INTRO TO AWS
(Amazon Web Services)

Sept 6, 2012
i290-1, Analyzing Big Data With Twitter

UC Berkeley School of Information
WHAT IS AWS?

- Very Large Shared “Cloud” Computing and Storage Resource
- Pre-Loaded Hadoop Clusters and Pig Environment
- Many Other Tools and Services
USING THE AWS TOOLS SUPPLIED BY THIS CLASS

- This is a PRIVILEGE; we are covering the cost for you.

- It can be EXPENSIVE; we all have to work together to control for this and prevent accidental over expenditures.

- We also will have to control when you have access to the clusters. (We are charged by cluster access time, not cycles.)

- You are only to use the services we describe and for the purposes we describe. You are NOT to use extra time for your own interests. (You can always get your own account.)
RUNNING PIG ON AWS: METHOD 1

• We assign each student or student pair user ID and password just for this class session.

• You can run a script, or run in interactive mode.

• When the script terminates, it stops the charges. But what if your script does not terminate? You need to monitor it.

• When you use interactive mode, you have to terminate the cluster activity yourself.
RUNNING PIG ON AWS: METHOD 2

• We have to do this at a set time so we can turn the clusters on and off.

• We create a cluster for each student or student pair.

• We “launch” that cluster. That means it is active and ready for you to use as much as you like.

• You have to access it via an ssh interface however.

• When the time is up, we terminate the cluster and the charges stop.
RUNNING WITH METHOD 1A
### Innovation.
Powered by Amazon Web Services.

- **Low Cost**
  Pay-as-you-go, no upfront expenses or long-term commitments.

- **Instant Elasticity**
  Instantly deploy your application. Scale resources up or down based on demand.

- **Open & Flexible**
  If it runs in a data center, it can run on AWS. You have full control.

- **Secure**
  Utilize a secure technology platform built and managed by Amazon.

---

### What is AWS?

<table>
<thead>
<tr>
<th>Customer Applications / AWS Marketplace</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deployment &amp; Management</td>
</tr>
<tr>
<td>Application Services</td>
</tr>
<tr>
<td>Foundation Services</td>
</tr>
<tr>
<td>Global Infrastructure</td>
</tr>
</tbody>
</table>

### Cost Savings with AWS

![Cost Savings with AWS Chart]

- **On Demand Capacity with AWS Actual Usage**
- **No Waste**
- **No Customer Dissatisfaction**

Learn more about the AWS Free Tier »

---

### Recent News

**Digital Media in the AWS Cloud**

Los Angeles, CA | September 19

AWS Elastic Resilience Now Available in Oregon

---

AWS.AMAZON.COM
Welcome

The AWS Management Console provides a graphical interface to Amazon Web Services. Learn more about how to use our services to meet your needs, or get started by selecting a service.

Getting started guides
Reference architectures
Free Usage Tier

Set Start Page

Console Home

Amazon Web Services

Compute & Networking
- Direct Connect NEW
  - Dedicated Network Connection to AWS
- EC2
  - Virtual Servers in the Cloud
- Elastic MapReduce
  - Managed Hadoop Framework
- Route 53
  - Scalable Domain Name System
- VPC
  - Isolated Cloud Resources

Deployment & Management
- CloudFormation
  - Templated AWS Resource Creation
- CloudWatch
  - Resource & Application Monitoring
- Elastic Beanstalk
  - AWS Application Container
- IAM
  - Secure AWS Access Control

App Services
- CloudSearch
  - Managed Search Service
- SES
  - Email Sending Service
- SNS
  - Push Notification Service
- SQS
  - Message Queue Service
- SWF
  - Workflow Service for Coordinating Application Components

Storage & Content Delivery
- CloudFront
  - Global Content Delivery Network
- Glacier NEW
  - Archive Storage in the Cloud
- S3
  - Scalable Storage in the Cloud
- Storage Gateway
  - Integrates on-premises IT environments with Cloud storage

Database
- DynamoDB
  - Predictable and Scalable NoSQL Data Store
- ElastiCache
  - In-Memory Caching
Select one of your buckets to the left to look at the objects it contains, or to upload objects into it.
THIS IS AN EXAMPLE: USE YOUR OWN BUCKET TO AVOID CONFLICTS
REMEMBER NOT TO WRITE TO THE SAME DIRECTORY TWICE.

```pig
sentences = LOAD 's3://tucson2/tucson.txt' USING PigStorage() AS (sentence:chararray);
hers = FILTER sentences BY sentence MATCHES '* her .*';
STORE hers INTO 's3://tucson2/tucson_sentences_with_hers';
```
S3 CONSOLE
Scripture tells us, "There is a river whose streams make glad the city of God, the holy place where the most high dwells. God is within her, she will not fall; God will help her at break of day."

On Saturday morning, Gabby, her staff, and many of her constituents gathered outside a supermarket to exercise their right to peaceful assembly and free speech. George and Dorothy Morris -- "Dot" to her friends -- were high school sweethearts who got married and had two daughters. They did everything together, traveling the open road in their R.V., enjoying what their friends called a 50-year honeymoon.

A New Jersey native, Phyllis Schneck retired to Tucson to beat the snow. But in the summer, she would return east, where her world revolved around her three children, her seven grandchildren, and two-year-old great-granddaughter. A gifted quilter, she'd often work under her favorite tree, or sometimes she'd sew aprons with the logos of the Jets and the Giants...

...to give out at the church where she volunteered. A Republican, she took a liking to Gabby and wanted to get to know her better.

...but his true passion was helping people. As Gabby's outreach director, he made the care of thousands of her constituents his own, seeing to it that seniors got the Medicare benefits that they had earned, that veterans got the medals and the care that they deserved, that government was working for ordinary folks.

And then there is nine-year-old Christina-Taylor Green. Christina was an A student. She was a dancer. She was a gymnast. She was a swimmer. She decided that she wanted to be the first woman to play in the Major Leagues, and as the only girl on her Little League team, no one put it past her.

She showed an appreciation for life uncommon for a girl her age. She'd remind her mother, "We are so blessed. We have the best life." And she'd pay those blessings back by participating in a charity that helped children who were less fortunate.

And I want to tell you -- her husband, Mark, is here, and he allows me to share this with you. Right after we went to visit, a few minutes after we left her room and some of her colleagues from Congress were in the room, Gabby opened her eyes for the first time. Gabby opened her eyes for the first time.
S3 Console
WARNING: BUCKET NAMES MUST USE LOWER CASE LETTERS
Welcome

The AWS Management Console provides a graphical interface to Amazon Web Services. Learn more about how to use our services to meet your needs, or get started by selecting a service.

Getting started guides
Reference architectures
Free Usage Tier

Set Start Page

Amazon Web Services

Compute & Networking
- Direct Connect
- EC2
- Elastic MapReduce
- Route 53
- VPC

Deployment & Management
- CloudFormation
- CloudWatch
- Elastic Beanstalk
- IAM

Storage & Content Delivery
- CloudFront
- Glacier
- S3
- Storage Gateway

App Services
- CloudSearch
- SES
- SNS
- SQS
- SWF

Database
- DynamoDB
- ElastiCache

MANAGEMENT CONSOLE
<table>
<thead>
<tr>
<th>Name</th>
<th>State</th>
<th>Creation Date</th>
<th>Elapsed Time</th>
<th>Normalized Instance Hours</th>
</tr>
</thead>
<tbody>
<tr>
<td>tucson pig interactive</td>
<td>TERMINATED</td>
<td>2012-09-05 23:37 PDT</td>
<td>0 hours 12 minutes</td>
<td>3</td>
</tr>
<tr>
<td>tucson pig 3</td>
<td>TERMINATED</td>
<td>2012-09-05 22:11 PDT</td>
<td>0 hours 10 minutes</td>
<td>3</td>
</tr>
<tr>
<td>Tucson Pig 2</td>
<td>FAILED</td>
<td>2012-09-05 15:45 PDT</td>
<td>0 hours 4 minutes</td>
<td>3</td>
</tr>
<tr>
<td>My Job Flow</td>
<td>FAILED</td>
<td>2012-09-05 15:41 PDT</td>
<td>0 hours 0 minutes</td>
<td>0</td>
</tr>
<tr>
<td>My Job Flow test 1</td>
<td>TERMINATED</td>
<td>2012-09-05 12:57 PDT</td>
<td>1 hour 29 minutes</td>
<td>6</td>
</tr>
<tr>
<td>111</td>
<td>FAILED</td>
<td>2012-09-05 02:30 PDT</td>
<td>0 hours 0 minutes</td>
<td>0</td>
</tr>
<tr>
<td>pigExample3</td>
<td>FAILED</td>
<td>2012-08-31 02:19 PDT</td>
<td>0 hours 3 minutes</td>
<td>3</td>
</tr>
<tr>
<td>pigExample2</td>
<td>FAILED</td>
<td>2012-08-30 15:33 PDT</td>
<td>0 hours 3 minutes</td>
<td>3</td>
</tr>
<tr>
<td>pigExamplple1</td>
<td>FAILED</td>
<td>2012-08-30 14:42 PDT</td>
<td>0 hours 3 minutes</td>
<td>3</td>
</tr>
</tbody>
</table>
## EMR Console

### Your Elastic MapReduce Job Flows

**Region:** US East (Virginia)

<table>
<thead>
<tr>
<th>Name</th>
<th>State</th>
<th>Creation Date</th>
<th>Elapsed Time</th>
<th>Normalized Instance Hours</th>
</tr>
</thead>
<tbody>
<tr>
<td>tucson pig interactive</td>
<td>TERMINATED</td>
<td>2012-09-05 23:37 PDT</td>
<td>0 hours 12 minutes</td>
<td>3</td>
</tr>
<tr>
<td>tucson pig 3</td>
<td>TERMINATED</td>
<td>2012-09-05 22:11 PDT</td>
<td>0 hours 10 minutes</td>
<td>3</td>
</tr>
</tbody>
</table>

### 1 Job Flow selected

**Job Flow:** j-30FF6EONHJTH9

**Last State Change Reason:** Terminated by user request

### Job Flow: j-30FF6EONHJTH9

**Description**

<table>
<thead>
<tr>
<th>Name</th>
<th>tucson pig 3</th>
<th>Creation Date:</th>
<th>2012-09-05 22:11 PDT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start Date:</td>
<td>2012-09-05 22:17 PDT</td>
<td>End Date:</td>
<td>2012-09-05 22:26 PDT</td>
</tr>
<tr>
<td>Availability Zone:</td>
<td>us-east-1a</td>
<td>Instance Count:</td>
<td>-</td>
</tr>
<tr>
<td>Master Instance Type:</td>
<td>-</td>
<td>Slave Instance Type:</td>
<td>-</td>
</tr>
<tr>
<td>Key Name:</td>
<td>-</td>
<td>Log URI:</td>
<td>s3n://tucson2/log/</td>
</tr>
<tr>
<td>Ami Version:</td>
<td>latest</td>
<td>Master Public DNS Name:</td>
<td>ec2-174-129-129-0.compute-1.amazonaws.com</td>
</tr>
<tr>
<td>Hadoop Version:</td>
<td>1.0.3</td>
<td>Keep Alive:</td>
<td>true</td>
</tr>
<tr>
<td>Termination Protected:</td>
<td>false</td>
<td>Subnet Id:</td>
<td>-</td>
</tr>
<tr>
<td>Supported Products:</td>
<td>-</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## EMR Console

### Your Elastic MapReduce Job Flows

<table>
<thead>
<tr>
<th>Name</th>
<th>State</th>
<th>Creation Date</th>
<th>Elapsed Time</th>
<th>Normalized Instance Hours</th>
</tr>
</thead>
<tbody>
<tr>
<td>tucson pig interactive</td>
<td>TERMINATED</td>
<td>2012-09-05 23:37 PDT</td>
<td>0 hours 12 minutes</td>
<td>3</td>
</tr>
<tr>
<td>tucson pig 3</td>
<td>TERMINATED</td>
<td>2012-09-05 22:11 PDT</td>
<td>0 hours 10 minutes</td>
<td>3</td>
</tr>
</tbody>
</table>

### Job Flow: j-30FF6EONHJTH9

**Last State Change Reason:** Terminated by user request

### Steps

<table>
<thead>
<tr>
<th>Step Name</th>
<th>State</th>
<th>Start Date</th>
<th>End Date</th>
<th>JAR</th>
<th>Main Class</th>
<th>Args</th>
</tr>
</thead>
</table>
EMR CONSOLE
I. Setting up Elastic MapReduce

1. You will receive a user ID, password, and private key file(*.pem). You MUST keep it securely. Do NOT share it with others.

2. Log in to the amazon URL you are supplied with, and enter the user ID and password you were given.

3. Go to the Elastic MapReduce Console Menu and click Create New Job Flow.
I. Setting up The Job Flow

4. Create a Job Flow Name, and Choose Pig Program at Job Type. Then continue.
I. Interactive or Script (Batch) Mode?

5a. Now you can choose either to run a script or to run an interactive session. If you are running a script, you need to be sure that the file and directory names in the script refer to the right places in the S3 file system. For the sample script shown above, we use the bucket s3://tucson2/ and make all our file and directory names relative to that.

THIS IS AN EXAMPLE: USE YOUR OWN_BUCKET TO AVOID CONFLICTS
REMEMBER NOT TO WRITE TO THE SAME DIRECTORY TWICE.
I. Interactive or Script (Batch) Mode?

5b. To choose interactive mode, select “Start an Interactive Pig Session”. Then continue.
I. Configure instance type

6. The instance type is the type of cluster. Bigger instances cost more money! NEVER choose a type larger than Large and LEAVE COUNT SET TO 2.
I. Select the key pair

7. This is only needed for interactive mode. Enter your private key file from Step 1. If you want a log directory, enter a log path with the s3 pathname, and Enable Debugging. Then continue.
I. No bootstrap.

8. Choose No bootstrap, then Continue.
I. Run the job!

9. Now you'll see the summary of your settings, and if all looks correct, start the job flow!
II. Running the Pig Script.

1. Wait for the server to start. You can watch it in the EMR console.

2. When the state changes to Waiting, click the Job Flow, and copy Master Public DNS Name from Description.
II. Running the Pig Script.

3. Now, back on your own computer, open a shell. (For example, terminal on mac, putty on windows.) You can now access the instance using ssh and the Master Public DNS Name from Step II.2 above, using your private key file and the user ID “hadoop”.

II. Running the Pig Script.

4. Now, you can run the script from your machine, on the AWS cluster.

$ pig -p INPUT=s3://<your-bucket>/<your-file-name> -p OUTPUT=s3://<your-bucket>/<output-file-name> s3://<your-bucket>/<your-script-name>
III. Terminating the Interactive Session (IMPORTANT!)

1. It is very important that you terminate your interactive session, so we are not charged for time that is not being used!!!

When you finish running the job flow, terminate the Job Flow from the EMR console.
USING THE AWS TOOLS SUPPLIED BY THIS CLASS

• This is a PRIVILEGE; we are not making you pay for it.

• It can be EXPENSIVE; we all have to work together to control for this and prevent accidental over expenditures.

• We also will have to control when you have access to the clusters. (We are charged by cluster access time, not cycles.)

• You are only to use the services we describe and for the purposes we describe. You are NOT to use extra time for your own interests. (You can always get your own account.)