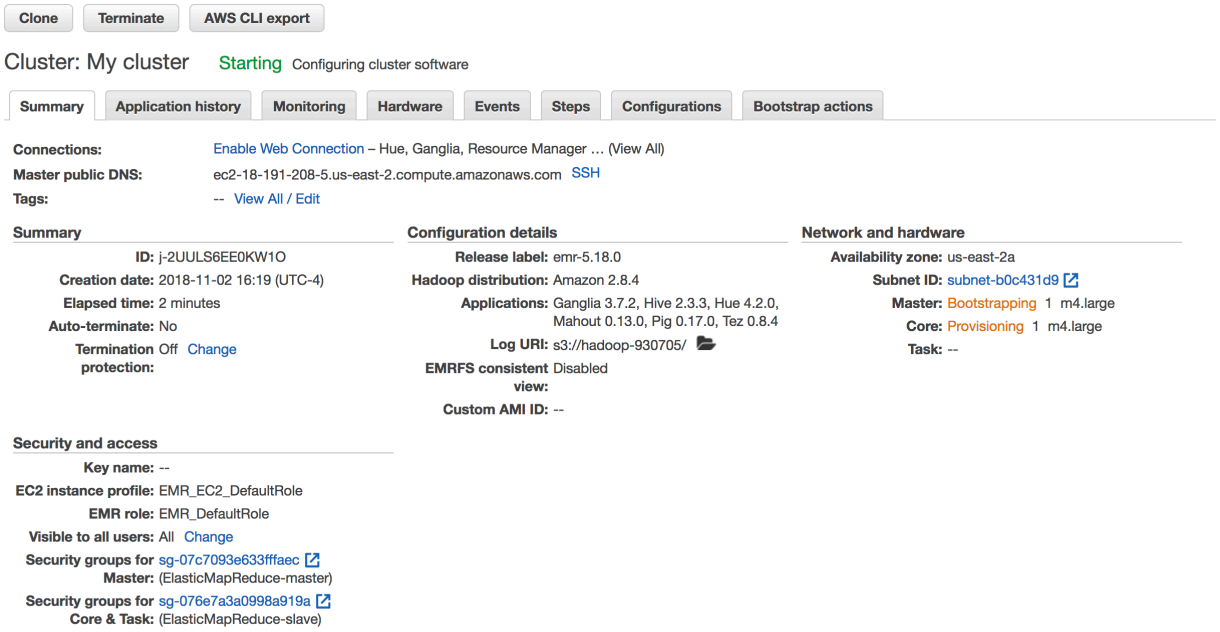
Analyze Big data with Hadoop

1. Create an Amazon S3 bucket

2. Launch an AWS EMR cluster



The several default application on the cluster:

- Apache Hadoop
- Ganglia
- Apache Taz
- Hue
- Pig

Data to be processed

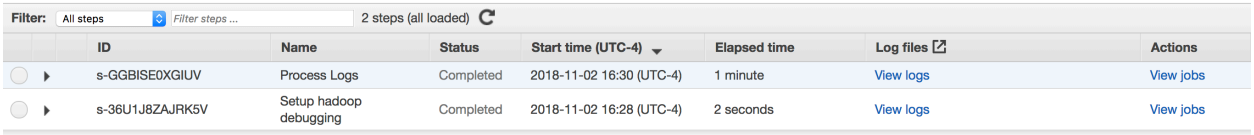
Amazon CloudFront is a web service that speeds up distribution of static and dynamic web content, such as .html, .css, .php, and image files. CloudFront delivers content through a worldwide network of data centers called *edge locations*. When a user requests content through CloudFront, the user is routed to the edge location that provides the lowest latency (time delay), so that content is delivered with the best possible performance. If the content is already in the edge location with the lowest latency, CloudFront delivers it immediately. If the content is not in that edge location, CloudFront retrieves it from an Amazon S3 bucket or an HTTP server (for example, a web server) that you have identified as the source for the definitive version of your content.

3. Running Hive script to process data

What the script is doing

The Hive script does the following:

- Creates a **Hive table** named *cloudfront_logs*.
- Reads the **CloudFront log files** from Amazon S3 and parses the files using the Regular Expression Serializer/Deserializer (*RegEx SerDe*).
- Writes the parsed results to the *cloudfront_logs* Hive table.
- Submits a HiveQL query against the data to retrieve the **total requests per operating system for a given time frame**.
- Writes the query results to your Amazon S3 output bucket.



Filter: All steps | Filter steps ... | 2 steps (all loaded)

ID	Name	Status	Start time (UTC-4)	Elapsed time	Log files	Actions
s-GGBISE0XGIUV	Process Logs	Completed	2018-11-02 16:30 (UTC-4)	1 minute	View logs	View jobs
s-36U1J8ZAJRK5V	Setup hadoop debugging	Completed	2018-11-02 16:28 (UTC-4)	2 seconds	View logs	View jobs

Output:

000000_0

```
Android855
Linux813
MacOS852
OSX799
Windows883
iOS794
```

4. Terminate the cluster
5.

Clone | Terminate | AWS CLI export

Cluster: My cluster Terminating Terminated by user request

Summary | Application history | Monitoring | Hardware | Events | Steps | Configurations | Bootstrap actions

Connections: --
Master public DNS: ec2-18-191-208-5.us-east-2.compute.amazonaws.com SSH
Tags: --

Summary
ID: j-2UULS6EE0KW1O
Creation date: 2018-11-02 16:19 (UTC-4)
Elapsed time: 26 minutes
Auto-terminate: No
Termination protection: Off

Configuration details
Release label: emr-5.18.0
Hadoop distribution: Amazon 2.8.4
Applications: Ganglia 3.7.2, Hive 2.3.3, Hue 4.2.0, Mahout 0.13.0, Pig 0.17.0, Tez 0.8.4
Log URI: s3://hadoop-930705/
EMRFS consistent view: Disabled
Custom AMI ID: --

Network and hardware
Availability zone: us-east-2a
Subnet ID: subnet-b0c431d9
Master: Running 1 m4.large
Core: Running 1 m4.large
Task: --

Summary

In this lab, I have learned the basic operation of AWS ECR, how to use Hadoop to process data and store it on AWS.