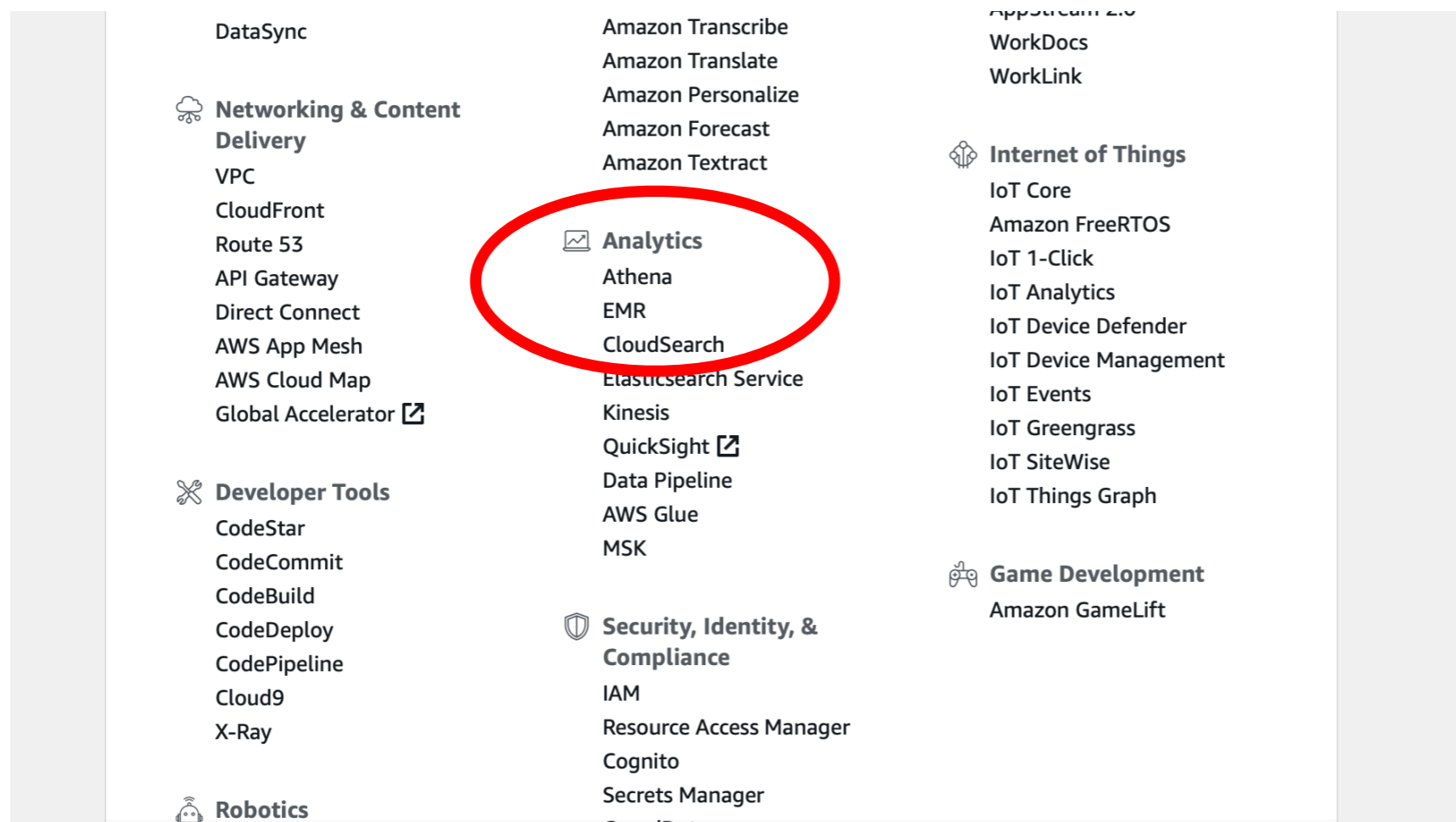


Using Amazon EMR

Data at Scale

Create a Cluster

- In AWS Management Console:
 - Select EMR under Analytics



Create a Cluster

- EMR gives you an overview
 - This will give you an overview of all previous clusters and their status

The screenshot shows the AWS Management Console interface for Amazon EMR. The left sidebar contains navigation options: Amazon EMR, Clusters, Security configurations, VPC subnets, Events, Notebooks, Help, and What's new. The main content area displays a table of clusters. At the top, there is a notification about using AWS Glue Data Catalog as an external Hive metastore. Below the notification are buttons for 'Create cluster', 'View details', 'Clone', and 'Terminate'. The table has a filter set to 'All clusters' and shows 4 clusters (all loaded). The table columns are: Name, ID, Status, Creation time (UTC-5), Elapsed time, and Normalized instance hours. The clusters listed are:

Name	ID	Status	Creation time (UTC-5)	Elapsed time	Normalized instance hours
My cluster	j-2LQPYLI63G9NW	Terminated with errors Instance failure	2019-03-11 23:25 (UTC-5)	20 hours	240
My cluster	j-23D4JM5T5RX04	Terminated with errors Validation error	2019-03-11 23:22 (UTC-5)	1 minute	0
My cluster	j-2EVXIAFL6PB1L	Terminated with errors Validation error	2019-03-11 23:01 (UTC-5)	3 minutes	0
My cluster	j-2T9BCSEUWQ3AG	Terminated User request	2019-03-11 22:23 (UTC-5)	15 minutes	24

Create a Cluster

- Create a cluster
 - The default will do
 - Choose your EC2 key pair

Create a Cluster

Amazon EMR

Clusters

Security configurations

VPC subnets

Events

Notebooks

Help

What's new

Clone Terminate AWS CLI export

Cluster: My cluster **Starting**

Summary Application history Monitoring Hardware Configurations Events Steps Bootstrap actions

Connections: --
Master public DNS: --
Tags: -- [View All / Edit](#)

Summary

ID: j-1LFPFEJQXN3J5
Creation date: 2019-04-09 11:52 (UTC-5)
Elapsed time: 0 seconds
Auto-terminate: No
Termination protection: Off [Change](#)

Network and hardware

Availability zone: --
Subnet ID: [subnet-cadbae96](#)
Master: Provisioning 1 m3.xlarge
Core: Provisioning 2 m3.xlarge
Task: --

Configuration details

Release label: emr-5.23.0
Hadoop distribution: Amazon 2.8.5
Applications: Ganglia 3.7.2, Hive 2.3.4, Hue 4.3.0, Mahout 0.13.0, Pig 0.17.0, Tez 0.9.1
Log URI: s3://aws-logs-842628122241-us-east-1/elasticmapreduce/
EMRFS consistent view: Disabled
Custom AMI ID: --

Security and access

Key name: nvirginia
EC2 instance profile: EMR_EC2_DefaultRole
EMR role: EMR_DefaultRole
Visible to all users: All [Change](#)

Info This feature will be deprecated soon.

Security groups for Master:

Security groups for Core & Task:



Create a Cluster

- Your cluster status is shown as provisioning
 - You need to wait (and refresh page) periodically
 - You can also look in the Hardware Tab

Amazon EMR

Clusters

Security configurations

VPC subnets

Events

Notebooks

Help

What's new

Clone Terminate AWS CLI export

Cluster: My cluster **Starting**

Summary Application history Monitoring Hardware Configurations Events Steps Bootstrap actions

Add task instance group

Instance groups

Filter: 2 instance groups (all loaded)

ID	Status	Node type & name	Instance type	Instance count
ig-2UQ0YZ00Z523W	Provisioning (2 Requested)	CORE Core Instance Group	m3.xlarge 8 vCore, 15 GiB memory, 80 SSD GB storage EBS Storage: none	0 Instances
ig-3LATA2U6QG2U3	Provisioning (1 Requested)	MASTER Master Instance Group	m3.xlarge 8 vCore, 15 GiB memory, 80 SSD GB storage EBS Storage: none	0 Instances

Create a Cluster

Termination protection: Off [Change](#)

Log URI: s3://aws-logs-842628122241-us-east-1/elasticmapreduce/ 

EMRFS consistent view: Disabled

Custom AMI ID: --

Network and hardware

Availability zone: us-east-1d

Subnet ID: [subnet-cadbae96](#) 

Master: **Running** 1 m3.xlarge

Core: **Running** 2 m3.xlarge

Task: --


Security and access

Key name: nvirginia

EC2 instance profile: EMR_EC2_DefaultRole

EMR role: EMR_DefaultRole

Visible to all users: All [Change](#)

 This feature will be deprecated soon.

Security groups for [sg-0588c2af5de317452](#) 

Master: (ElasticMapReduce-master)

Security groups for [sg-0b7b1f7a1c9b79ceb](#) 

Core & Task: (ElasticMapReduce-slave)

Connect to your cluster

- You now can connect to your instance
- Click the “connect tab” on your master in order to connect
 - Open ssh
 - ```
ssh -i "nvirginia.pem"
hadoop@ec2-3-86-68-146.compute-1.amazonaws.com
```
  - The name of your instance is of course different



# Connect to your Cluster

```
[Peter-Canisius:Documents thomasschwarz$ ssh -i ~/ssh_open/nvirginia.pem hadoop@ec2-34-229-103-137.compute-1.amazonaws.com
The authenticity of host 'ec2-34-229-103-137.compute-1.amazonaws.com (34.229.103.137)' can't be established.
ECDSA key fingerprint is SHA256:Z2D7dVEn+e0e4mjvtTqK/89xh4M5RakQ/n+iDrImVg.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added 'ec2-34-229-103-137.compute-1.amazonaws.com,34.229.103.137' (ECDSA) to the list of known hosts.
Last login: Tue Apr 9 20:52:26 2019
```

```
 _ | _ | _)
 _ | (_ | / Amazon Linux AMI
 _ | \ _ | _ |
```

```
https://aws.amazon.com/amazon-linux-ami/2018.03-release-notes/
7 package(s) needed for security, out of 12 available
Run "sudo yum update" to apply all updates.
```

```
EEEEEEEEEEEEEEEEEEEE MMMMMMM MMMMMMM RRRRRRRRRRRRRRR
E::::::::::::::::::::E M::::::::M M::::::::M R:::::::::R
EE::::::::EEEEEEEE::::E M::::::::M M::::::::M R::::RRRRRR:::::R
 E::::E EEEEE M::::::::M M::::::::M RR::::R R::::R
 E::::E M:::::M:::M M:::M:::::M R:::R R::::R
 E:::::EEEEEEEEEE M:::::M M:::M M:::M M:::::M R:::RRRRRR:::::R
 E:::::::::::::E M:::::M M:::M:::M M:::::M R:::::::::RR
 E:::::EEEEEEEEEE M:::::M M:::::M M:::::M R:::RRRRRR:::::R
 E::::E M:::::M M:::M M:::::M R:::R R::::R
 E::::E EEEEE M:::::M MMM M:::::M R:::R R::::R
EE:::::EEEEEEEE::::E M:::::M M:::::M R:::R R::::R
E:::::::::::::E M:::::M M:::::M RR::::R R::::R
EEEEEEEEEEEEEEEEEEEE MMMMMMM MMMMMMM RRRRRRR RRRRRR
```

# Running Pig

- Type
  - `pig -x mapreduce`
  - Don't use just "pig"
  - You get a number of messages
  - And then the grunt prompt

```
thomasschwarz — ec2-user@ip-172-31-47-53:~/Data — ssh -i ssh_open/nvirg...
[[ec2-user@ip-172-31-47-53 ~]$ cd Data
[[ec2-user@ip-172-31-47-53 Data]$ ls
[[ec2-user@ip-172-31-47-53 Data]$ ls -la
total 3744
drwxrwxr-x 2 ec2-user ec2-user 4096 Apr 9 17:10 .
drwx----- 4 ec2-user ec2-user 4096 Apr 9 17:16 ..
-rw----- 1 ec2-user ec2-user 3821568 Apr 9 17:16 .posts.xml.swp
[[ec2-user@ip-172-31-47-53 Data]$ rm .posts*
[[ec2-user@ip-172-31-47-53 Data]$ vi baseball.txt
[[ec2-user@ip-172-31-47-53 Data]$ pig
19/04/09 17:18:24 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
19/04/09 17:18:24 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
19/04/09 17:18:24 INFO pig.ExecTypeProvider: Trying ExecType : TEZ_LOCAL
19/04/09 17:18:24 INFO pig.ExecTypeProvider: Trying ExecType : TEZ
19/04/09 17:18:24 INFO pig.ExecTypeProvider: Picked TEZ as the ExecType
19/04/09 17:18:24 INFO pig.Main: Loaded log4j properties from file: /etc/pig/conf/log4j.properties
53 [main] INFO org.apache.pig.Main - Apache Pig version 0.17.0 (r: unknown)
compiled Mar 19 2019, 23:58:33
19/04/09 17:18:24 INFO pig.Main: Apache Pig version 0.17.0 (r: unknown) compiled
Mar 19 2019, 23:58:33
54 [main] INFO org.apache.pig.Main - Logging error messages to: /mnt/var/log
/pig/pig_1554830304830.log
19/04/09 17:18:24 INFO pig.Main: Logging error messages to: /mnt/var/log/pig/pig
_1554830304830.log
75 [main] INFO org.apache.pig.impl.util.Utills - Default bootup file /home/ec
2-user/.pigbootup not found
19/04/09 17:18:24 INFO util.Utills: Default bootup file /home/ec2-user/.pigbootup
not found
19/04/09 17:18:25 INFO Configuration.deprecation: mapred.job.tracker is deprecat
ed. Instead, use mapreduce.jobtracker.address
812 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine
- Connecting to hadoop file system at: hdfs://ip-172-31-47-53.ec2.internal:802
0
19/04/09 17:18:25 INFO executionengine.HExecutionEngine: Connecting to hadoop fi
le system at: hdfs://ip-172-31-47-53.ec2.internal:8020
1506 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG
-default-a69265a8-32bf-42c2-9f02-83682ff466e6
19/04/09 17:18:26 INFO pig.PigServer: Pig Script ID for the session: PIG-default
-a69265a8-32bf-42c2-9f02-83682ff466e6
19/04/09 17:18:27 INFO impl.TimelineClientImpl: Timeline service address: http:/
/ip-172-31-47-53.ec2.internal:8188/ws/v1/timeline/
2413 [main] INFO org.apache.pig.backend.hadoop.PigATSCClient - Created ATS Hook
19/04/09 17:18:27 INFO hadoop.PigATSCClient: Created ATS Hook
grunt>
```



# Running Pig

- In grunt:
  - Load piggybank
- `register file:/usr/lib/pig/lib/piggybank.jar`

# Running Pig

- You can interact with your local and the hadoop file system
- Use sh command to run local file system commands

```
| [grunt> sh ls
| stockdata
```

# Running Pig

- To move your file to the Hadoop file system, use `copyFromLocal`
- You interact with your Hadoop distributed file system using `fs` and then a negative sign

```
[grunt> copyFromLocal stockdata stockdata
```

```
[grunt> copyFromLocal stockdata stockdata
[grunt> fs -ls
Found 1 items
-rw-r--r-- 1 hadoop hadoop 105 2019-04-09 21:05 stockdata
```

# Running Pig

- Now you can create your first relation from your file
  - Assign a name
  - Use Load
  - Specify how to read it
    - PigStorage()
    - TextLoader

```
grunt> A = LOAD 'stockdata' USING PigStorage() AS (name:chararray, price:double, volume:long);
grunt> DUMP A;
```

# Running Pig

- Now you can create your first relation from your file
  - PigStorage has an optional argument, namely the delimitator
  - Scheme uses common types:
    - chararray, double, float, long, int, ...
- Then you can run your first map-reduce job: Dump
  - Takes some time

```
grunt> A = LOAD 'stockdata' USING PigStorage() AS (name:chararray, price:double, volume:long);
grunt> DUMP A;
```



# Running Pig

- Can use illustrate:
  - Another map-reduce job

```
[grunt> A = LOAD 'stockdata' USING PigStorage() AS (name:chararray, price:double, volume:long);
[grunt> ILLUSTRATE A;
13891 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting t
```

```
19/04/09 21:41:04 INFO mapReduceLayer.PigMapOnly$Map: Aliases being p
```

| A | name:chararray | price:double | volume:long |
|---|----------------|--------------|-------------|
|   | ACER           | 5.32         | 13491432    |





# Running Pig

- Can project with Generate

```
(ACH, 100.23)
grunt> B = foreach A Generate name, price;
```

```
(A, 35.21)
(AA, 92.12)
(AABA, 1.54)
(ACGL, 12.98)
(ACER, 5.32)
(ACH, 100.23)
```



# Running Pig

- Can order with other column

```
[grunt> C = GROUP A by $2;
```

```
(2134, {(ACGL, 12.98, 2134)})
(239148, {(AABA, 1.54, 239148)})
(1233214, {(ACH, 100.23, 1233214)})
(1234243, {(A, 35.21, 1234243)})
(13491432, {(ACER, 5.32, 13491432)})
(98739879, {(AA, 92.12, 98739879)})
```



# Stop Running Pig

- Use quit to exit pig
- exit from the master
- Go to the AWS management console
  - Select EMR
    - Select your cluster
    - Set it to terminating



# Stop Running Pig

The screenshot shows the AWS Management Console interface for the EMR Clusters page. The browser address bar shows the URL: `console.aws.amazon.com/elasticmapreduce/home?region=us-east-1`. The page header includes navigation menus for Services, Resource Groups, and user information (tbobsj, N. Virginia, Support).

A notification banner at the top states: "You can use the AWS Glue Data Catalog as your external Hive metastore for Apache Spark, Apache Hive, and Presto workloads on Amazon EMR release 5.10.0 and later. To get started, simply select the AWS Glue Data Catalog for table metadata when creating your cluster."

Below the notification, there are buttons for "Create cluster", "View details", "Clone", and "Terminate". A filter dropdown is set to "All clusters", showing "6 clusters (all loaded)".

|                                     | Name       | ID               | Status                                     | Creation time (UTC-5)    | Elapsed time        | Normalized instance hours |
|-------------------------------------|------------|------------------|--------------------------------------------|--------------------------|---------------------|---------------------------|
| <input checked="" type="checkbox"/> | My cluster | j-29JNU8VKK7UD7  | Waiting<br>Cluster ready                   | 2019-04-09 15:43 (UTC-5) | 1 hour, 6 minutes   | 0                         |
| <input type="checkbox"/>            | My cluster | j-1LFPFEJQXN3J5  | Terminated<br>User request                 | 2019-04-09 11:52 (UTC-5) | 2 hours, 48 minutes | 72                        |
| <input type="checkbox"/>            | My cluster | j-2LQPPLYI63G9NW | Terminated with errors<br>Instance failure | 2019-03-11 23:25 (UTC-5) | 20 hours            | 240                       |
| <input type="checkbox"/>            | My cluster |                  |                                            | UTC-5)                   | 1 minute            | 0                         |
| <input type="checkbox"/>            | My cluster |                  |                                            | UTC-5)                   | 3 minutes           | 0                         |
| <input type="checkbox"/>            | My cluster |                  |                                            | UTC-5)                   | 15 minutes          | 24                        |

A "Terminate clusters" dialog box is open in the foreground, asking for confirmation to terminate the cluster. The dialog contains the following text:

**Terminate clusters**

Are you sure you want to terminate this cluster?

- j-29JNU8VKK7UD7 (My cluster)

Any pending work or data residing on the cluster will be lost, such as data stored in HDFS. This action is irreversible.

Buttons: [Cancel](#) [Terminate](#)



# Stop Running Pig

The screenshot shows the AWS Management Console interface for Amazon EMR. The top navigation bar includes the AWS logo, 'Services', 'Resource Groups', and user information. A sidebar on the left lists navigation options like 'Amazon EMR', 'Clusters', 'Security configurations', etc. The main content area displays a table of EMR clusters with columns for Name, ID, Status, Creation time, Elapsed time, and Normalized instance hours. A 'Terminating' cluster is highlighted with an orange circle.

**Filter:** All clusters  6 clusters (all loaded) ↻

|                          | Name                                     | ID              | Status                                     | Creation time (UTC-5)    | Elapsed time        | Normalized instance hours |
|--------------------------|------------------------------------------|-----------------|--------------------------------------------|--------------------------|---------------------|---------------------------|
| <input type="checkbox"/> | <span>▶</span> <span>○</span> My cluster | j-29JNU8VKK7UD7 | Terminating<br>User request                | 2019-04-09 15:43 (UTC-5) | 1 hour, 8 minutes   | 24                        |
| <input type="checkbox"/> | <span>▶</span> My cluster                | j-1LFPFEJQXN3J5 | Terminated<br>User request                 | 2019-04-09 11:52 (UTC-5) | 2 hours, 48 minutes | 72                        |
| <input type="checkbox"/> | <span>▶</span> <span>●</span> My cluster | j-2LQPYLI63G9NW | Terminated with errors<br>Instance failure | 2019-03-11 23:25 (UTC-5) | 20 hours            | 240                       |
| <input type="checkbox"/> | <span>▶</span> <span>●</span> My cluster | j-23D4JM5T5RX04 | Terminated with errors<br>Validation error | 2019-03-11 23:22 (UTC-5) | 1 minute            | 0                         |
| <input type="checkbox"/> | <span>▶</span> <span>●</span> My cluster | j-2EVXIAFL6PB1L | Terminated with errors<br>Validation error | 2019-03-11 23:01 (UTC-5) | 3 minutes           | 0                         |
| <input type="checkbox"/> | <span>▶</span> My cluster                | j-2T9BCSEUWQ3AG | Terminated<br>User request                 | 2019-03-11 22:23 (UTC-5) | 15 minutes          | 24                        |

# Checking for damage

- Select your name on the drop down
- Select account
- Select cost explorer