

CSE 344 Introduction to Data Management

Section 9: AWS, Hadoop, Pig Latin

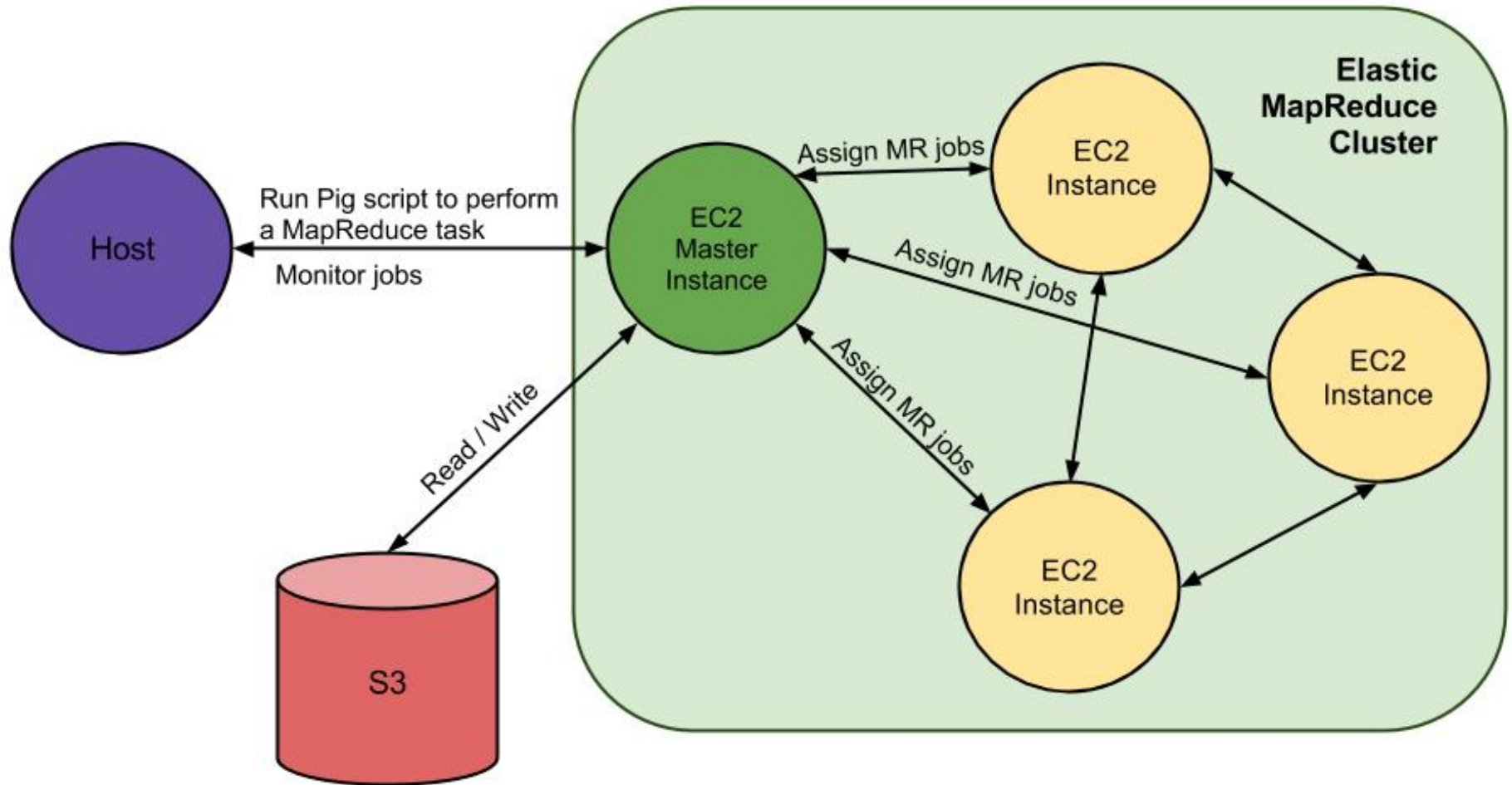
TA: Daseul Lee (dslee@cs)

Homework 8

- Big Data analysis on billion triple dataset using Amazon Web Service (AWS)
 - Billion Triple Set: contains web information, obtained by a crawler
 - (subject, predicate, object)
 - Working with up to 0.5 TB of data
- You will write pig queries for each task and use MapReduce to perform data analysis.
- Due Friday 12/6
- 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:

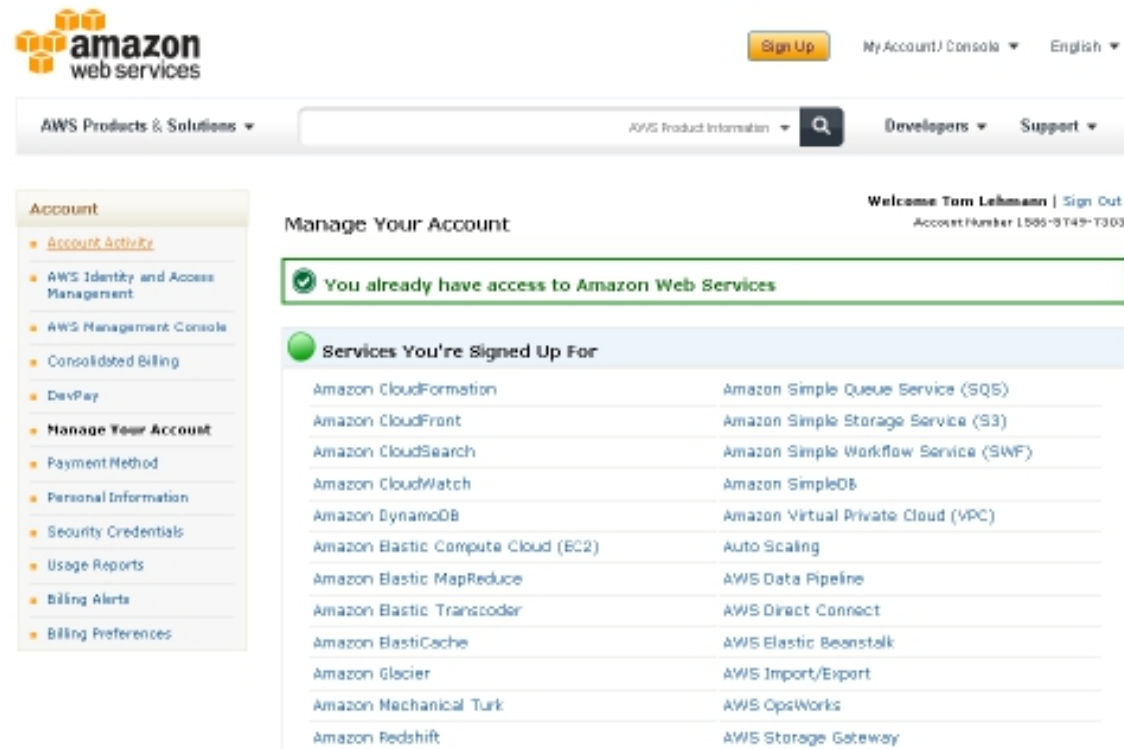
- Write to cluster's HDFS
- Output directory name is /user/hadoop/outdir. You'll need to create /user/hadoop
- Need to double download
 1. from HDFS to master node's filesystem with *hadoop dfs -copyToLocal*
 2. from master node to local machine with scp

Set-up

(Disclaimer: Important details are
found in the spec)

1. Setting up AWS account

- Sign up/in: <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 on the left, and 'Sign Up', 'My Account / Console', and 'English' links are on the right. Below the header, a navigation bar includes 'AWS Products & Solutions', a search bar, and links for 'AWS Product Information', 'Developers', and 'Support'. The left sidebar contains an 'Account' menu with options like 'Account Activity', 'AWS Identity and Access Management', 'AWS Management Console', 'Consolidated Billing', 'DevPay', 'Manage Your Account' (highlighted), 'Payment Method', 'Personal Information', 'Security Credentials', 'Usage Reports', 'Billing Alerts', and 'Billing Preferences'. The main content area, titled 'Manage Your Account', shows a green checkmark and the message '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 CloudFront, Amazon CloudSearch, Amazon CloudWatch, Amazon DynamoDB, Amazon Elastic Compute Cloud (EC2), Amazon Elastic MapReduce, Amazon Elastic Transcoder, Amazon ElastiCache, Amazon Glacier, Amazon Mechanical Turk, Amazon Redshift, Amazon Simple Queue Service (SQS), Amazon Simple Storage Service (S3), Amazon Simple Workflow Service (SWF), Amazon SimpleDB, Amazon Virtual Private Cloud (VPC), Auto Scaling, AWS Data Pipeline, AWS Direct Connect, AWS Elastic Beanstalk, AWS Import/Export, AWS OpsWorks, and AWS Storage Gateway.

amazon web services

Sign Up My Account / Console English

AWS Products & Solutions AWS Product Information Developers Support

Account

- Account Activity
- AWS Identity and Access Management
- AWS Management Console
- Consolidated Billing
- DevPay
- Manage Your Account**
- Payment Method
- Personal Information
- Security Credentials
- Usage Reports
- Billing Alerts
- Billing Preferences

Manage Your Account

Welcome Tom Lehmann | Sign Out
Account Number: 1585-9149-1303

You already have access to Amazon Web Services

Services You're Signed Up For

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

1. Setting up AWS account

- 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 to avoid extra charges!



Sign Up

My Account | Console ▾ English ▾

AWS Products & Solutions ▾

AWS Product Information ▾



Developers ▾

Support ▾

Your Account

- Account Activity
- Usage Reports
- Security Credentials
- Personal Information
- Payment Method
- Consolidated Billing
- AWS Identity and Access Management
- AWS Management Console
- DevPay Activity

AWS Credits

If you have received a promotion credits code or a grant for using AWS, you can easily update your account here.

Enter your claim code below and click **Redeem**. We'll add the credits to your AWS account.

Redeem

General Credit Terms and Conditions:

- 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.

2. Setting up an EC2 key pair

- Go to EC2 Management Console
<https://console.aws.amazon.com/ec2/>
- Pick region in navigation bar (top right)
- Click on *Key Pairs* and click *Create Key Pair*
- Enter name and click *Create*
- Download of .pem private key
 - lets you access EC2 instance
 - Only time you can download the key

2. Setting up an EC2 key pair (Linux/Mac)

- Change the file permission

```
$ chmod 600 </path/to/saved/keypair/file.pem>
```

2. Setting up an EC2 key pair (Windows)

- AWS instruction:
<http://docs.aws.amazon.com/gettingstarted/latest/computebasics-linux/getting-started-deploy-app-connect.html>
- Use PuTTYGen to convert a key pair from .pem to .ppk (part 1 – 2)
- Use PuTTY to establish a connection to EC2 master instance (part 3 – 6)

2. Setting up an EC2 key pair

- Note: Some students were having problem running job flows (next task after setting EC2 key pair) because of no active key found
- If so, go to AWS security credentials page and make sure that you see a key under the access key, if not just click Create a new Access Key.

<https://portal.aws.amazon.com/gp/aws/securityCredentials>

3. Starting an AWS cluster

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

The screenshot shows the 'Create a New Job Flow' wizard in the AWS Management Console. The wizard has a progress bar at the top with five steps: DEFINE JOB FLOW (active), SPECIFY PARAMETERS, CONFIGURE EC2 INSTANCES, ADVANCED OPTIONS, and BOOTSTRAP ACTIONS. Below the progress bar, the text reads: '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.'

The 'Job Flow Name' field is labeled 'Job Flow Name *' and contains the text 'My Job Flow'. Below this field is a note: 'Choose a descriptive name for the job flow. It does not have to be unique.'

The 'Hadoop Version' field is labeled 'Hadoop Version *' and has a dropdown menu showing 'Hadoop 1.0.3 (Amazon Distribution)'. Below this field is a note: 'Choose a descriptive name for the job flow. It does not have to be unique.'

The 'Create a Job Flow' section has two radio buttons: 'Run your own application' (selected) and 'Run a sample application'. Below these is a dropdown menu labeled 'Choose a Job Type'.

On the right side, there is a yellow box with the following text: '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.'

At the bottom right, there is a 'Continue' button with a right arrow. In the bottom right corner, there is a note: '* Required field'.

3. Starting an AWS Cluster

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

3. 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

3. Starting an AWS Cluster

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


3. 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 <div>Choose Bootstrap Action ▾ Learn More</div>	Optional Arguments <div></div>
Name <div></div>	
Amazon S3 Location <div></div>	
<div></div>	

 Add another Bootstrap Action

3. 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 buttons for "Create New Job Flow", "Terminate", and "Debug". On the right side, there are buttons for "Show/Hide", "Refresh", and "Help". Below these buttons, there's a "Viewings" dropdown menu set to "All". To the right of the dropdown, there are navigation arrows and the text "1 to 1 of 1 Job Flows". Below this, there's 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: "Name" is "TL_superflow", "State" is "STARTING", "Creation Date" is "2013-03-06 21:52 PST", "Elapsed Time" is "0 hours 0 minutes", and "Normalized Instance Hours" is "0".

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

3. 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:	-

3. 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>
```

4. Running Pig interactively

- Once you successfully made a connection to EC2 cluster, type pig, and it will show
grunt>
- Time to write some pig queries!



4. Running Pig interactively

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

5. Monitoring Hadoop jobs

Possible options are:

1. Using ssh tunneling (recommended)
2. Using LYNX
3. Using SOCKS proxy

6. 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.