

CSE 414 Database Systems

Section 9: AWS, Hadoop, Pig Latin

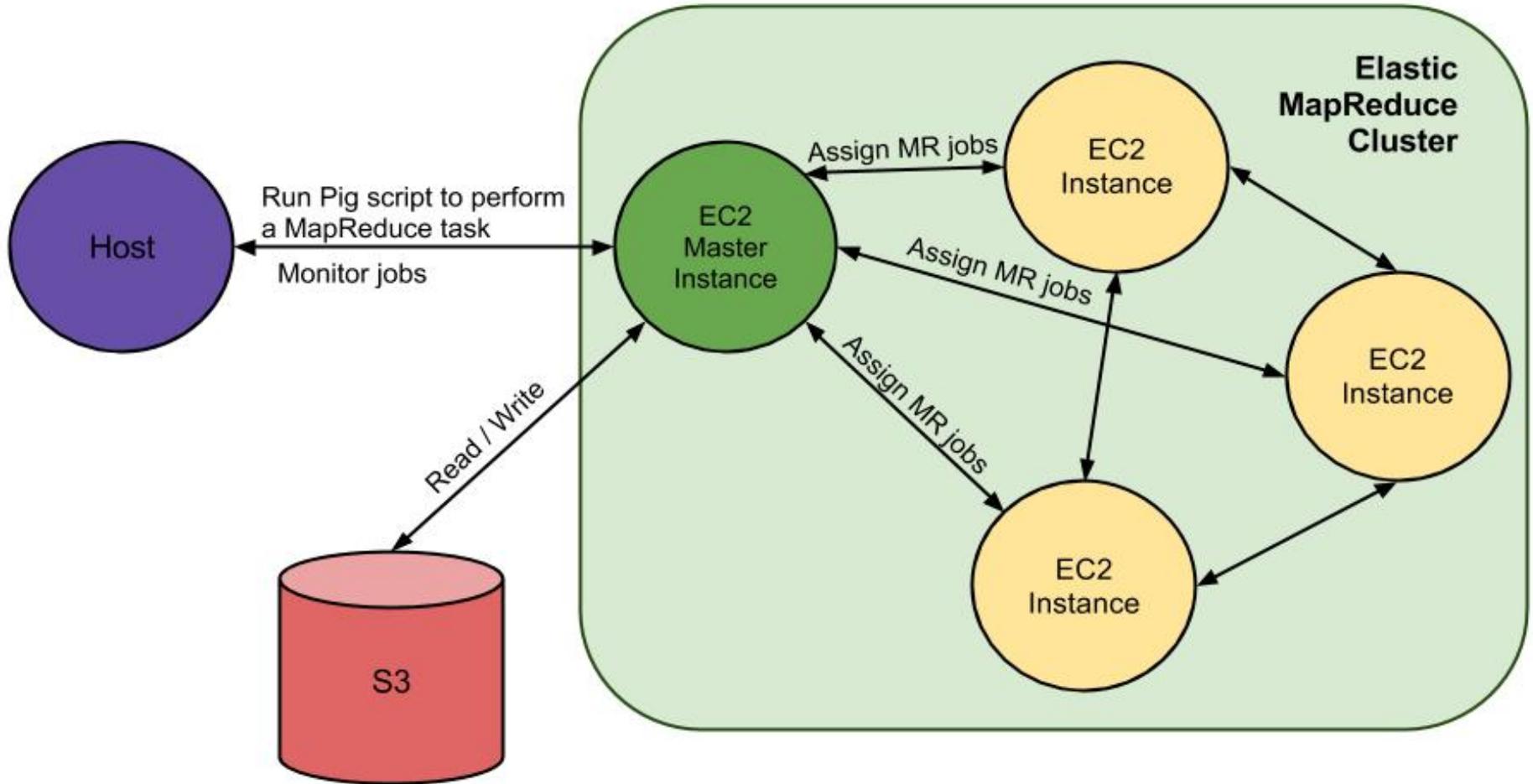
TA: Daseul Lee (dslee@cs)

Homework 8

- Big Data analysis on Amazon Web Service (AWS)
 - Working with up to 0.5 TB of data
 - Billion Triple Set
- Due Friday 6/7
- No late days!

Overview

- AWS offers various cloud computing services. In this assignment, we will use:
 - **Elastic MapReduce**: Managed Hadoop Framework
 - **EC2** (Elastic Computing Cluster): virtual servers in the cloud
 - **S3** (Simple Storage Service): scalable storage in the cloud



Where is your input file?

- Your input files come from Amazon S3
- You will use three sets, each of different size
 - `s3n://uw-cse344-test/cse344-test-file` -- 250KB
 - `s3n://uw-cse344/btc-2010-chunk-000` -- 2GB
 - `s3n://uw-cse344` -- 0.5TB
- See `example.pig` for how to load the dataset

```
raw = LOAD 's3n://uw-cse344-test/cse344-test-file' USING TextLoader as (line:chararray);
```

Where is your output stored?

- Two options

1. Hadoop File System

The AWS Hadoop cluster maintains its own HDFS instance, which dies with the cluster -- this fact is not inherent in HDFS. **Don't forget to copy** them to your local machine before terminating the job.

2. S3

S3 is persistent storage. But S3 costs money while it stores data. **Don't forget to delete** them once you are done.

- It will output a set of files stored under a directory. Each file is generated by a reduce worker to avoid contention on a single output file.

How can you get the output files?

1. Easier and expensive way:

- Create your own S3 bucket(file system), write the output there
- Output filenames become s3n://your-bucket/outdir
- Can download the files via S3 Management Console
- But S3 does cost money, even when the data isn't going anywhere. DELETE YOUR DATA ONCE YOU'RE DONE!

2. Harder and cheapskate way:

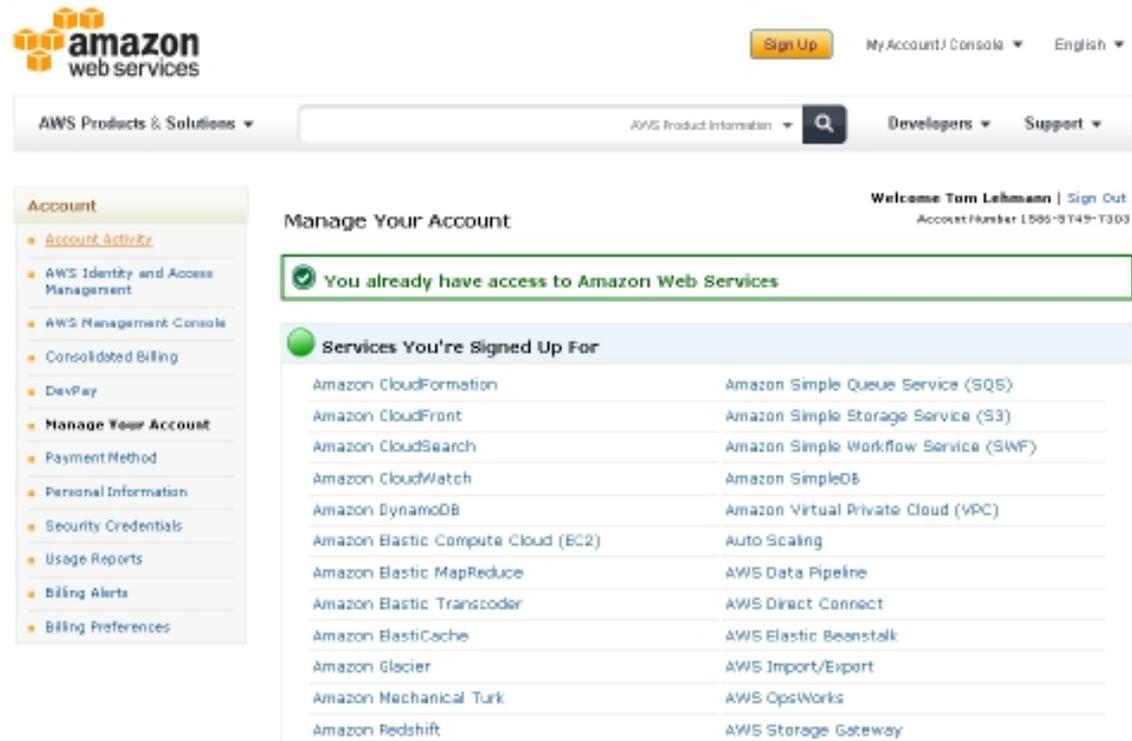
- Write to cluster's HDFS
- Output directory name is /user/hadoop/outdir. You'll need to create /user/hadoop
- Need to double download
 1. from HDFS to master node's filesystem with *hadoop dfs -copyToLocal*
 2. from master node to local machine with *scp*

Set-up

(Disclaimer: Important details are
found in the spec)

Connecting to AWS

- <https://aws.amazon.com/>
- Make sure you are signed up for (1) Elastic MapReduce (2) EC2 (3) S3

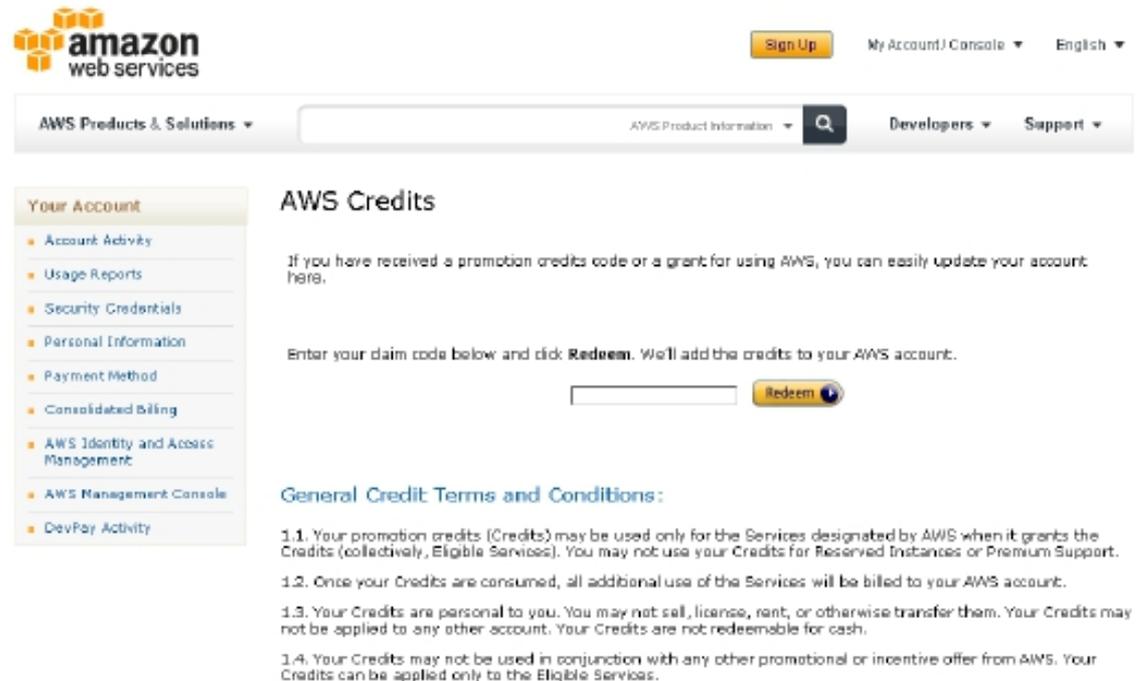


The screenshot displays the AWS Management Console interface. At the top, the Amazon Web Services logo is visible on the left, and navigation links for 'Sign Up', 'My Account / Console', and 'English' are on the right. Below the logo, there's a search bar and links for 'AWS Products & Solutions', 'AWS Product Information', 'Developers', and 'Support'. The main content area is titled 'Manage Your Account' and includes a welcome message for 'Tom Lehmann' with a 'Sign Out' link and account number '1585-9749-7303'. A green notification box states 'You already have access to Amazon Web Services'. Below this, a section titled 'Services You're Signed Up For' lists various AWS services in two columns.

Amazon CloudFormation	Amazon Simple Queue Service (SQS)
Amazon CloudFront	Amazon Simple Storage Service (S3)
Amazon CloudSearch	Amazon Simple Workflow Service (SWF)
Amazon CloudWatch	Amazon SimpleDB
Amazon DynamoDB	Amazon Virtual Private Cloud (VPC)
Amazon Elastic Compute Cloud (EC2)	Auto Scaling
Amazon Elastic MapReduce	AWS Data Pipeline
Amazon Elastic Transcoder	AWS Direct Connect
Amazon ElastiCache	AWS Elastic Beanstalk
Amazon Glacier	AWS Import/Export
Amazon Mechanical Turk	AWS OpsWorks
Amazon Redshift	AWS Storage Gateway

Free Credit

- <https://aws.amazon.com/awscredits/>
- Should have received your AWS credit code by email
- \$100 worth of credits should be enough
- **Don't forget to terminate your job flows!**



The screenshot shows the AWS Credits redemption page. At the top, there is the Amazon Web Services logo and navigation links for 'Sign Up', 'My Account / Console', and 'English'. Below the logo is a search bar and links for 'AWS Products & Solutions', 'AWS Product Information', 'Developers', and 'Support'. On the left side, there is a 'Your Account' sidebar with links to 'Account Activity', 'Usage Reports', 'Security Credentials', 'Personal Information', 'Payment Method', 'Consolidated Billing', 'AWS Identity and Access Management', 'AWS Management Console', and 'DevPay Activity'. The main content area is titled 'AWS Credits' and contains the following text: 'If you have received a promotion credits code or a grant for using AWS, you can easily update your account here.' Below this is a form with a text input field and a 'Redeem' button. Underneath the form is the heading 'General Credit Terms and Conditions:' followed by four numbered terms: 1.1. Your promotion credits (Credits) may be used only for the Services designated by AWS when it grants the Credits (collectively, Eligible Services). You may not use your Credits for Reserved Instances or Premium Support. 1.2. Once your Credits are consumed, all additional use of the Services will be billed to your AWS account. 1.3. Your Credits are personal to you. You may not sell, license, rent, or otherwise transfer them. Your Credits may not be applied to any other account. Your Credits are not redeemable for cash. 1.4. Your Credits may not be used in conjunction with any other promotional or incentive offer from AWS. Your Credits can be applied only to the Eligible Services.

Have AWS create a key pair for you

- Go to EC2 Management Console
- <https://console.aws.amazon.com/ec2/>
- Pick region in navigation bar (top right)
- Click on Key Pairs
- Click Create Key Pair
- Enter name and click Create
- Download of .pem private key - this is needed to access any of your instances

Have AWS create a key pair for you

- People using Windows need to set up PuTTY
- <http://docs.aws.amazon.com/gettingstarted/latest/wah-linux/getting-started-deploy-app-connect.html>
- Everybody else can use this command to change the permission
\$ chmod 600 </path/to/saved/keypair/file.pem>

Starting an AWS cluster

- <http://console.aws.amazon.com/elasticmapreduce/home>
- Click *Amazon Elastic Map Reduce* Tab
- Click *Create New Job Flow*

Create a New Job Flow Cancel

DEFINE JOB FLOW | SPECIFY PARAMETERS | CONFIGURE EC2 INSTANCES | ADVANCED OPTIONS | BOOTSTRAP ACTIONS | REVIEW

Name your job flow and select its type. If you don't have an application to run, use one of our samples to get started.

Job Flow Name*:

Choose a descriptive name for the job flow. It does not have to be unique.

Hadoop Version*:

Create a Job Flow*: Run your own application
 Run a sample application

Run your own application: Select the type of application to run: Hive, Custom JAR, Streaming, Pig or HBase.
Run a sample application: Select the sample application to run.

* Required field

Starting an AWS Cluster

- Name the Job Flow
- Select Pig Program as Job Type
- Select Run your own application
- CONTINUE

Starting an AWS Cluster

- Select Start an Interactive Pig Session
- CONTINUE

Create a New Job Flow Cancel

DEFINE JOB FLOW **SPECIFY PARAMETERS** CONFIGURE EC2 INSTANCES ADVANCED OPTIONS BOOTSTRAP ACTIONS REVIEW

Choose between either executing an existing Pig script or starting an interactive Pig session.

Execute a Pig Script

Run a Pig script which has been uploaded to S3. With this option the job flow starts, automatically executes the script, then terminates the job flow automatically when the script has completed.

Script Location*:
The location of your Pig script in Amazon S3.

Input Location:
The URL of the Amazon S3 Bucket that contains the input files.

Output Location:
The URL of the Amazon S3 Bucket to store output files. Should be unique.

Extra Args:

Start an Interactive Pig Session

Start a job flow with Pig setup for interactive use. Interactive use requires you to have an SSH client to access the master host via the user "hadoop". When you are finished your session, manually terminate the job flow from the list of running jobs.

[< Back](#) [Continue](#) * Required field

Starting an AWS Cluster

- Select only 1 core instance
- CONTINUE
- Set your previously created Key Pair to be the Amazon EC2 Key Pair
- CONTINUE

Starting an AWS Cluster

- Configure your Bootstrap Actions
- Action Type: Memory Intensive Configuration

Configure your Bootstrap Actions

Use the table below to define the name, location and optional arguments for any Bootstrap Actions you want associated with this Job Flow.

Bootstrap Action	
Action Type Choose Bootstrap Action <input type="button" value="Learn More"/>	Optional Arguments
Name <input type="text"/>	<input type="text"/>
Amazon S3 Location <input type="text"/>	

 Add another Bootstrap Action

Starting an AWS Cluster

- CONTINUE
- Create Job Flow
- Refresh page to see your job flow (might take a few minutes...)

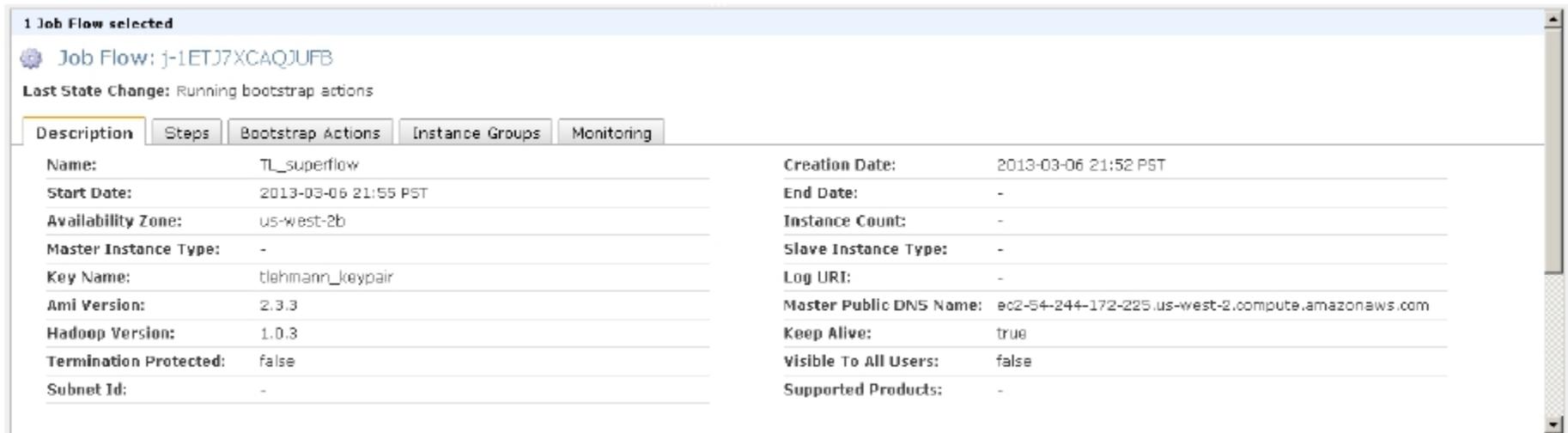


The screenshot shows the AWS Elastic MapReduce console interface. At the top, there's a header "Your Elastic MapReduce Job Flows". Below the header, there are several action buttons: "Create New Job Flow", "Terminate", and "Debug". On the right side, there are "Show/Hide", "Refresh", and "Help" buttons. Below the buttons, there's a "Viewings" dropdown menu set to "All". The main content is a table with the following columns: "Name", "State", "Creation Date", "Elapsed Time", and "Normalized Instance Hours". The table contains one row with the following data: "TL_superflow", "STARTING", "2013-03-06 21:52 PST", "0 hours 0 minutes", and "0".

Name	State	Creation Date	Elapsed Time	Normalized Instance Hours
TL_superflow	STARTING	2013-03-06 21:52 PST	0 hours 0 minutes	0

Starting an AWS Cluster

- Click on your Job Flow
- Retrieve the Master Public DNS Name



1 Job Flow selected

Job Flow: j-1ETJ7XCAQJUB

Last State Change: Running bootstrap actions

Description Steps Bootstrap Actions Instance Groups Monitoring

Name:	TL_superflow	Creation Date:	2013-03-06 21:52 PST
Start Date:	2013-03-06 21:55 PST	End Date:	-
Availability Zone:	us-west-2b	Instance Count:	-
Master Instance Type:	-	Slave Instance Type:	-
Key Name:	tlehmann_keypair	Log URI:	-
Ami Version:	2.3.3	Master Public DNS Name:	ec2-54-244-172-225.us-west-2.compute.amazonaws.com
Hadoop Version:	1.0.3	Keep Alive:	true
Termination Protected:	false	Visible To All Users:	false
Subnet Id:	-	Supported Products:	-

Starting an AWS Cluster

- Windows users use PuTTY to connect to cluster
- Everybody else runs this from command line

```
ssh -o "ServerAliveInterval 10" -i </path/to/saved/keypair/file.pem>  
hadoop@<master.public-dns-name.amazonaws.com>
```

Starting an AWS Cluster

- Type pig, and it will show
grunt>
- Time to write some pig queries!



example.pig

- Found in the project archive
- Loads and parses billion triple dataset
- Triples (subject, predicate, object)
- Group object by attribute, sort in descending order based on count of tuple
- Check out the README for more information

Monitoring Hadoop jobs

Possible options are:

1. Using ssh tunneling
2. Using LYNX
3. Using SOCKS proxy

Terminating Cluster

- Go to Management Console
- Select Job Flow
- Click Terminate
- Wait a few minutes ...
- Eventually status should be

 TERMINATED

Final Comment

- Start early
- Important: read the spec carefully!
If you get stuck or have an unexpected outcome, it is likely that you miss some step or there may be important directions/notes in the spec.
- Running jobs may take up to several hours
 - Extra credit problem takes about ~4 hours.