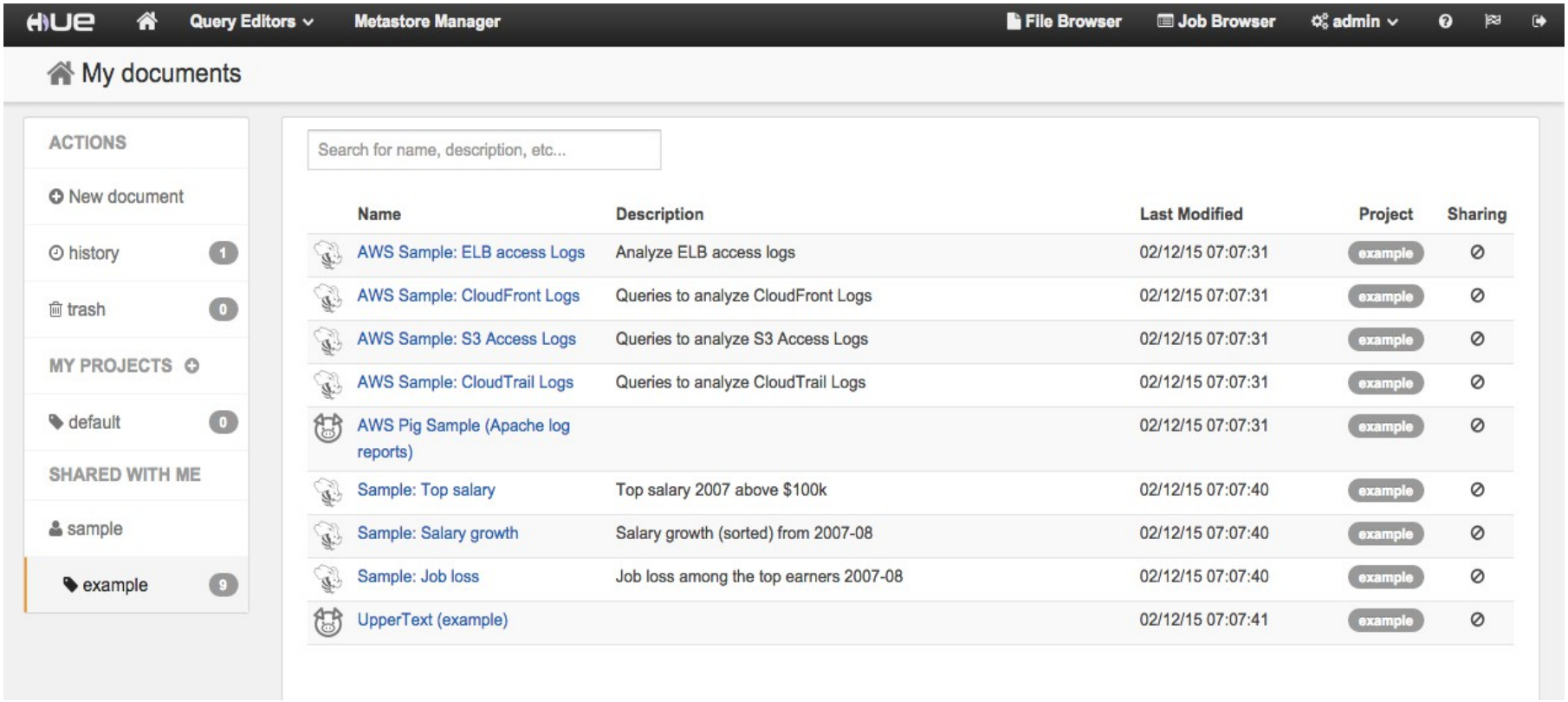
Big Data Analytics Using Hadoop Cluster On Amazon EMR

February 2015

Dr.Thanachart Numnonda
IMC Institute
thanachart@imcinstitute.com

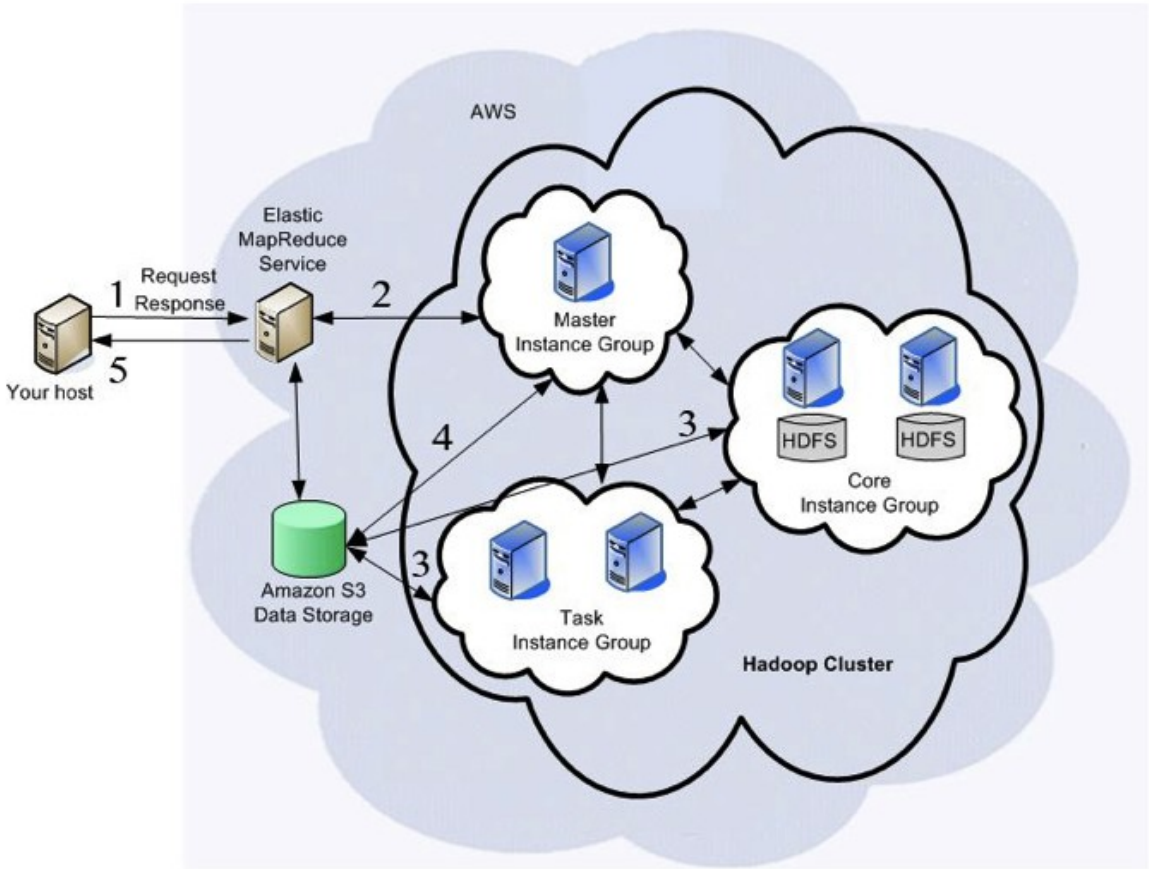
Modifiy from Original Version by Danairat T.
Certified Java Programmer, TOGAF – Silver
danairat@gmail.com

Running this lab using Amazon EMR

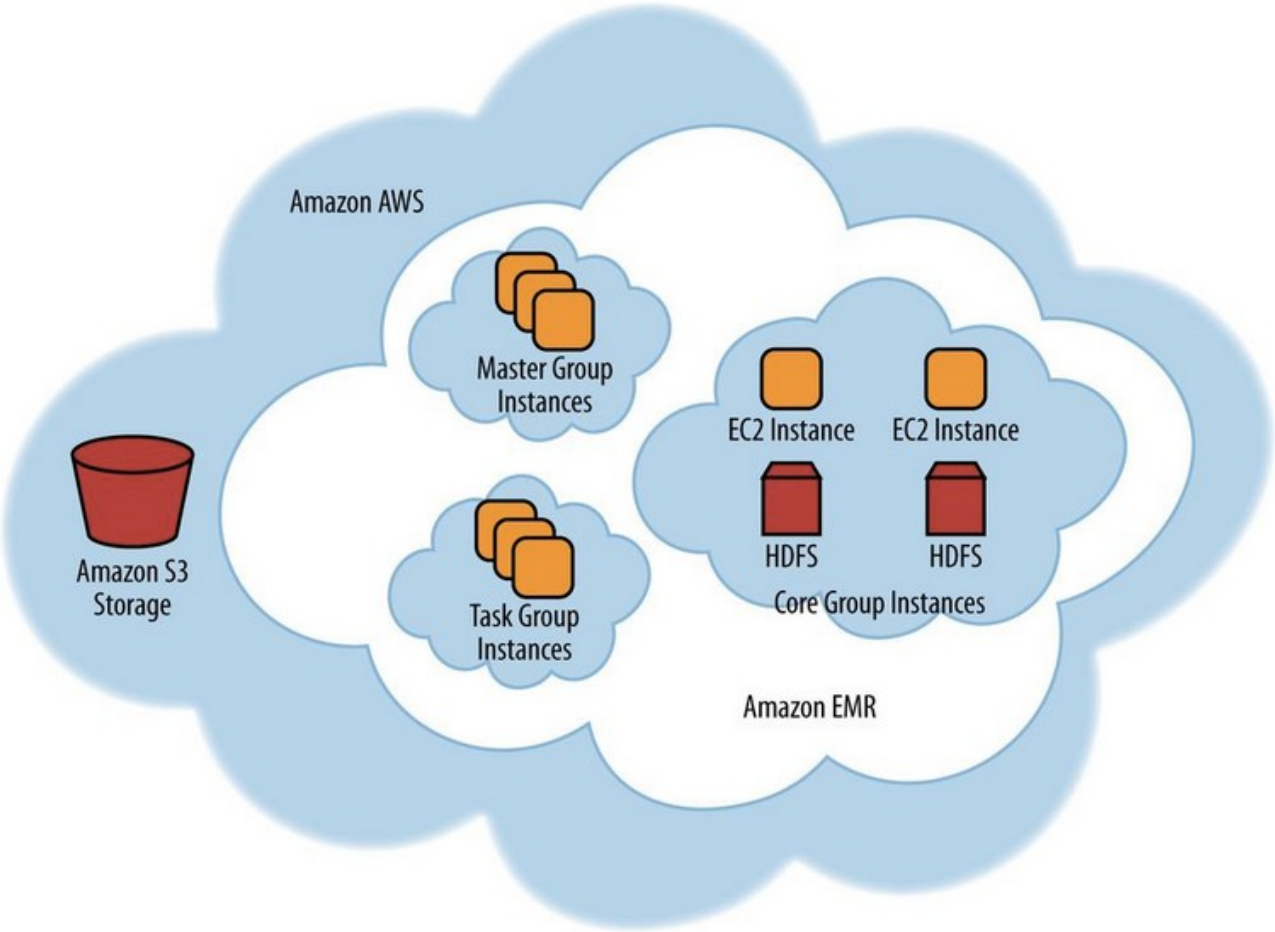


Hands-On: Create an EMR cluster

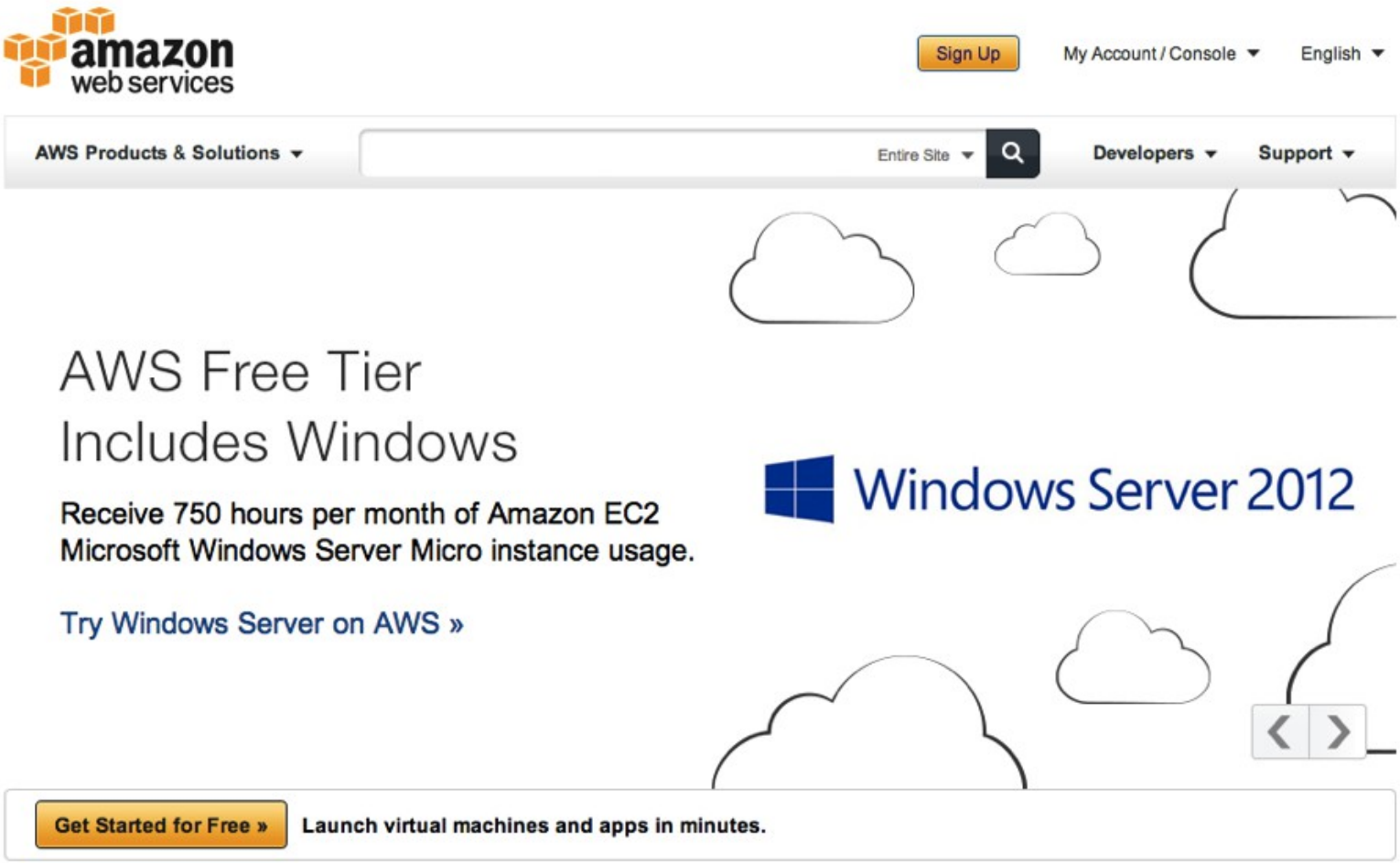
Architecture Overview of Amazon EMR



Amazon EMR Cluster



Creating an AWS account



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

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

Additional Information

Getting Started Guide

Documentation

All S3 Resources

Forums

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	[redacted]	Show	Active (Make Inactive)

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	[redacted]	Show	Active (Make Inactive)

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

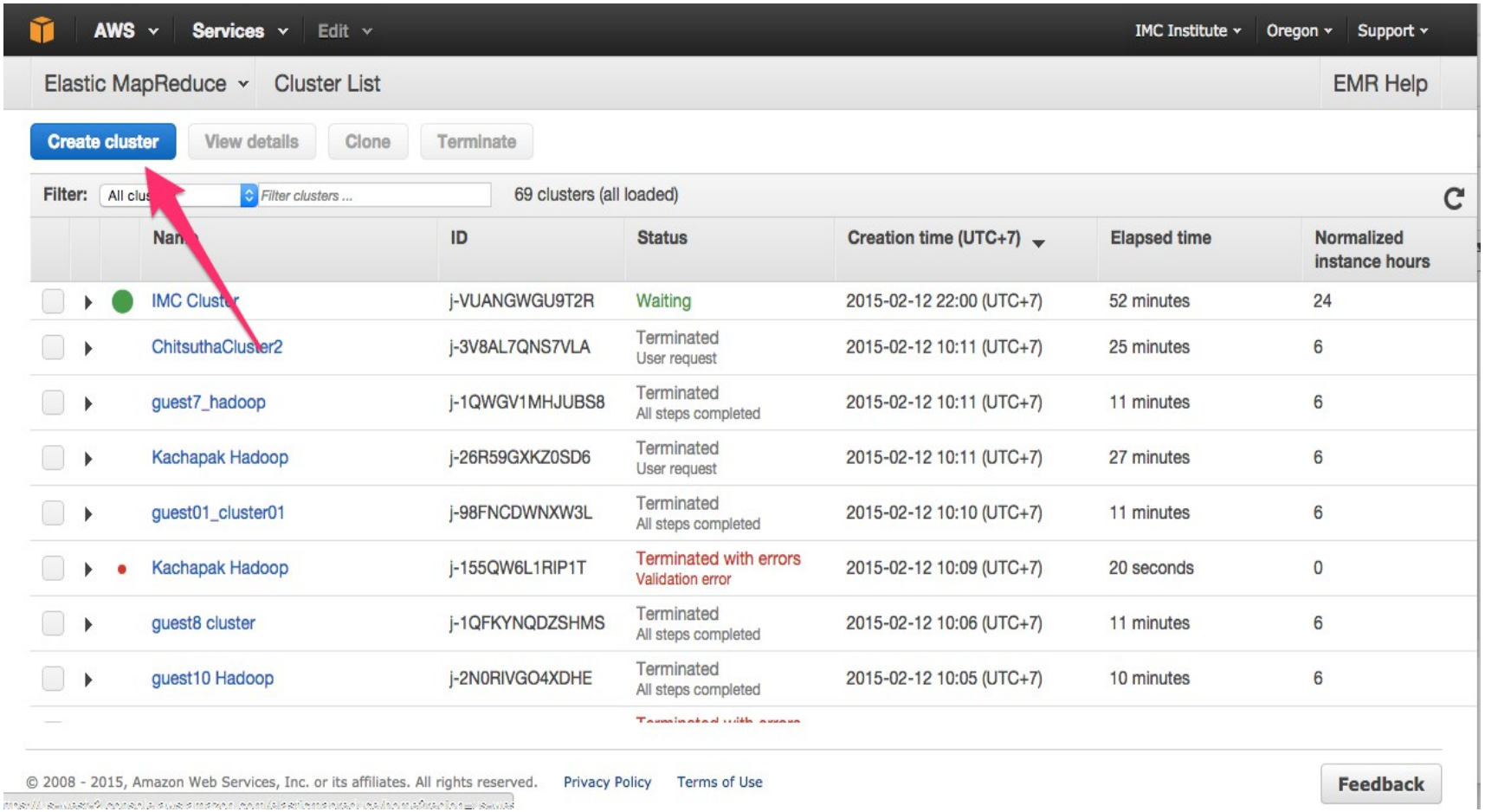
Mobile Services

- **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



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

Configure sample application

Cluster Configuration

Cluster name IMC Cluster

Termination protection Yes / No

Prevents accidental termination of the cluster: to shut down the cluster, you must turn off termination protection. Learn more

Logging Enabled

Copy the cluster's log files automatically to S3. Learn more

Log folder S3 location

s3://imcbucket/elasticmapreduce/

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

Debugging Enabled

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

Tags

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

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

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 Select an application

Configure and add

Creating a cluster in EMR (cont.)

Leave the Hardware Configuration as default

Choose an exisitng 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 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		
Core	Core instance group - 2	m3.xlarge	2		
Task	Task instance group - 3	m3.xlarge	0		

Add task instance group

Security and Access

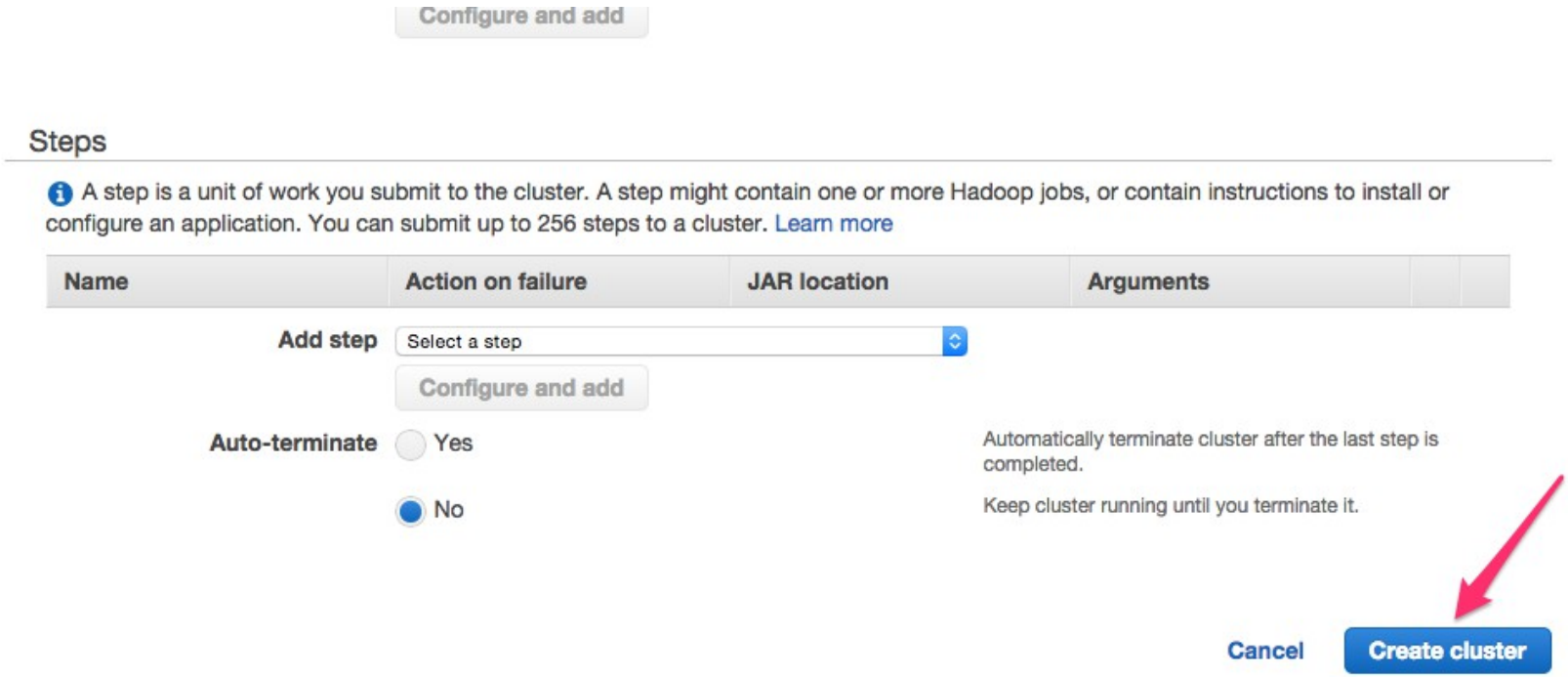
EC2 key pair imckey — Use an existing EC2 key pair to SSH into the master node of the Amazon EMR cluster. Learn more

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



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

ID: j-VUANGWGU9T2R

Creation date: 2015-02-12 22:00 (UTC+7)

Elapsed time: 1 hour, 3 minutes

Auto-terminate: No

Termination protection: On Change

Configuration Details

AMI version: 3.3.2

Hadoop distribution: Amazon 2.4.0

Applications: Hive 0.13.1, Pig 0.12.0, Hue

Log URI: s3://imcbucket/logs/

EMRFS consistent view: Disabled

Network and Hardware Resize

Availability zone: us-west-2b

Subnet ID: subnet-c0510ca8

Master: Running 1 m3.xlarge

Core: Running 2 m3.xlarge

Task: --

Security and Access

Key name: imckey

EC2 instance profile: --

EMR role: --

Visible to all users: All Change

Security groups for Master: sg-46638029 (ElasticMapReduce-master)

Security groups for Core & Task: sg-4563802a (ElasticMapReduce-slave)

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

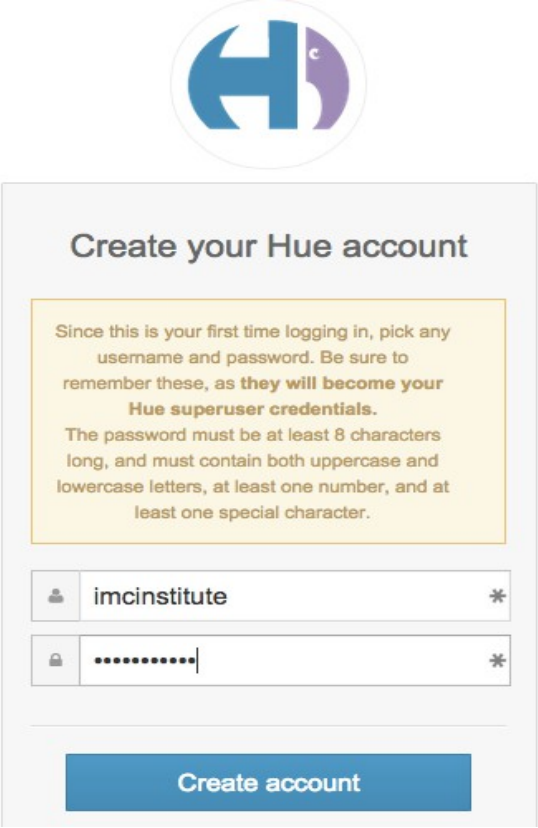
Name of Interface	URI
Hadoop version 2.x	
Hadoop ResourceManager	http://*master-public-dns-name*:9026/
Hadoop HDFS NameNode	http://*master-public-dns-name*:9101/
Ganglia Metrics Reports	http://*master-public-dns-name*/ganglia/
HBase Interface	http://*master-public-dns-name*:60010/master-status
Hue Web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

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



Launch the Hue Web Interface (Cont.)

ec2-54-201-231-114.us-west-2.compute.amazonaws.com:8888/home

Query Editors | Metastore Manager | File Browser | Job Browser | admin

My documents

ACTIONS
- New document
- history 1
- trash 0

MY PROJECTS
- default 0

SHARED WITH ME
- sample
- example 9

Search for name, description, etc...

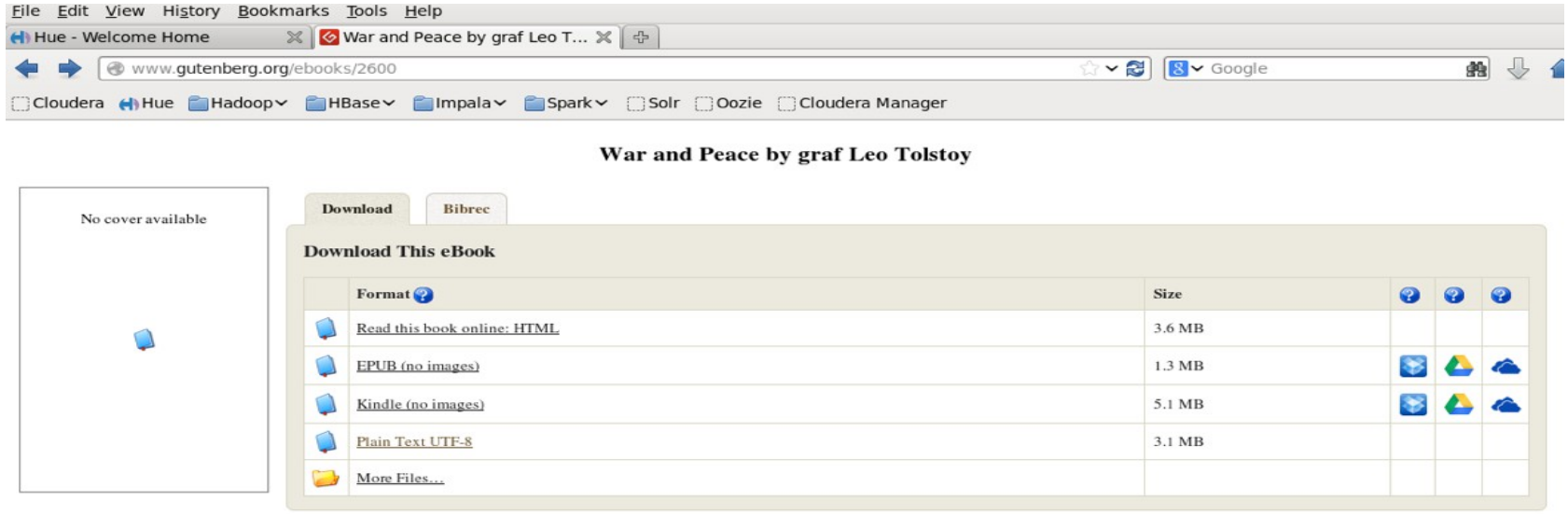
Name	Description	Last Modified	Project	Sharing
AWS Sample: ELB access Logs	Analyze ELB access logs	02/12/15 07:07:31	example	
AWS Sample: CloudFront Logs	Queries to analyze CloudFront Logs	02/12/15 07:07:31	example	
AWS Sample: S3 Access Logs	Queries to analyze S3 Access Logs	02/12/15 07:07:31	example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

Hands-On: Importing/Exporting Data to HDFS

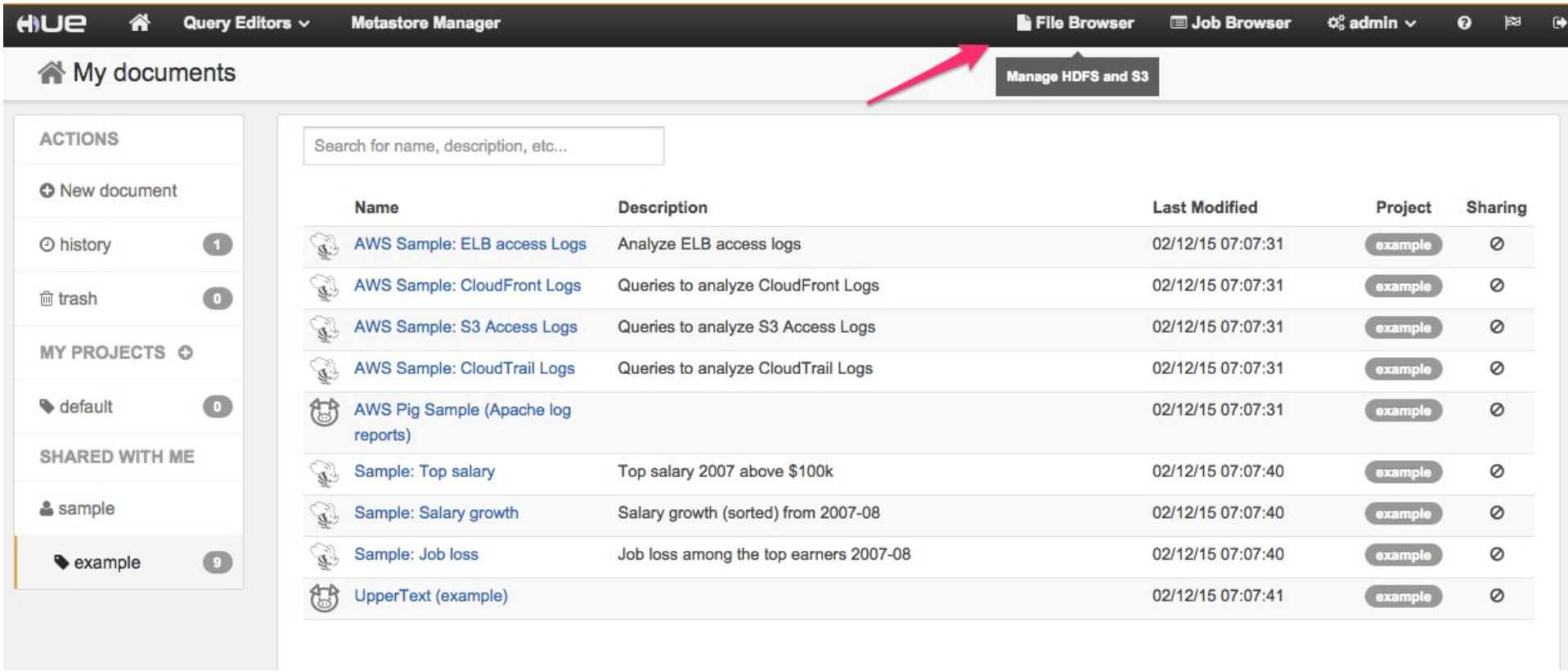
Importing Data to Hadoop

Download War and Peace Full Text

www.gutenberg.org/ebooks/2600

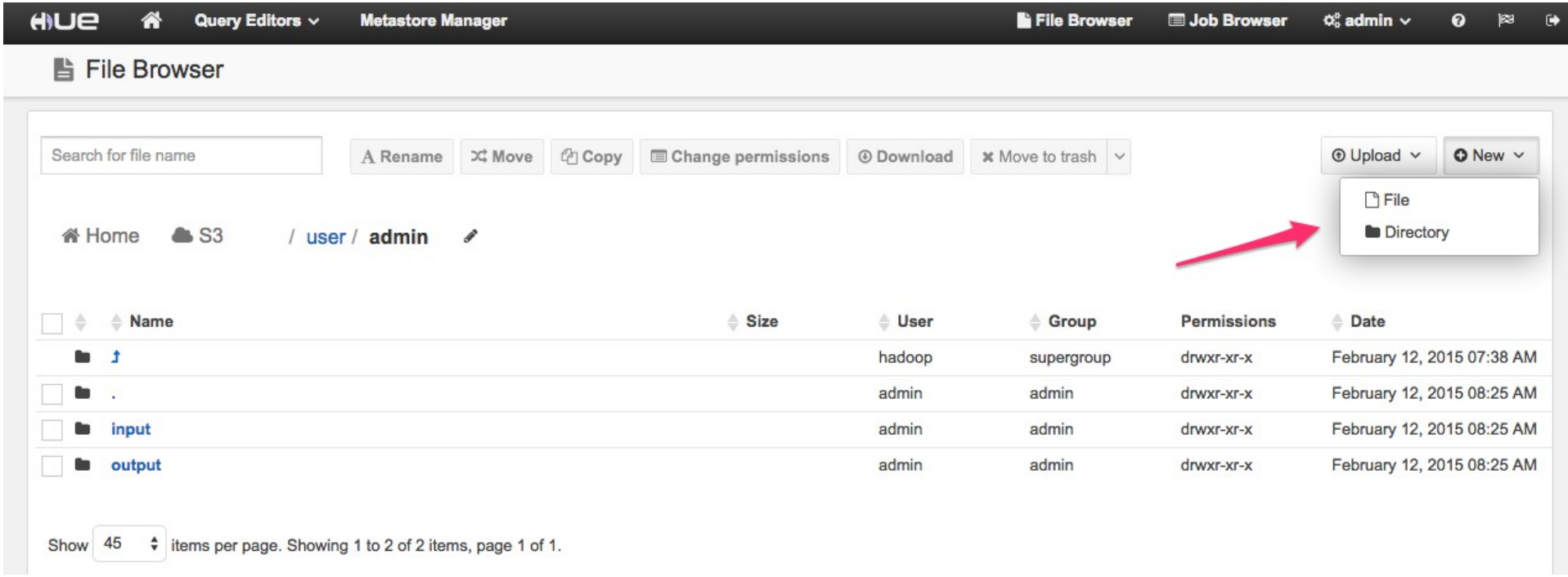


Review file in Hadoop HDFS using File Browse



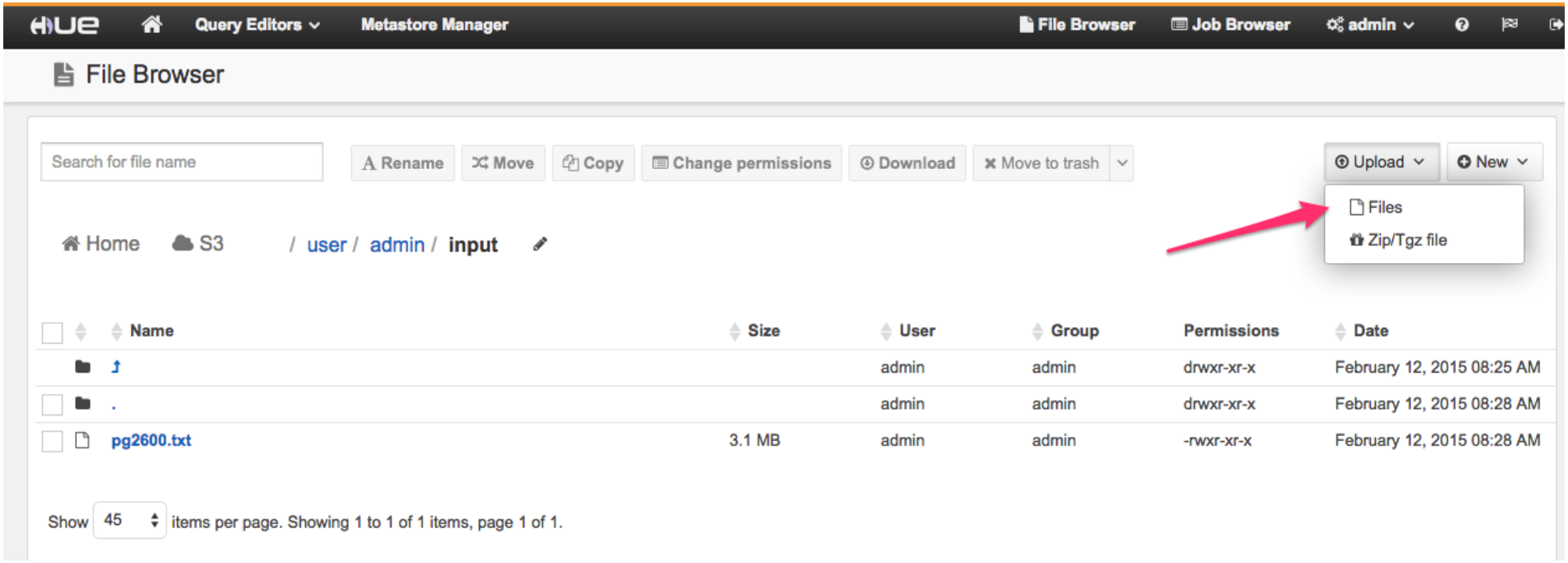
Create new directory

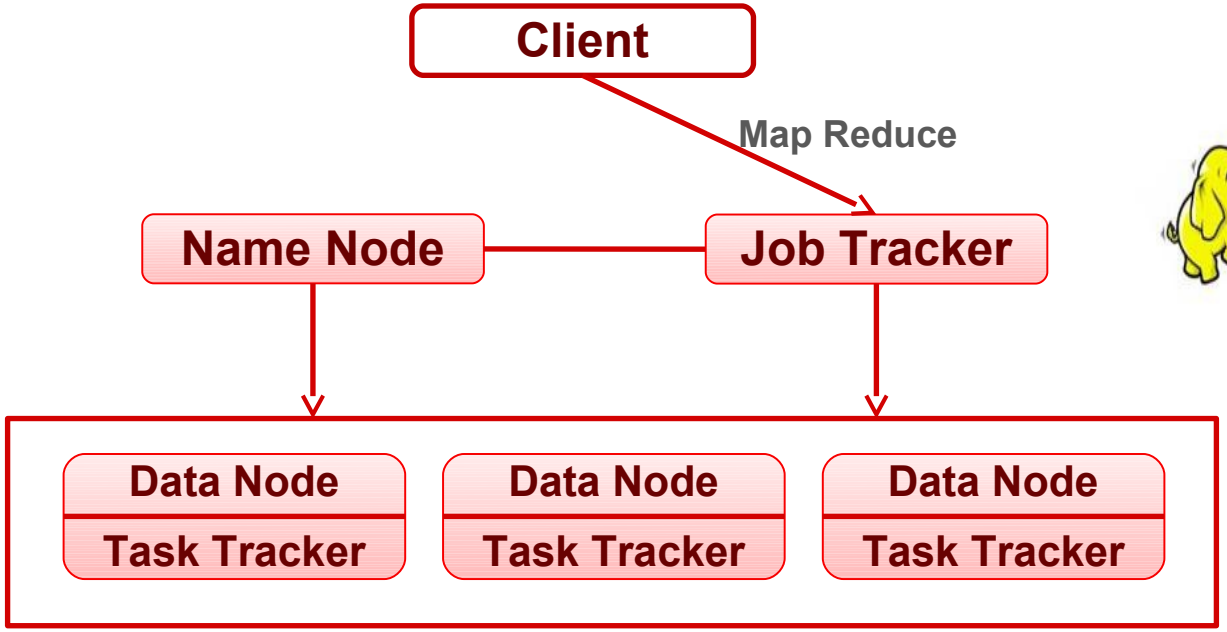
Create two new directory name: input and output



Upload Files

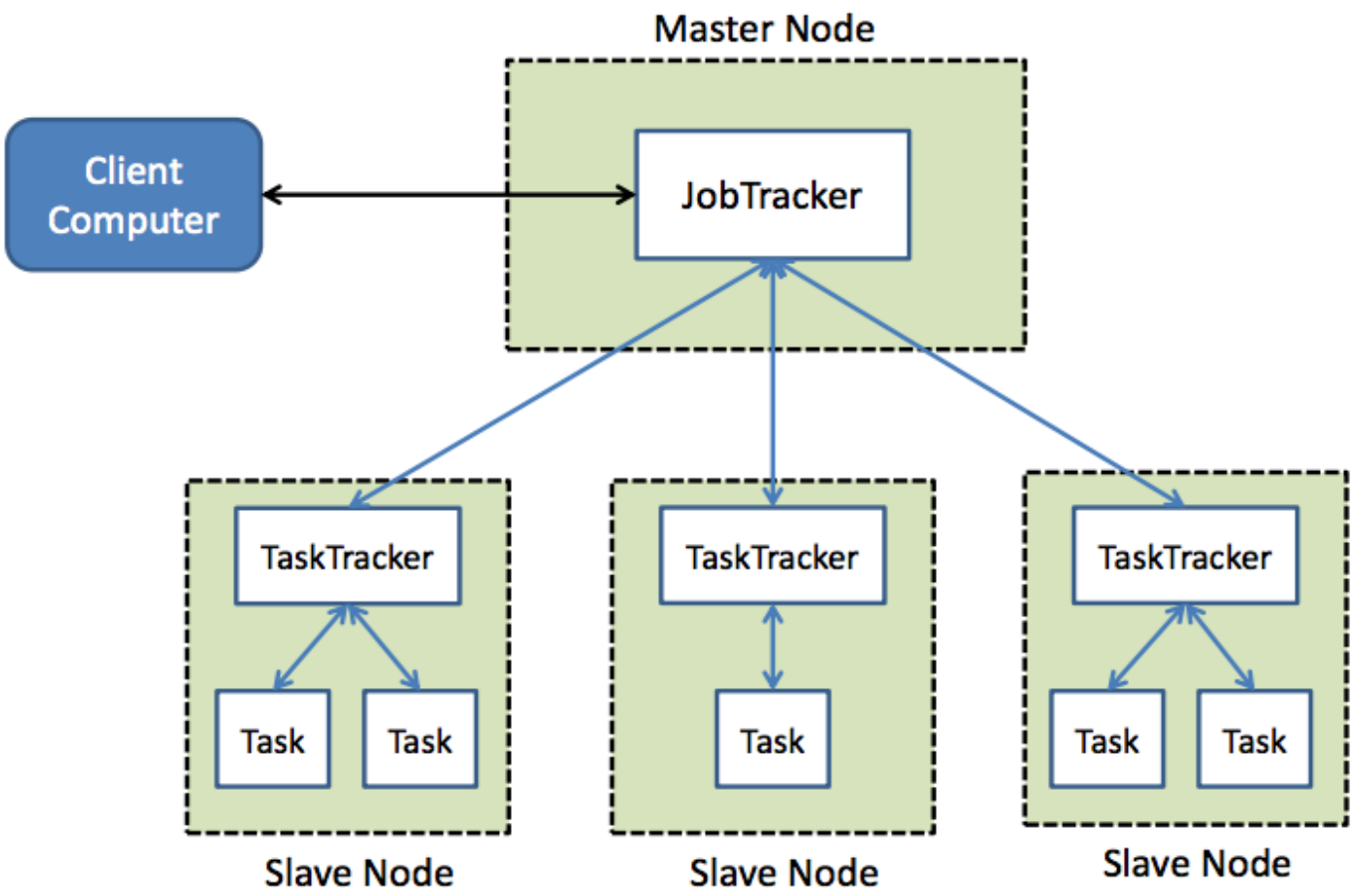
Upload file: pg2600.txt into input directory





Lecture: Understanding Map Reduce Processing

High Level Architecture of MapReduce



Hands-On: Writing you own Map Reduce Program

Wordcount (HelloWord in Hadoop)

```
package org.myorg;

import java.io.IOException;
import java.util.*;

import org.apache.hadoop.fs.Path;
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);
      }
    }
  }
```

Wordcount (HelloWord in Hadoop)

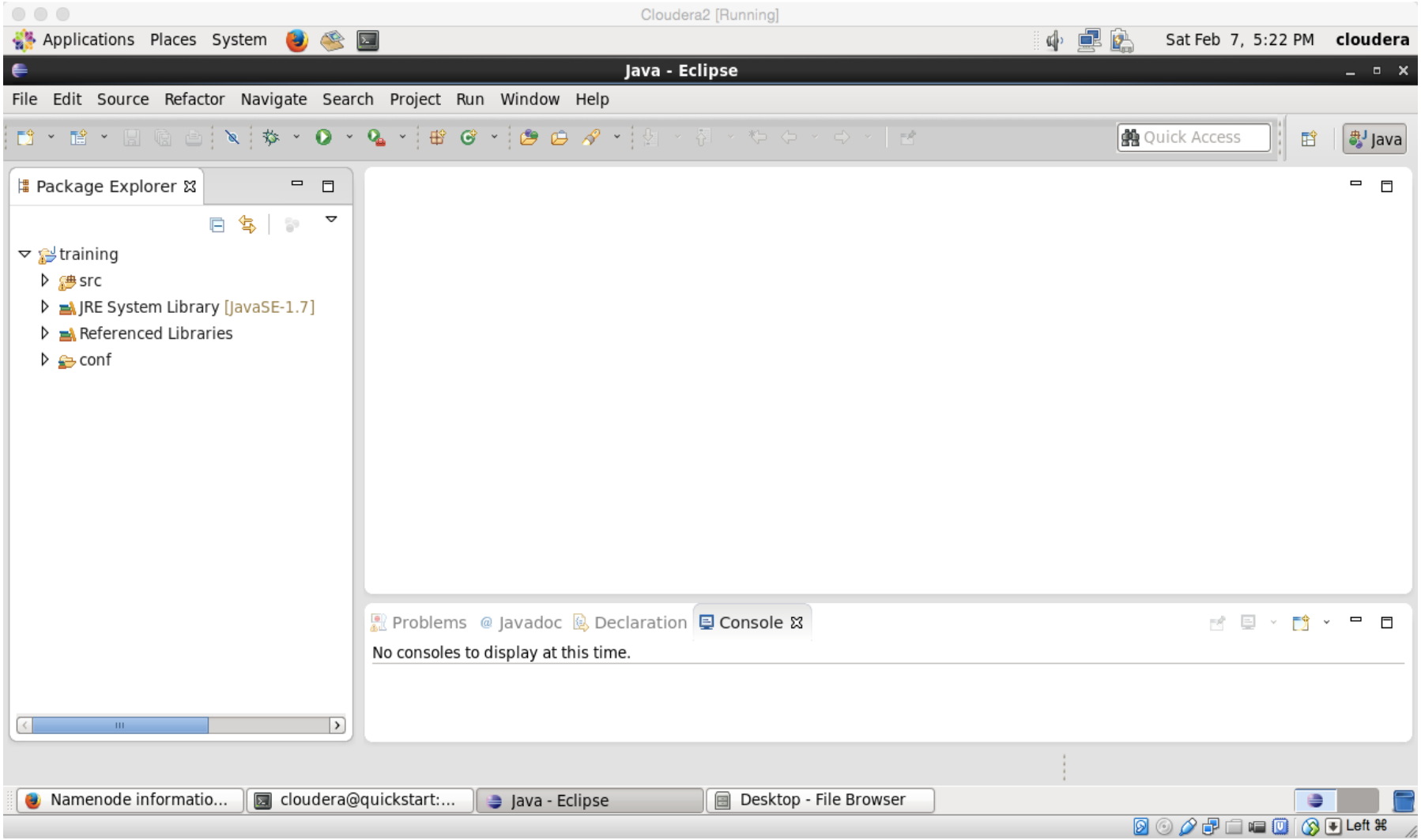
```
27.
28.   public static class Reduce extends MapReduceBase implements Reducer<Text, IntWritable, Text, IntWritable> {
29.     public void reduce(Text key, Iterator<IntWritable> values, OutputCollector<Text, IntWritable> output, Reporter reporter) throws IOException {
30.       int sum = 0;
31.  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

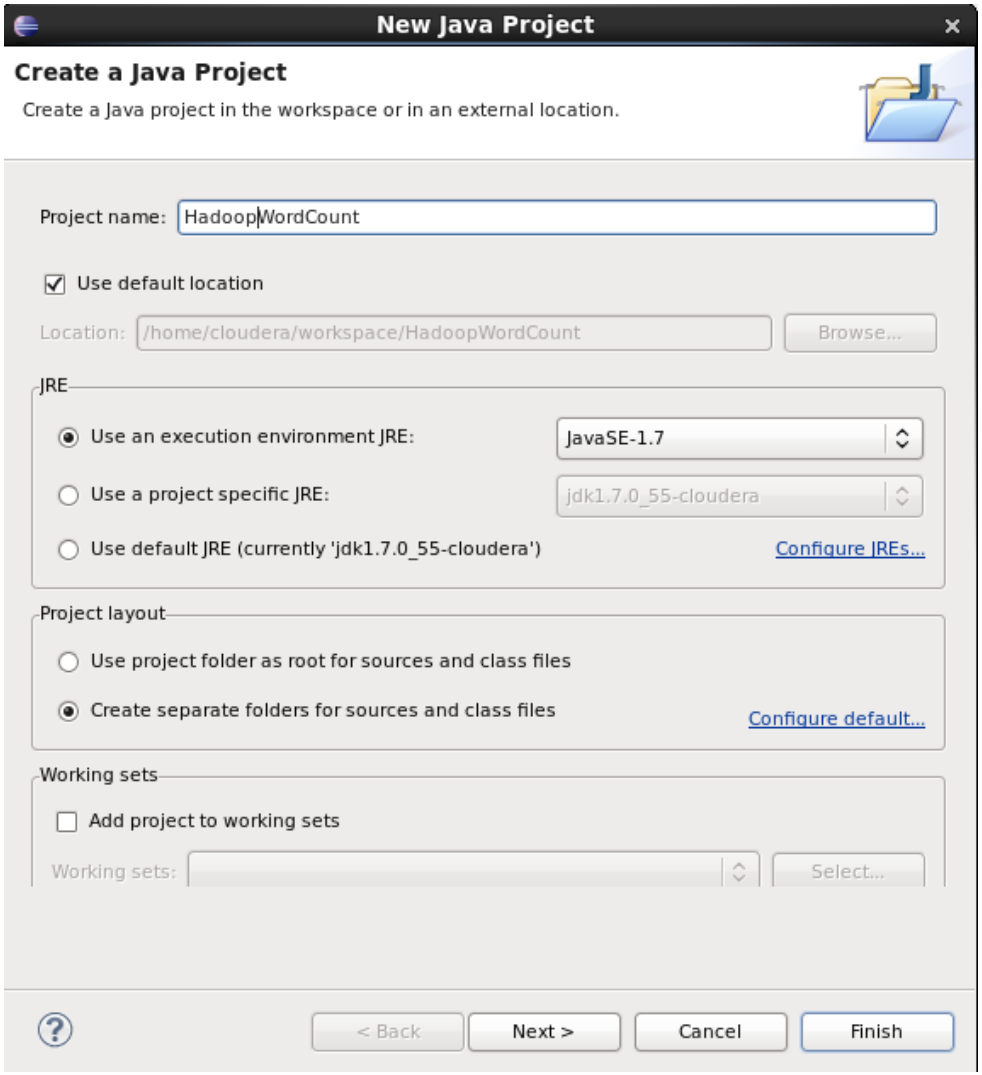
Hands-On: Writing Map/Reduce Program on Eclipse

Starting Eclipse in a local machine



Create a Java Project

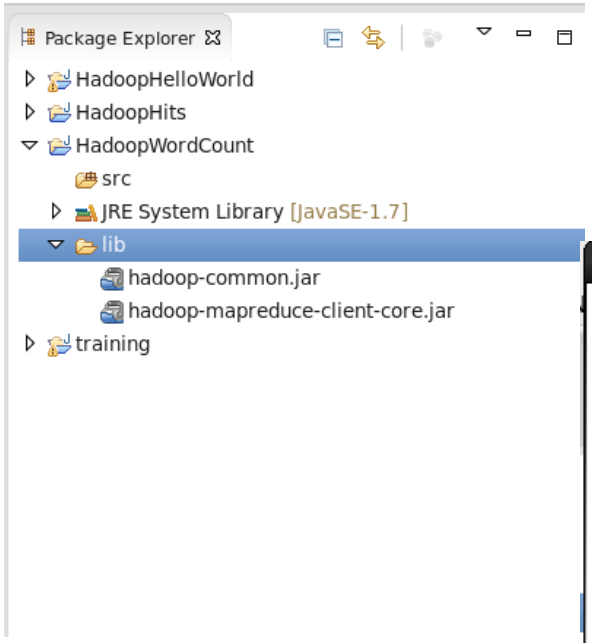
Let's name it HadoopWordCount

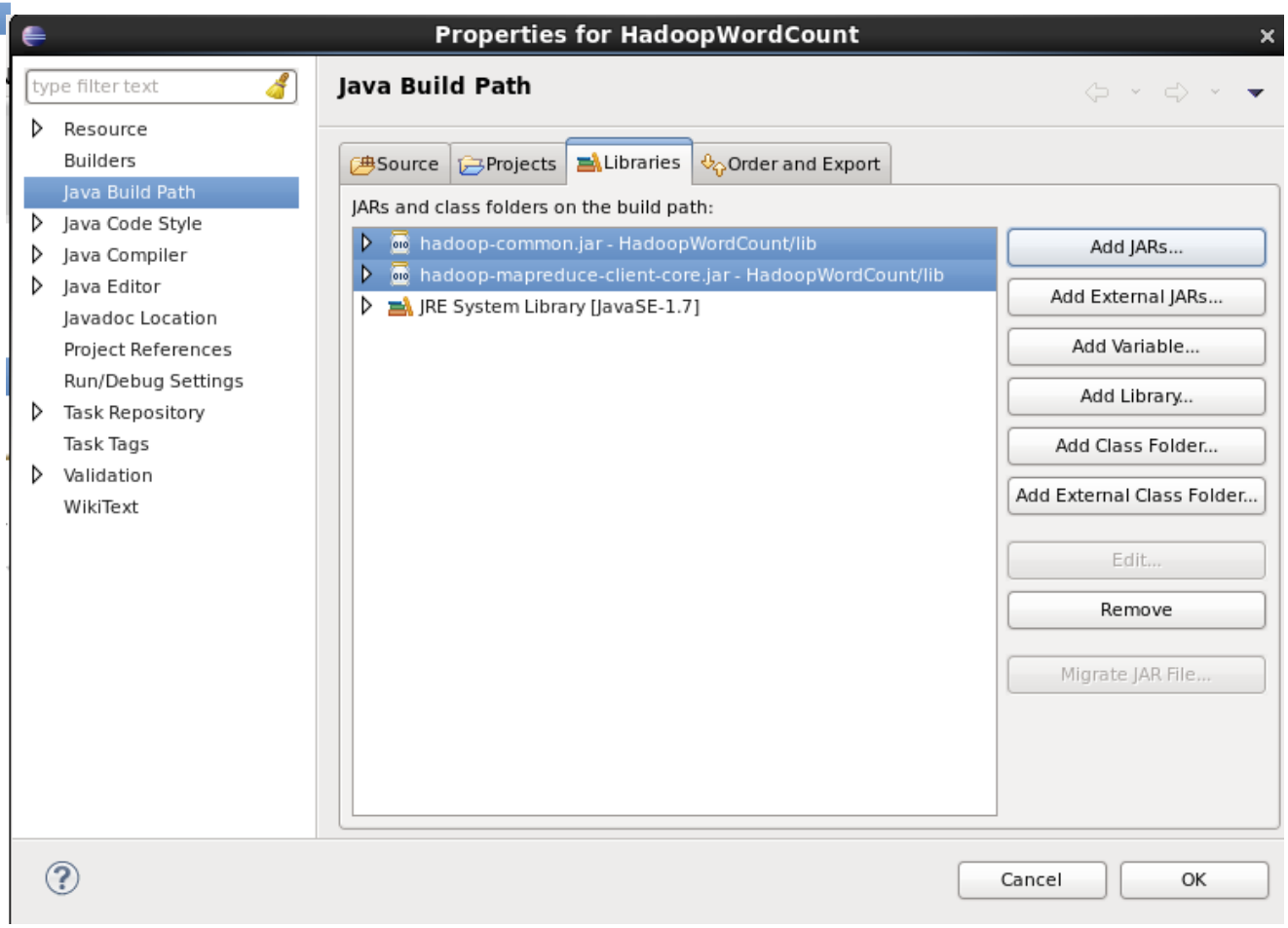


Add dependencies to the project

- 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





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

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, IntWrit

        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, the 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 botton, 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:

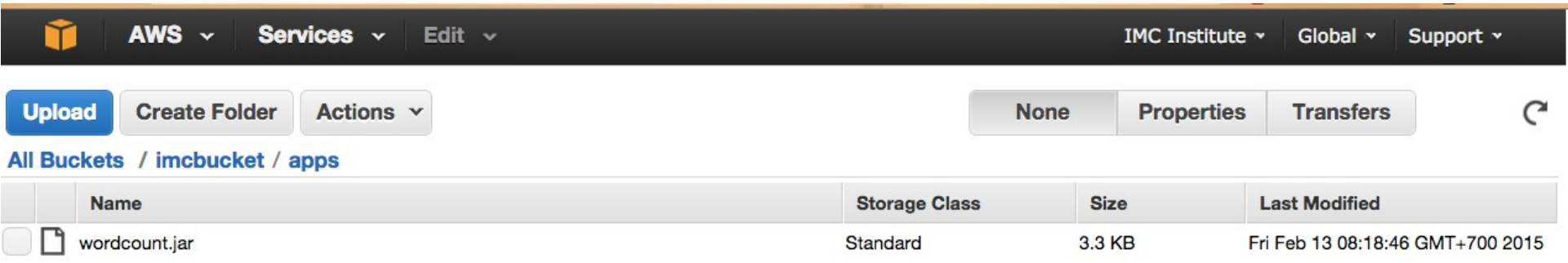
Main class: org.myorg.WordCount — Browse...

< 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



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

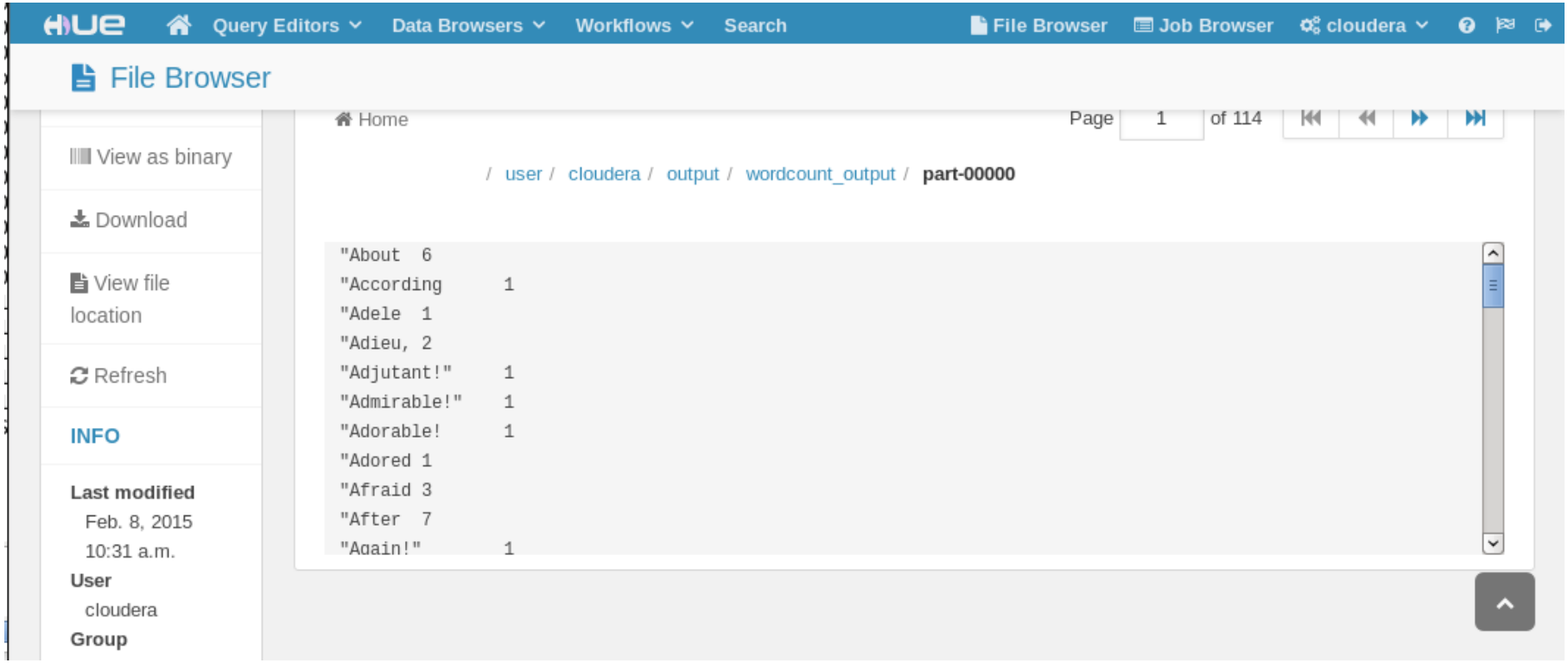
AWS Services Edit IMC Institute Global Support

Upload Create Folder Actions None Properties Transfers

All Buckets / imcbucket / output / wordcount_result

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

Reviewing MapReduce Output Result



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

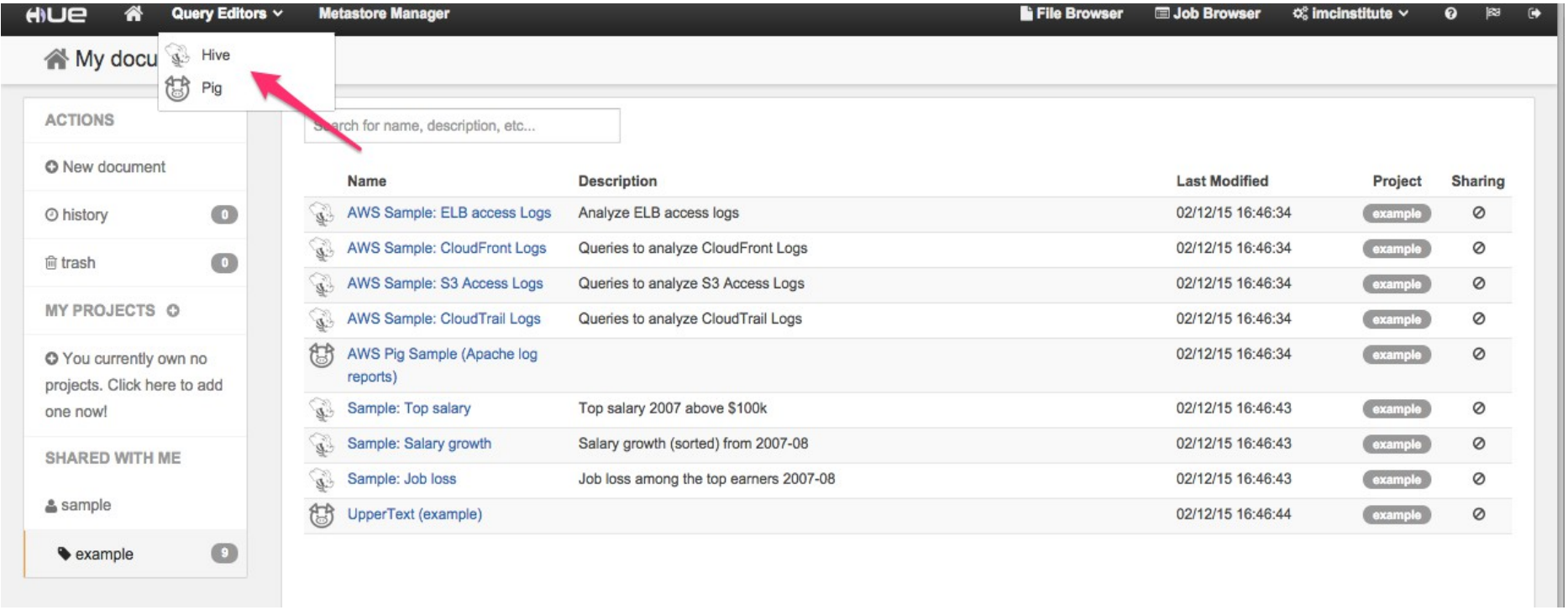
Starting Hive

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

Logging initialized using configuration in jar:file:/home/hadoop/.versions/hive-0.1
3.1/lib/hive-common-0.13.1-amzn-1.jar!/hive-log4j.properties
hive>
```

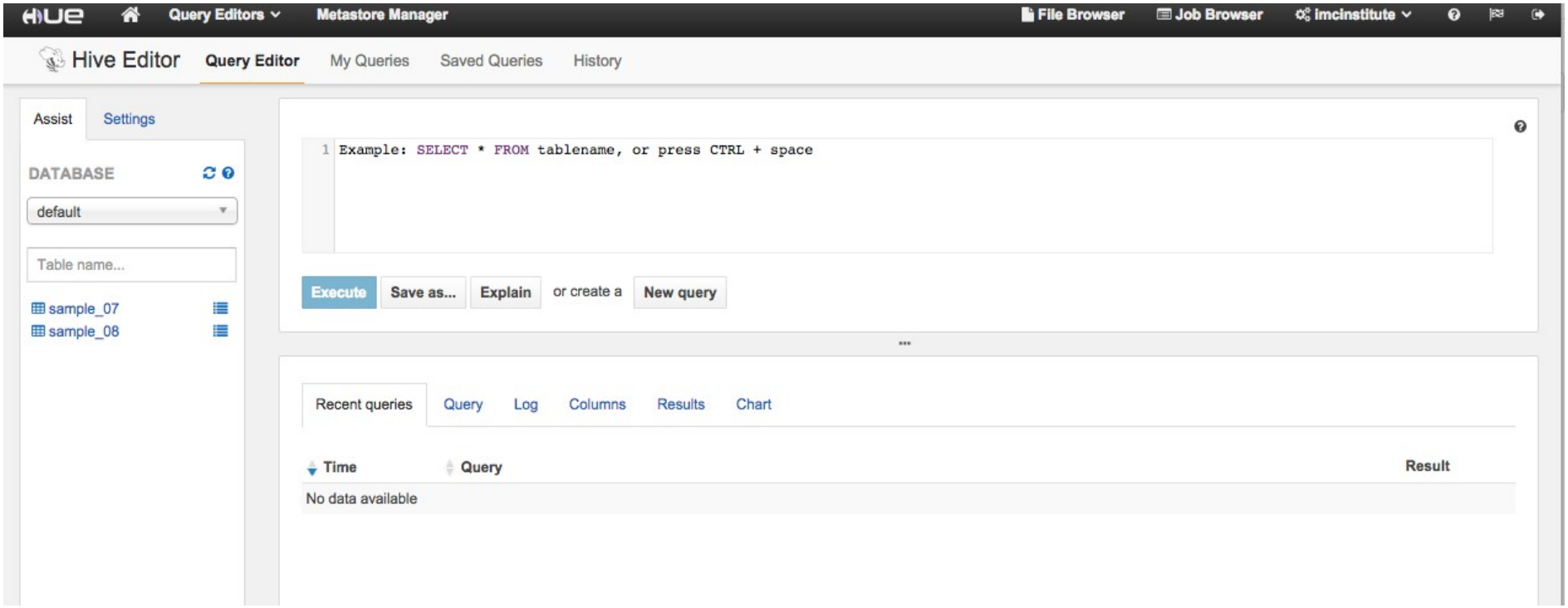
Quit from Hive

```
hive> quit;
```

Starting Hive Editor from Hue



Starting Hive Editor from Hue

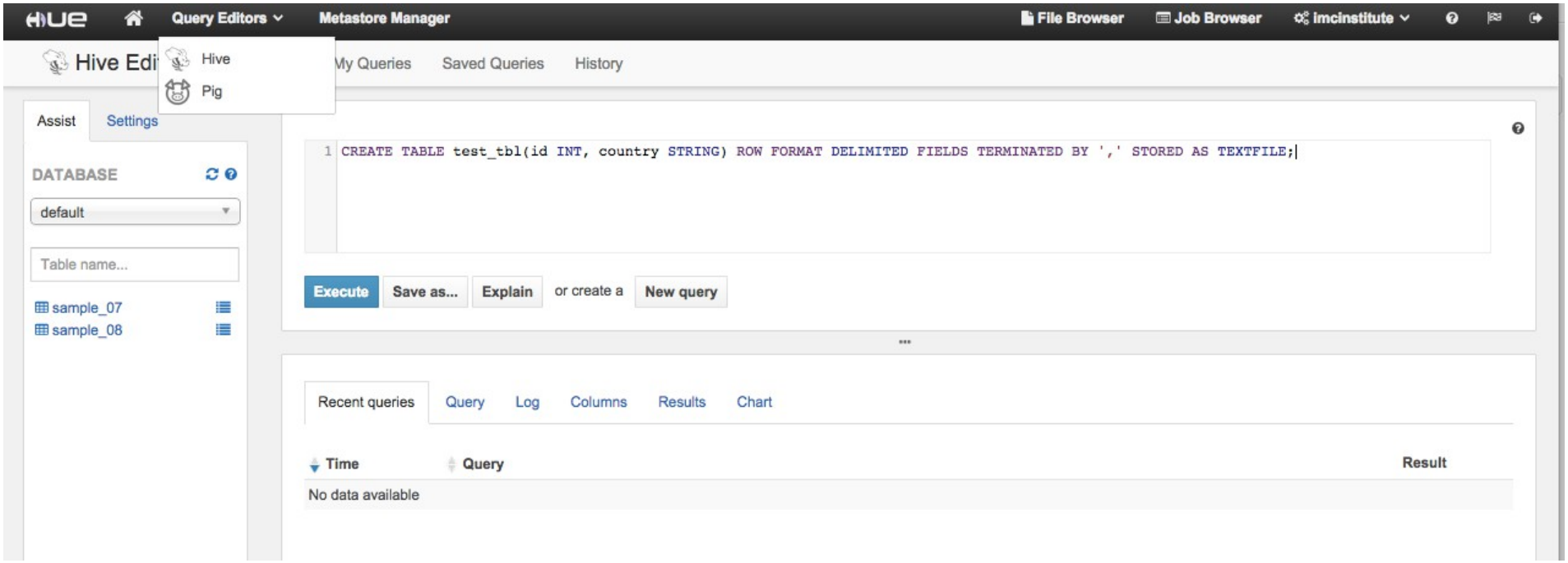


Creating Hive Table

```
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



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

Using Hue Query Editor

Query Editors | Metastore Manager | File Browser | Job Browser | imcinstitute

Hive Editor | Query Editor | My Queries | Saved Queries | History

My Queries

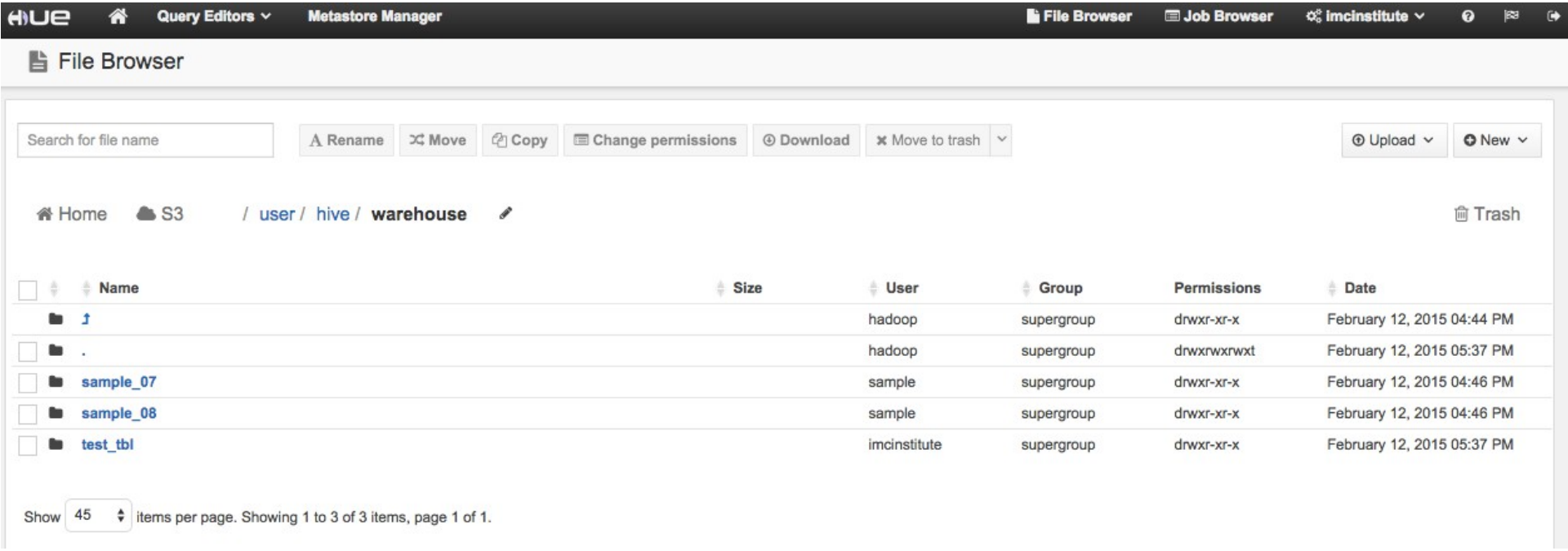
Search for query | View result | Edit | Copy | Usage history | Move to trash | View trash | New query

Recent Saved Queries 0 | Recent Run Queries 1

Time	Name	Query	State
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



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" = "([^]*) ([^]*) ([^]*) (-|\\[^\\]*\\])
([^ \"]*|\"[^\"]*\") (-|[0-9]*) (-|[0-9]*)(?: ([^ \"]*|\".*\")
([^ \"]*|\".*\"))?"
)
STORED AS TEXTFILE;
```

Lecture
Understanding Pig

Introduction

A high-level platform for creating MapReduce programs Using Hadoop

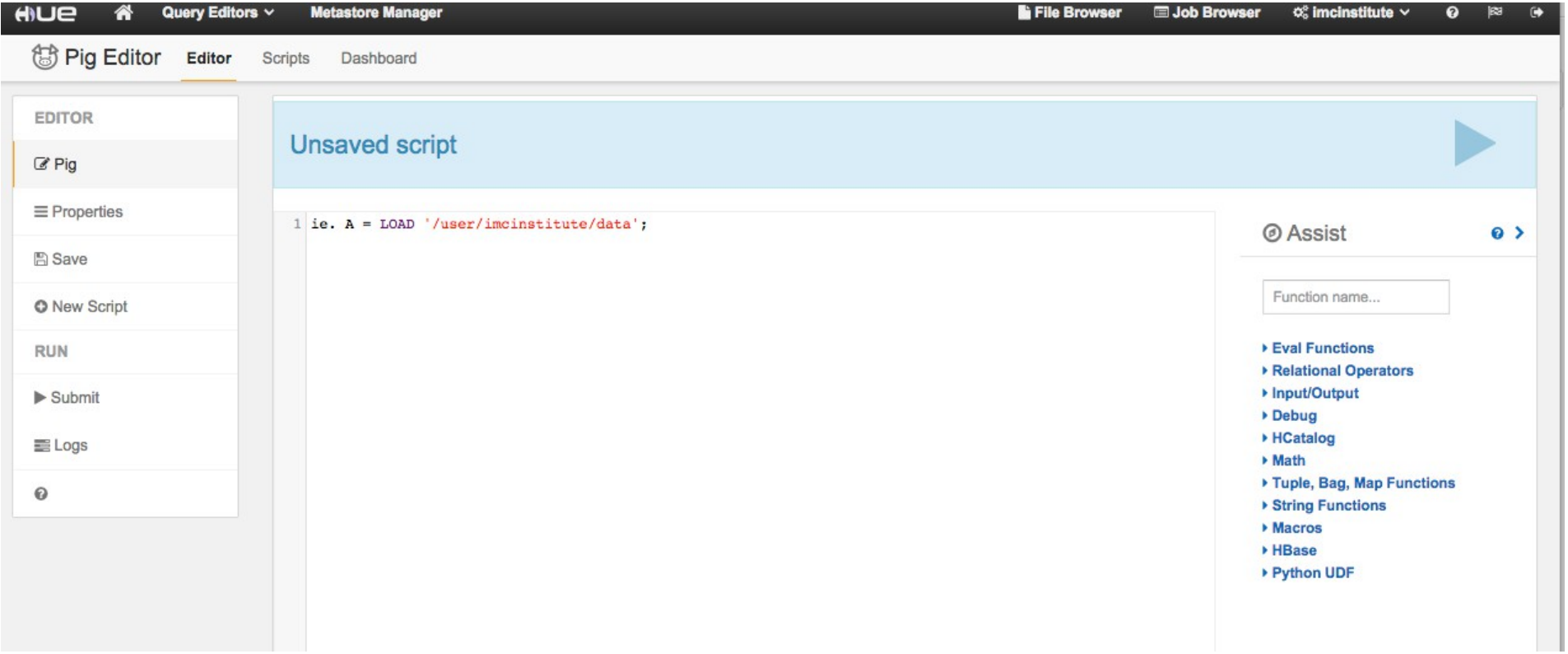
Pig is a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs. The salient property of Pig programs is that their structure is amenable to substantial parallelization, which in turns 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 bo
otup file /home/hadoop/.pigbootup not found
2015-02-13 01:52:33,562 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2015-02-13 01:52:33,562 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2015-02-13 01:52:33,565 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.HExecutionEngine - Connecting to hadoop file system at: file:///
2015-02-13 01:52:33,920 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - 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.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt>
```

Starting Pig from Hue

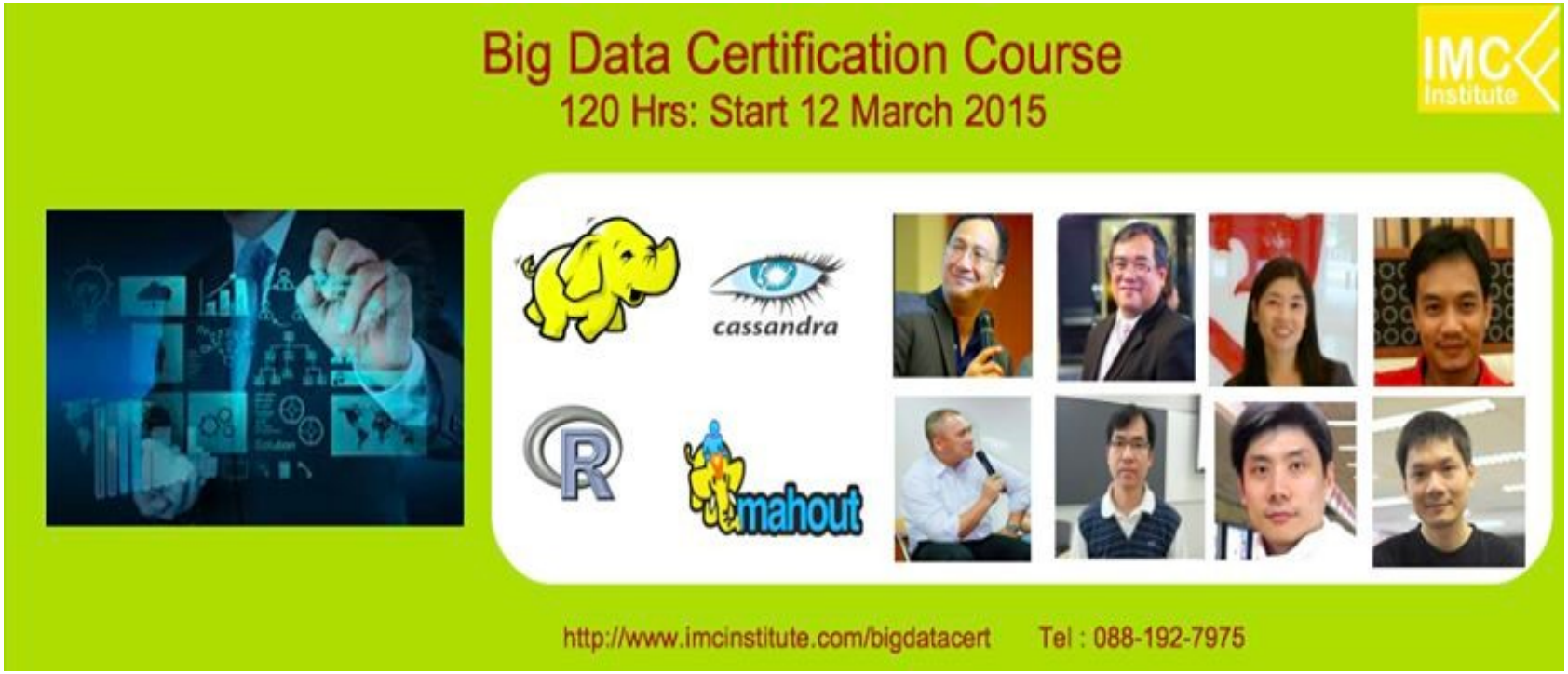


Writing a Pig Script

```
countryFilter.pig

A = load 'hdi-data.csv' using PigStorage(',')  AS (id:int, country:chararray, hdi:float,
lifeex:int, mysch:i
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
IMC Institute
cassandra
mahout
http://www.imcinstitute.com/bigdatacert
Tel : 088-192-7975

Thank you

www.imcinstitute.com
www.facebook.com/imcinstitute