">↑</a>		admin	admin	drwxr-xr-x	February 12, 2015 08:25 AM
<input type="checkbox"/>	<a href="#">.</a>		admin	admin	drwxr-xr-x	February 12, 2015 08:28 AM
<input type="checkbox"/>	<a href="#">pg2600.txt</a>	3.1 MB	admin	admin	-rwxr-xr-x	February 12, 2015 08:28 AM

At the bottom, it says 'Show 45 items per page. Showing 1 to 1 of 1 items, page 1 of 1.'



# Lecture: Understanding Map Reduce Processing

# High Level Architecture of MapReduce



# Hands-On: Writing you own Map Reduce Program

---

# Wordcount (HelloWord in Hadoop)

```
1.  package org.myorg;
2.
3.  import java.io.IOException;
4.  import java.util.*;
5.
6.  import org.apache.hadoop.fs.Path;
7.  import org.apache.hadoop.conf.*;
8.  import org.apache.hadoop.io.*;
9.  import org.apache.hadoop.mapred.*;
10. import org.apache.hadoop.util.*;
11.
12. public class WordCount {
13.
14.     public static class Map extends MapReduceBase implements Mapper<LongWritable, Text, Text,
15.     IntWritable> {
16.         private final static IntWritable one = new IntWritable(1);
17.         private Text word = new Text();
18.
19.         public void map(LongWritable key, Text value, OutputCollector<Text, IntWritable> output,
20.         Reporter reporter) throws IOException {
21.             String line = value.toString();
22.             StringTokenizer tokenizer = new StringTokenizer(line);
23.             while (tokenizer.hasMoreTokens()) {
24.                 word.set(tokenizer.nextToken());
25.                 output.collect(word, one);
26.             }
27.         }
28.     }
29. }
```



# Wordcount (HelloWord in Hadoop)

```
27.  
28.    public static class Reduce extends MapReduceBase implements Reducer<Text, IntWritable, Text,  
29.    IntWritable> {  
30.        public void reduce(Text key, Iterator<IntWritable> values, OutputCollector<Text, IntWritable>  
31.        output, Reporter reporter) throws IOException {  
32.            int sum = 0;  
33.            while (values.hasNext()) {  
34.                sum += values.next().get();  
35.            }  
36.            output.collect(key, new IntWritable(sum));  
37.        }  
    }
```

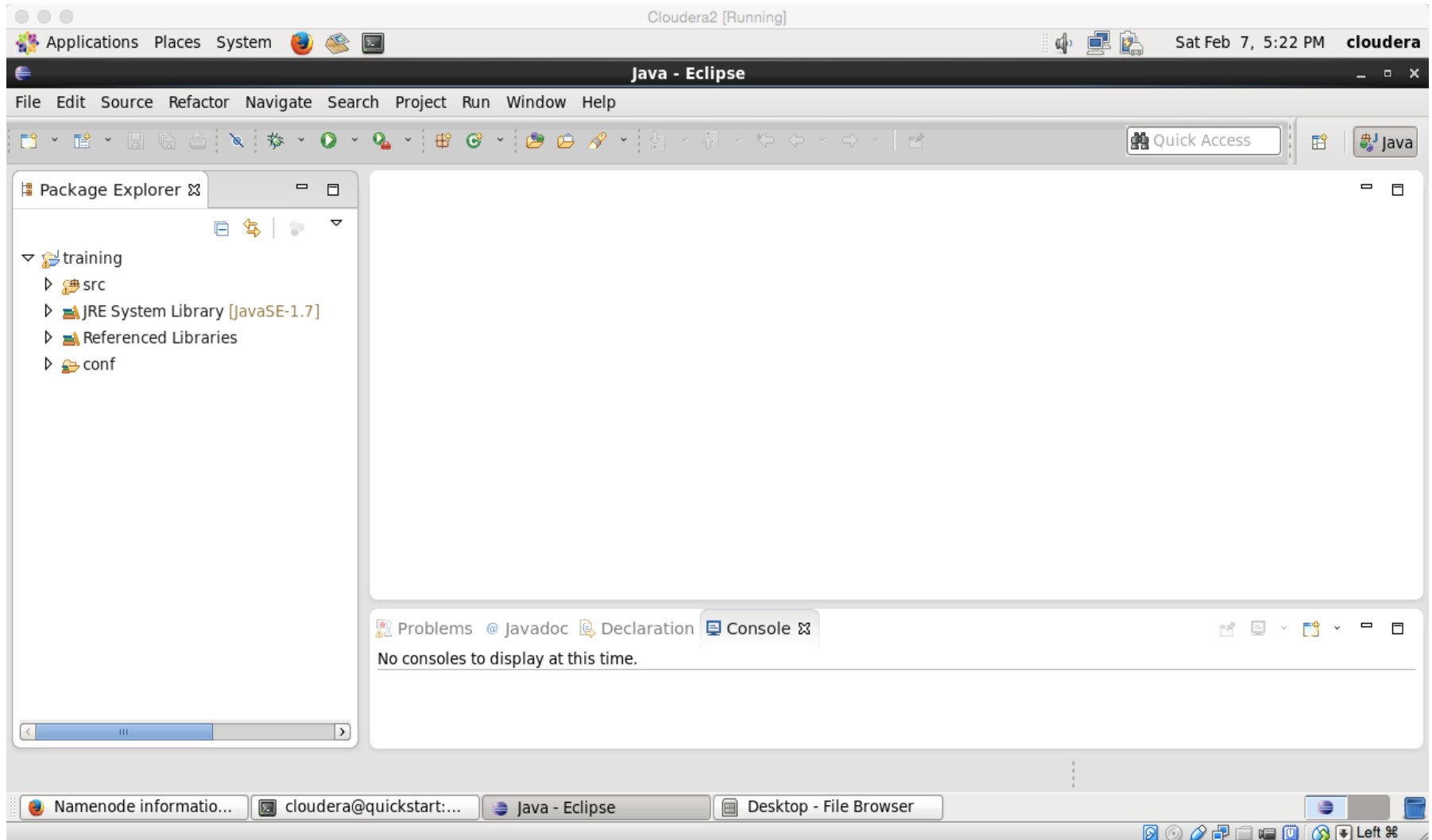
# Wordcount (HelloWord in Hadoop)

```
38. public static void main(String[] args) throws Exception {
39.     JobConf conf = new JobConf(WordCount.class);
40.     conf.setJobName("wordcount");
41.
42.     conf.setOutputKeyClass(Text.class);
43.     conf.setOutputValueClass(IntWritable.class);
44.
45.     conf.setMapperClass(Map.class);
46.
47.     conf.setReducerClass(Reduce.class);
48.
49.     conf.setInputFormat(TextInputFormat.class);
50.     conf.setOutputFormat(TextOutputFormat.class);
51.
52.     FileInputFormat.setInputPaths(conf, new Path(args[1]));
53.     FileOutputFormat.setOutputPath(conf, new Path(args[2]));
54.
55.     JobClient.runJob(conf);
57. }
58. }
59.
```

# **Hands-On: Writing Map/Reduce Program on Eclipse**

---

# Starting Eclipse in a local machine



# Create a Java Project

Let's name it HadoopWordCount

**New Java Project**

Create a Java project in the workspace or in an external location.

Project name:

☒ Use default location

Location:  [Browse...](#)

JRE

☒ Use an execution environment JRE:

☐ Use a project specific JRE:

☐ Use default JRE (currently 'jdk1.7.0\_55-cloudera') [Configure JREs...](#)

Project layout

☐ Use project folder as root for sources and class files

☒ Create separate folders for sources and class files [Configure default...](#)

Working sets

☐ Add project to working sets

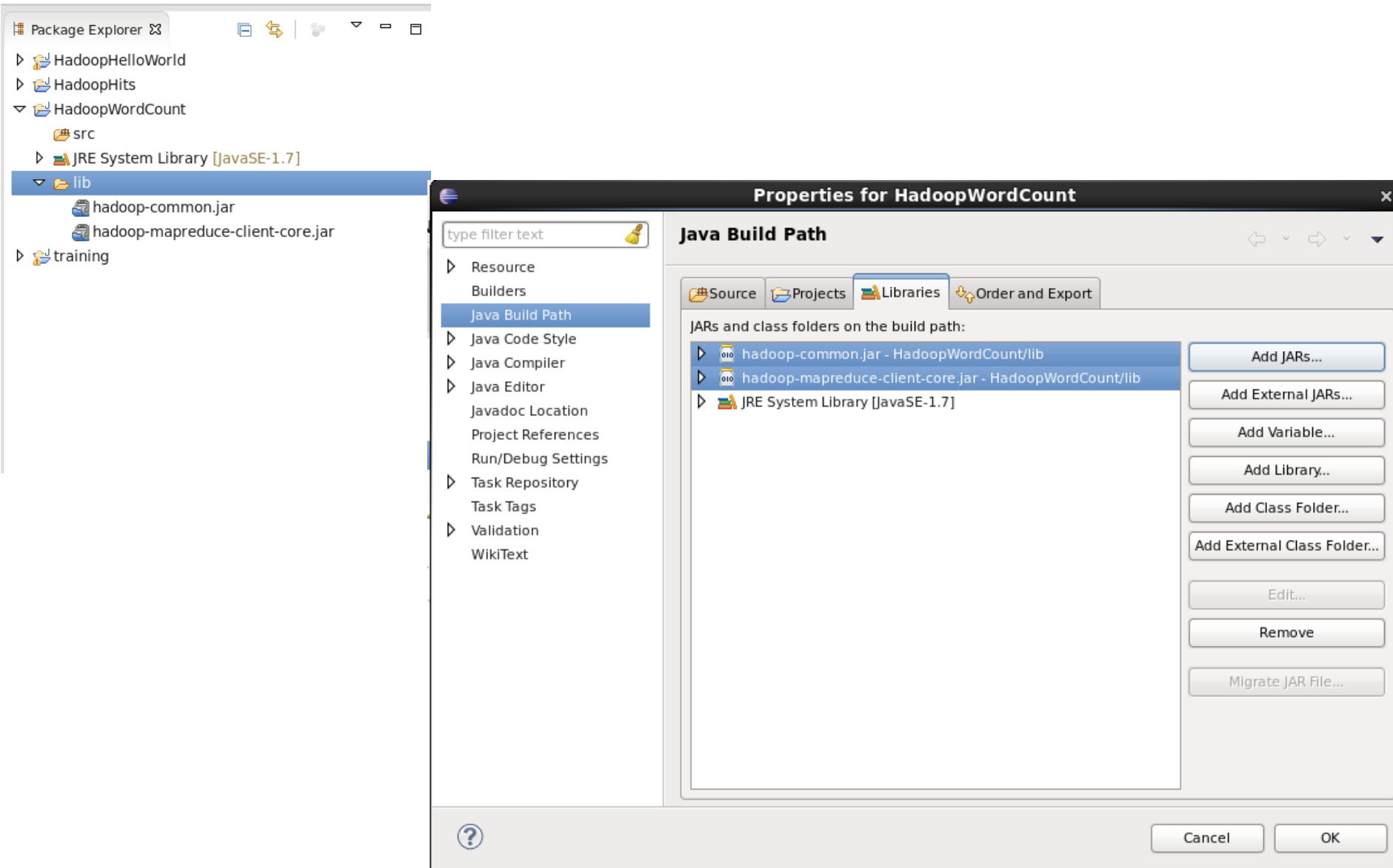
Working sets:  [Select...](#)

[?](#) [< Back](#) [Next >](#) [Cancel](#) [Finish](#)

# Add dependencies to the project

- Note you may need to download Hadoop-core-jar.zip
- Add the following two JARs to your build path
- [hadoop-common.jar](#) and [hadoop-mapreduce-client-core.jar](#).
- By perform the following steps
  - Add a folder named [lib](#) to the project
  - Copy the mentioned JARs in this folder
  - Right-click on the project name >> select **Build Path** >> then **Configure Build Path**
  - Click on Add Jars, select these two JARs from the [lib](#) folder

# Add dependencies to the project



The screenshot shows the Eclipse IDE interface. On the left, the Package Explorer displays the project structure for 'HadoopWordCount', including 'src', 'JRE System Library [JavaSE-1.7]', and a 'lib' folder containing 'hadoop-common.jar' and 'hadoop-mapreduce-client-core.jar'. The 'lib' folder is selected.

The main window shows the 'Properties for HadoopWordCount' dialog, with the 'Java Build Path' tab active. The 'Libraries' sub-tab is selected, showing a list of JARs and class folders on the build path:

- hadoop-common.jar - HadoopWordCount/lib
- hadoop-mapreduce-client-core.jar - HadoopWordCount/lib
- JRE System Library [JavaSE-1.7]

On the right side of the dialog, there are several buttons for managing the build path:

- Add JARs...
- Add External JARs...
- Add Variable...
- Add Library...
- Add Class Folder...
- Add External Class Folder...
- Edit...
- Remove
- Migrate JAR File...

At the bottom of the dialog, there are 'Cancel' and 'OK' buttons.

# Writing a source code

- Right click the project, the select **New >> Package**
- Name the package as org.myorg
- Right click at org.myorg, the select **New >> Class**
- Name the package as WordCount
- Writing a source code as shown in previoud slides



Java - HadoopWordCount/src/org/myorg/WordCount.java - Eclipse

File Edit Source Refactor Navigate Search Project Run Window Help

Quick Access Java

Package Explorer

- HadoopHelloWorld
- HadoopHits
- HadoopWordCount
  - src
    - org.myorg
      - WordCount.java
  - JRE System Library [JavaSE-1.7]
  - Referenced Libraries
  - lib
    - hadoop-common.jar
    - hadoop-mapreduce-client-core.jar
  - training

WordCount.java

```
import org.apache.hadoop.conf.*;
import org.apache.hadoop.io.*;
import org.apache.hadoop.mapred.*;
import org.apache.hadoop.util.*;

public class WordCount {

    public static class Map extends MapReduceBase implements Mapper<LongWritable, Text, Text, IntWritable> {

        private final static IntWritable one = new IntWritable(1);
        private Text word = new Text();


        public void map(LongWritable key, Text value, OutputCollector<Text, IntWritable> output,
            Reporter reporter) throws IOException {
            String line = value.toString();
            StringTokenizer tokenizer = new StringTokenizer(line);
            while (tokenizer.hasMoreTokens()) {
                word.set(tokenizer.nextToken());
                output.collect(word, one);
            }
        }
    }
}
```


Problems Javadoc Declaration Console

# Building a Jar file

- Right click the project, then select **Export**
- Select **Java** and then **JAR** file
- Provide the JAR name, as `wordcount.jar`
- Leave the **JAR package options** as default
- In the **JAR Manifest Specification** section, in the bottom, specify the **Main** class
- In this case, select WordCount
- Click on **Finish**
- The JAR file will be build and will be located at cloudera/workspace

Note: you may need to re-size the dialog font size by select  
Windows >> Preferences >> Appearance >> Colors and Fonts


**JAR Export**
✕

**JAR Manifest Specification**


Customize the manifest file for the JAR file.

Specify the manifest:

☒ **Generate the manifest file**

☐ Save the manifest in the workspace

☐ Use the saved manifest in the generated JAR description file

Manifest file:  Browse...

☐ **Use existing manifest from workspace**

Manifest file:  Browse...

Seal contents:

☐ Seal the JAR Details...

☒ Seal some packages Nothing sealed Details...

Select the class of the application entry point:

Main class:  Browse...

?
< Back
Next >
Cancel
Finish

# **Hands-On: Running Map Reduce and Deploying to Hadoop Runtime Environment**

---

# Running a Jar file

- Create a folder [applications](#) on Amazon S3
- Upload [wordcount.jar](#) to [s3://imcbucket/apps](#)

The screenshot shows the AWS Management Console interface. At the top, there's a navigation bar with 'AWS', 'Services', and 'Edit' dropdowns. On the right, it says 'IMC Institute', 'Global', and 'Support'. Below this, there are buttons for 'Upload', 'Create Folder', and 'Actions'. To the right of these are tabs for 'None', 'Properties', and 'Transfers'. The breadcrumb path is 'All Buckets / imcbucket / apps'. Below this is a table with columns: Name, Storage Class, Size, and Last Modified. The table contains one entry: 'wordcount.jar' with 'Standard' storage class, '3.3 KB' size, and 'Fri Feb 13 08:18:46 GMT+700 2015' as the last modified date.

Name	Storage Class	Size	Last Modified
wordcount.jar	Standard	3.3 KB	Fri Feb 13 08:18:46 GMT+700 2015


# Running a Jar file (cont)

- Open the Master node using SSH command
  - `ssh hadoop@ec2-54-213-220-37.us-west-2.compute.amazonaws.com -i imckey.pem`
- Run the following commands
  - `$ mkdir apps`
  - `$ hadoop fs -get s3://imcbucket/applications/wordcount.jar apps`
  - `$ hadoop jar apps/wordcount.jar org.myorg.WordCount s3://imcbucket/input/* s3://imcbucket/output/wordcount_result`

# Reviewing MapReduce Output Result

Browse to the `s3://imcbucket/output/wordcount_result`

Open part-xxxx files









 **AWS** ▾ **Services** ▾ **Edit** ▾

IMC Institute ▾ Global ▾ Support ▾

**Upload** **Create Folder** **Actions** ▾

**None** **Properties** **Transfers**

[All Buckets](#) / [imcbucket](#) / [output](#) / [wordcount\\_result](#)

	Name	Storage Class	Size	Last Modified
<input type="checkbox"/>	 _SUCCESS	Standard	0 bytes	Fri Feb 13 08:04:32 GMT+700 2015
<input type="checkbox"/>	 part-00000	Standard	0 bytes	Fri Feb 13 08:04:20 GMT+700 2015
<input type="checkbox"/>	 part-00001	Standard	8 bytes	Fri Feb 13 08:04:22 GMT+700 2015
<input type="checkbox"/>	 part-00002	Standard	18 bytes	Fri Feb 13 08:04:23 GMT+700 2015
<input type="checkbox"/>	 part-00003	Standard	9 bytes	Fri Feb 13 08:04:23 GMT+700 2015
<input type="checkbox"/>	 part-00004	Standard	0 bytes	Fri Feb 13 08:04:23 GMT+700 2015
<input type="checkbox"/>	 part-00005	Standard	0 bytes	Fri Feb 13 08:04:29 GMT+700 2015
<input type="checkbox"/>	 part-00006	Standard	9 bytes	Fri Feb 13 08:04:31 GMT+700 2015

# Reviewing MapReduce Output Result

The screenshot shows the HUE File Browser interface. The top navigation bar includes the HUE logo, a home icon, and links to Query Editors, Data Browsers, Workflows, Search, File Browser, Job Browser, and cloudera. The main content area displays the file path `/ user / cloudera / output / wordcount_output / part-00000`. The file content is displayed as a table of word counts.

"About	6
"According	1
"Adele	1
"Adieu,	2
"Adjutant!"	1
"Admirable!"	1
"Adorable!"	1
"Adored	1
"Afraid	3
"After	7
"Again!"	1

On the left sidebar, there are options to View as binary, Download, View file location, and Refresh. Below these is an INFO section showing the file was last modified on Feb. 8, 2015 at 10:31 a.m. by user cloudera.



# Lecture

---

# Understanding Hive



# Introduction

## A Petabyte Scale Data Warehouse Using Hadoop



**Hive is developed by Facebook, designed to enable easy data summarization, ad-hoc querying and analysis of large volumes of data. It provides a simple query language called Hive QL, which is based on SQL**

# Hands-On: Creating Table and Retrieving Data using Hive

---

# Running Hive from the Master node

## Starting Hive

```
[hadoop@ip-172-31-39-229 ~]$ hive
```

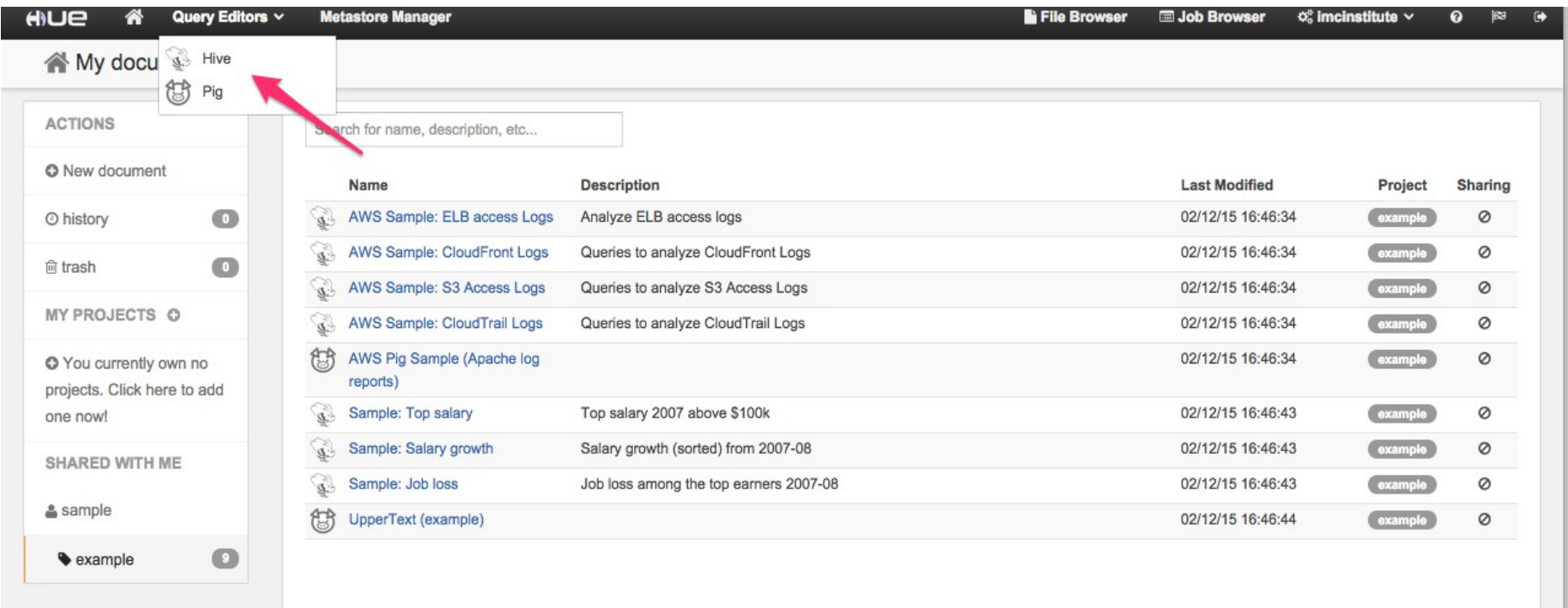
```
Logging initialized using configuration in jar:file:/home/hadoop/.versions/hive-0.13.1/lib/hive-common-0.13.1-amzn-1.jar!/hive-log4j.properties
```

```
hive> █
```

## Quit from Hive

```
hive> quit;
```

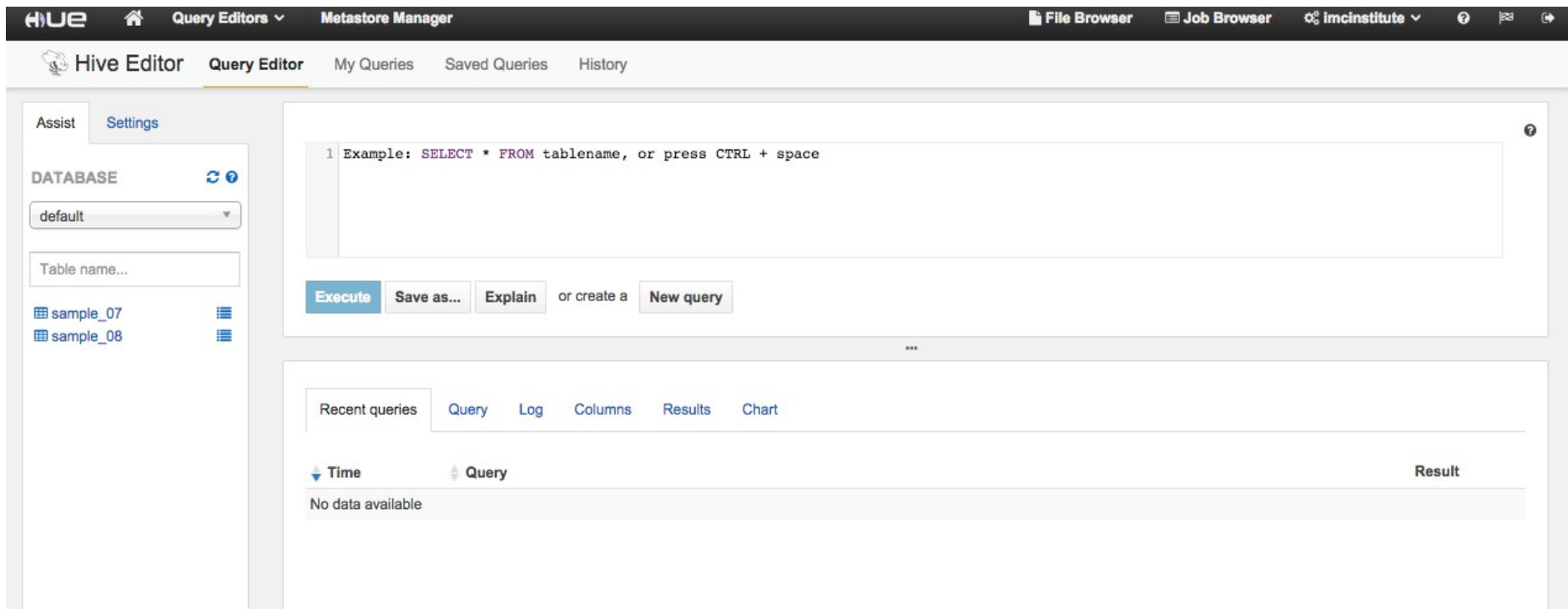
# Starting Hive Editor from Hue



The screenshot shows the Hue web interface. The top navigation bar includes 'HUE', 'Query Editors', 'Metastore Manager', 'File Browser', 'Job Browser', and 'imcinstitute'. The left sidebar contains 'My documents', 'ACTIONS' (New document, history, trash), 'MY PROJECTS', and 'SHARED WITH ME'. The 'Query Editors' dropdown menu is open, showing 'Hive' and 'Pig' options, with a red arrow pointing to 'Hive'. The main content area displays a table of sample queries.

Name	Description	Last Modified	Project	Sharing
<a href="#">AWS Sample: ELB access Logs</a>	Analyze ELB access logs	02/12/15 16:46:34	example	
<a href="#">AWS Sample: CloudFront Logs</a>	Queries to analyze CloudFront Logs	02/12/15 16:46:34	example	
<a href="#">AWS Sample: S3 Access Logs</a>	Queries to analyze S3 Access Logs	02/12/15 16:46:34	example	
<a href="#">AWS Sample: CloudTrail Logs</a>	Queries to analyze CloudTrail Logs	02/12/15 16:46:34	example	
<a href="#">AWS Pig Sample (Apache log reports)</a>		02/12/15 16:46:34	example	
<a href="#">Sample: Top salary</a>	Top salary 2007 above \$100k	02/12/15 16:46:43	example	
<a href="#">Sample: Salary growth</a>	Salary growth (sorted) from 2007-08	02/12/15 16:46:43	example	
<a href="#">Sample: Job loss</a>	Job loss among the top earners 2007-08	02/12/15 16:46:43	example	
<a href="#">UpperText (example)</a>		02/12/15 16:46:44	example	

# Starting Hive Editor from Hue



The screenshot shows the Hue web interface for the Hive Editor. The top navigation bar includes the Hue logo, a home icon, and links to Query Editors, Metastore Manager, File Browser, Job Browser, and Imcinstitute. Below this, the Hive Editor section has tabs for Query Editor (selected), My Queries, Saved Queries, and History. On the left sidebar, there are links for Assist and Settings, a DATABASE dropdown menu set to 'default', a 'Table name...' input field, and a list of tables including 'sample\_07' and 'sample\_08'. The main query editor area contains a text input field with the example query: `1 Example: SELECT * FROM tablename, or press CTRL + space`. Below the input field are buttons for 'Execute', 'Save as...', 'Explain', and 'New query'. At the bottom, there is a 'Recent queries' section with tabs for Query, Log, Columns, Results, and Chart. The 'Query' tab is active, showing a table with columns 'Time', 'Query', and 'Result'. The table is currently empty, displaying 'No data available'.

# Creating Hive Table

```
hive (default)> CREATE TABLE test_tbl(id INT, country STRING) ROW  
FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE;
```

```
OK
```

```
Time taken: 4.069 seconds
```

```
hive (default)> show tables;
```

```
OK
```

```
test_tbl
```

```
Time taken: 0.138 seconds
```

```
hive (default)> describe test_tbl;
```

```
OK
```

```
id    int
```

```
country string
```

```
Time taken: 0.147 seconds
```

```
hive (default)>
```

See also: <https://cwiki.apache.org/Hive/languagemanual-ddl.html>

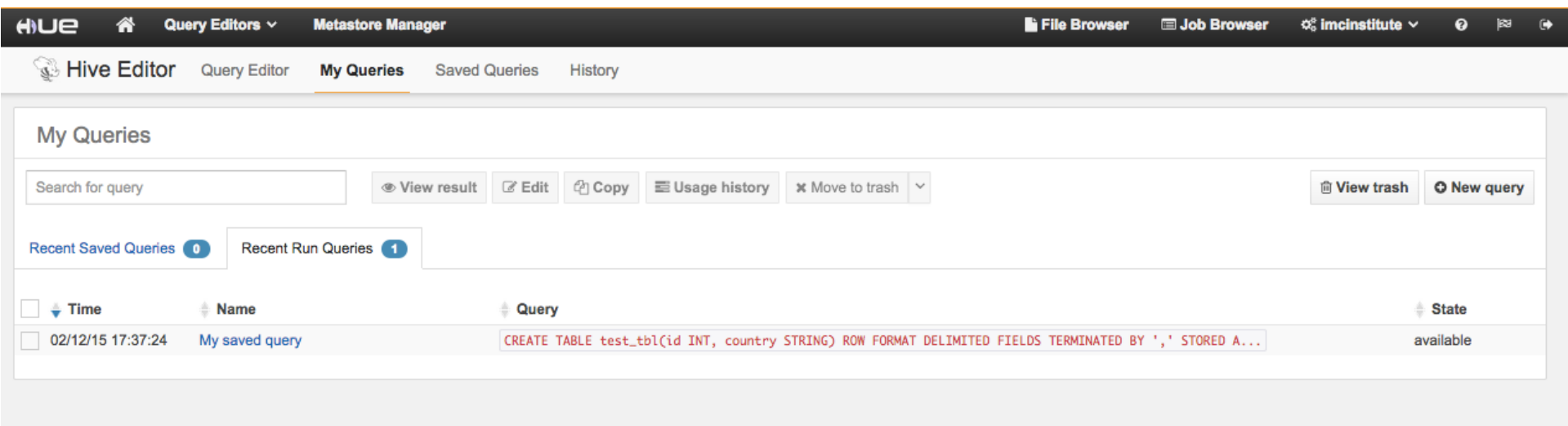
# Using Hue Query Editor

The screenshot displays the Hue Query Editor interface. The top navigation bar includes the Hue logo, a home icon, and tabs for 'Query Editors' and 'Metastore Manager'. On the right, there are links for 'File Browser', 'Job Browser', and the 'imcinstitute' user profile. A dropdown menu for 'Query Editors' is open, showing options for 'Hive' and 'Pig'. The left sidebar contains an 'Assist' button, a 'Settings' link, and a 'DATABASE' section with a 'default' dropdown and a 'Table name...' input field. Below this, a list of tables 'sample\_07' and 'sample\_08' is visible. The main query editor area contains a single line of SQL: `1 CREATE TABLE test_tbl(id INT, country STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE;|`. Below the query, there are buttons for 'Execute', 'Save as...', 'Explain', and 'New query'. The bottom section shows a 'Recent queries' tab with sub-tabs for 'Query', 'Log', 'Columns', 'Results', and 'Chart'. A table with columns 'Time', 'Query', and 'Result' is shown, with the message 'No data available' in the first row.

See also: <https://cwiki.apache.org/Hive/languagemanual-ddl.html>



# Using Hue Query Editor



The screenshot shows the Hue Query Editor interface. The top navigation bar includes 'HUE', 'Home', 'Query Editors', 'Metastore Manager', 'File Browser', 'Job Browser', and 'Incubate'. The 'Hive Editor' section has tabs for 'Query Editor', 'My Queries' (selected), 'Saved Queries', and 'History'.

Under the 'My Queries' tab, there is a search bar and several action buttons: 'View result', 'Edit', 'Copy', 'Usage history', 'Move to trash', 'View trash', and 'New query'.

Below the buttons, there are two sections: 'Recent Saved Queries' (0) and 'Recent Run Queries' (1). The 'Recent Run Queries' section contains a table with the following data:

<input type="checkbox"/>	Time	Name	Query	State
<input type="checkbox"/>	02/12/15 17:37:24	My saved query	CREATE TABLE test_tbl(id INT, country STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED A...	available

See also: <https://cwiki.apache.org/Hive/languagemanual-ddl.html>

# Reviewing Hive Table in HDFS

**HUE** [Home](#) [Query Editors](#) [Metastore Manager](#) [File Browser](#) [Job Browser](#) [imcinstitute](#) [?](#) [🚩](#) [🔗](#)

## File Browser

Search for file name  [Rename](#) [Move](#) [Copy](#) [Change permissions](#) [Download](#) [Move to trash](#) [Upload](#) [New](#)

[Home](#) [S3](#) / [user](#) / [hive](#) / [warehouse](#) [Trash](#)

<input type="checkbox"/>	Name	Size	User	Group	Permissions	Date
<input type="checkbox"/>	<a href="#">f</a>		hadoop	supergroup	drwxr-xr-x	February 12, 2015 04:44 PM
<input type="checkbox"/>	<a href="#">.</a>		hadoop	supergroup	drwxrwxrwt	February 12, 2015 05:37 PM
<input type="checkbox"/>	<a href="#">sample_07</a>		sample	supergroup	drwxr-xr-x	February 12, 2015 04:46 PM
<input type="checkbox"/>	<a href="#">sample_08</a>		sample	supergroup	drwxr-xr-x	February 12, 2015 04:46 PM
<input type="checkbox"/>	<a href="#">test_tbl</a>		imcinstitute	supergroup	drwxr-xr-x	February 12, 2015 05:37 PM

Show  items per page. Showing 1 to 3 of 3 items, page 1 of 1.

# Alter and Drop Hive Table

```
hive (default)> alter table test_tbl add columns (remarks STRING);
```

```
hive (default)> describe test_tbl;
```

```
OK
```

```
id    int
```

```
country string
```

```
remarks string
```

```
Time taken: 0.077 seconds
```

```
hive (default)> drop table test_tbl;
```

```
OK
```

```
Time taken: 0.9 seconds
```

See also: <https://cwiki.apache.org/Hive/adminmanual-metastoreadmin.html>

# Loading Data to Hive Table

## Creating Hive table

```
$ hive
```

```
hive (default)> CREATE TABLE test_tbl(id INT, country STRING) ROW  
FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE;
```

## Loading data to Hive table

```
hive (default)> LOAD DATA LOCAL INPATH '/tmp/country.csv' INTO TABLE test_tbl;  
Copying data from file:/tmp/test_tbl_data.csv  
Copying file: file:/tmp/test_tbl_data.csv  
Loading data to table default.test_tbl  
OK  
Time taken: 0.241 seconds  
hive (default)>
```

# Querying Data from Hive Table

```
hive (default)> select * from test_tbl;
```

```
OK
```

```
1    USA
```

```
62   Indonesia
```

```
63   Philippines
```

```
65   Singapore
```

```
66   Thailand
```

```
Time taken: 0.287 seconds
```

```
hive (default)>
```

# Insert Overwriting the Hive Table

```
hive (default)> LOAD DATA LOCAL INPATH  
'/home/cloudera/Downloads/test_tbl_data_updated.csv' overwrite INTO  
TABLE test_tbl;
```

```
Copying data from file:/tmp/test_tbl_data_updated.csv
```

```
Copying file: file:/tmp/test_tbl_data_updated.csv
```

```
Loading data to table default.test_tbl
```

```
Deleted hdfs://localhost:54310/user/hive/warehouse/test_tbl
```

```
OK
```

```
Time taken: 0.204 seconds
```

```
hive (default)>
```

# MovieLens

<http://grouplens.org/datasets/movielens/>

grouplens about datasets publications blog

## MovieLens

GroupLens Research has collected and made available rating data sets from the MovieLens web site (<http://movielens.org>). The data sets were collected over various periods of time, depending on the size of the set. Before using these data sets, please review their README files for the usage licenses and other details.

### MovieLens 100k

100,000 ratings from 1000 users on 1700 movies.

- [README.txt](#)
- [ml-100k.zip](#)
- [Index of unzipped files](#)

### MovieLens 1M

1 million ratings from 6000 users on 4000 movies.

## Datasets

[MovieLens](#)

[HetRec 2011](#)

[WikiLens](#)

[Book-Crossing](#)

[Jester](#)

[EachMovie](#)

# Create the Hive Table for movielen

```
hive (default)> CREATE TABLE u_data (  
    userid INT,  
    movieid INT,  
    rating INT,  
    unixtime STRING)  
ROW FORMAT DELIMITED  
FIELDS TERMINATED BY '\t'  
STORED AS TEXTFILE;  
  
hive (default)> LOAD DATA LOCAL INPATH  
'/home/cloudera/Downloads/u.data' overwrite INTO TABLE u_data;
```



# Create the Hive Table for Apache LOf

```
hive (default)> CREATE TABLE apachelog (  
    host STRING,  
    identity STRING,  
    user STRING,  
    time STRING,  
    request STRING,  
    status STRING,  
    size STRING,  
    referer STRING,  
    agent STRING)  
  
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.RegexSerDe'  
WITH SERDEPROPERTIES (  
    "input.regex" = "([^]*) ([^]*) ([^]*) (-|\\\[^\\" data-bbox="101 201 829 903"/>
```



# Lecture

---

# Understanding Pig

# Introduction

A high-level platform for creating MapReduce programs Using Hadoop



**Pig is a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs. The salient property of Pig programs is that their structure is amenable to substantial parallelization, which in turns enables them to handle very large data sets.**

# Hands-On: Running a Pig script

---

# Starting Pig Command Line

```
[hadoop@ip-172-31-39-229 ~]$ pig -x local
2015-02-13 01:52:33,278 [main] INFO  org.apache.pig.Main - Apache Pig version 0.12.0 (rexported) compiled Oct 27 2014, 17:56:51
2015-02-13 01:52:33,278 [main] INFO  org.apache.pig.Main - Logging error message s to: /mnt/var/log/apps/pig.log
2015-02-13 01:52:33,298 [main] INFO  org.apache.pig.impl.util.Utils - Default bootstrap file /home/hadoop/.pigbootstrap not found
2015-02-13 01:52:33,562 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2015-02-13 01:52:33,562 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2015-02-13 01:52:33,565 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: file:///
2015-02-13 01:52:33,920 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2015-02-13 01:52:33,922 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> █
```

# Starting Pig from Hue

The screenshot displays the Hue Pig Editor interface. The top navigation bar includes the Hue logo, a home icon, and links to Query Editors, Metastore Manager, File Browser, Job Browser, and the IMC Institute logo. Below this, the Pig Editor tab is active, showing sub-tabs for Editor, Scripts, and Dashboard. The left sidebar contains an EDITOR section with options for Pig, Properties, Save, New Script, and a RUN section with Submit and Logs. The main editor area shows an 'Unsaved script' header and a single line of Pig Latin code: `1 ie. A = LOAD '/user/imcinstitute/data';`. On the right, an 'Assist' panel provides a search box for function names and a list of categories: Eval Functions, Relational Operators, Input/Output, Debug, HCatalog, Math, Tuple, Bag, Map Functions, String Functions, Macros, HBase, and Python UDF.

HUE

Query Editors Metastore Manager

File Browser Job Browser IMC Institute

Pig Editor Editor Scripts Dashboard

EDITOR

Pig

Properties

Save

New Script

RUN

Submit

Logs

?

Unsaved script

```
1 ie. A = LOAD '/user/imcinstitute/data';
```

Assist

Function name...

- Eval Functions
- Relational Operators
- Input/Output
- Debug
- HCatalog
- Math
- Tuple, Bag, Map Functions
- String Functions
- Macros
- HBase
- Python UDF

# Writing a Pig Script

countryFilter.pig

```
A = load 'hdi-data.csv' using PigStorage(',') AS (id:int, country:chararray, hdi:float,
lifeex:int, mysch:
nt, eysch:int, gni:int);
B = FILTER A BY gni > 2000;
C = ORDER B BY gni;
dump C;
```

# Running a Pig Script

```
[hdadmin@localhost ~]$ cd Downloads
```

```
[hdadmin@localhost ~]$ pig -x local
```

```
grunt > run countryFilter.pig
```

```
....
```

```
(150,Cameroon,0.482,51,5,10,2031)
```

```
(126,Kyrgyzstan,0.615,67,9,12,2036)
```

```
(156,Nigeria,0.459,51,5,8,2069)
```

```
(154,Yemen,0.462,65,2,8,2213)
```

```
(138,Lao People's Democratic Republic,0.524,67,4,9,2242)
```

```
(153,Papua New Guinea,0.466,62,4,5,2271)
```

```
(165,Djibouti,0.43,57,3,5,2335)
```

```
(129,Nicaragua,0.589,74,5,10,2430)
```

```
(145,Pakistan,0.504,65,4,6,2550)
```



# Big Data Certification Course

120 Hrs: Start 12 March 2015



<http://www.imcinstitute.com/bigdatacert>

Tel : 088-192-7975

# Thank you

[www.imcinstitute.com](http://www.imcinstitute.com)

[www.facebook.com/imcinstitute](https://www.facebook.com/imcinstitute)