**Preparation/Analysis Big Data - BIA 6305**

**Leonardo Ji**

**Assignment 3**

**Part 1 Hadoop Apache Ecosystem**

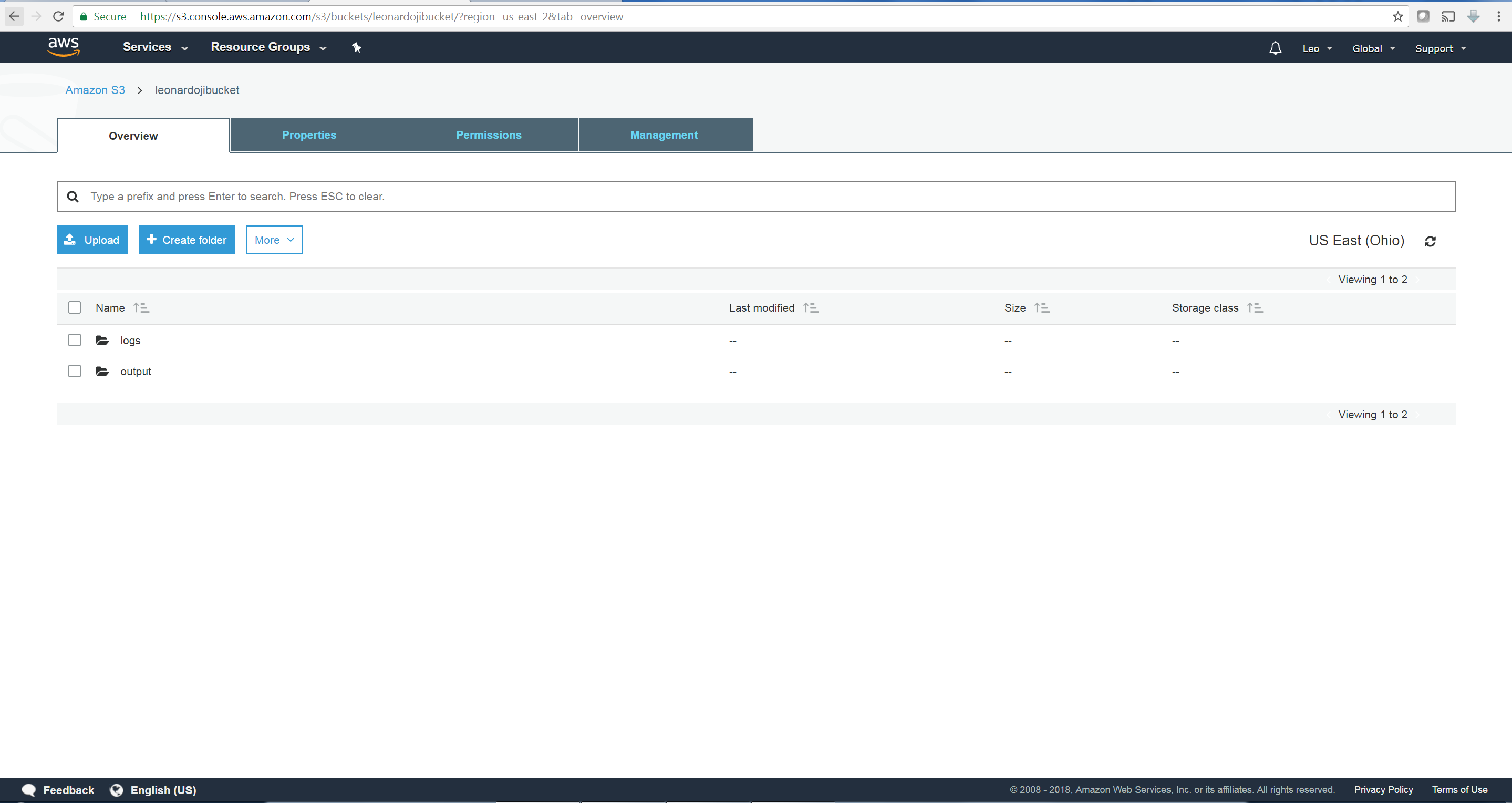
The table listed some of the open source software or systems related to Apache Hadoop.

|  |  |  |
| --- | --- | --- |
| **Name** | **Type** | **Definition** |
| HDFS | Distributed File system | The Hadoop Distributed File System (HDFS) is used to store large files across multiple machines. It is derived from Google File System (GFS). |
| Ambari | System Deployment | Hadoop web management UI. It allows provision, deploy, management of a Hadoop cluster and nodes. |
| Mesos | System Deployment | Mesos is a cluster manager that provides resource sharing and isolation across cluster applications. |
| Bigtop | System Deployment | Developed by Cloudera. BigTop is used for some vendors to build their own distributions based on Apache Hadoop |
| Hbase | Database | Hadoop non-relationship database. Inspired by Google BigTable. |
| Cassandra | Database | Another non-SQL distributed database. Inspired by how Amazon’s Dynamo storing data. |
| Neo4j | Database | An open-source graph database written in Java. It is an embedded, disk-based, fully transactional Java persistence engine that stores data structured in graphs rather than in tables. |
| YARN | Distributed Programming | YARN stands for “Yet-Another-Resource-Negotiator”. YARN is a resource manager and node manager to execute MapReduce application. |
| Pig | Distributed Programming | Compiles Pig script into MapReduce jobs that runs on Hadoop. |
| Spark | Distributed Programming | Data Analytics platform on top of HDFS. Faster in memory approach vs Map Reduce as in Hadoop. Spark ML library also provide training/running machine learning models. |
| Storm | Distributed Programming | A distributed real-time computation system for processing fast, large streams of data. Storm is an architecture based on master-workers paradigm. |
| Hive | SQL | Data warehouse developed by Facebook. It provides SparkQL, a SQL-like language to query data on Hadoop. |
| Impala | SQL | The Apache-licensed Impala project brings scalable parallel database technology to Hadoop, enabling users to issue low-latency SQL queries to data stored in HDFS and Apache HBase without requiring data movement or transformation. It's a Google Dremel clone (Big Query google). |
| Presto | SQL | Facebook created and open sourced SQL engine it says is on average 10 times faster than Hive for running queries across large data sets stored in Hadoop and elsewhere. |
| Drill | SQL | Drill is the open source version of Google's Dremel system which is available as an infrastructure service called Google BigQuery. |
| Kafka | Data | Distributed publish-subscribe system for processing large amounts of streaming data. Kafka is a Message Queue developed by LinkedIn that persists messages to disk in a very performant manner. |
| Thrift | Service Programming | A cross-language RPC framework for service creations. It’s the service base for Facebook technologies (the original Thrift contributor). Thrift provides a framework for developing and accessing remote services. |
| Zookeeper | Service Programming | It’s a coordination service that gives you the tools you need to write correct distributed applications. ZooKeeper was developed at Yahoo! Research. Several Hadoop projects are already using ZooKeeper to coordinate the cluster and provide highly-available distributed services. Perhaps most famous of those are Apache HBase, Storm, Kafka. |
| Avro | Service Programming | Apache Avro is a framework for modeling, serializing and making Remote Procedure Calls (RPC). Avro data is described by a schema, and one interesting feature is that the schema is stored in the same file as the data it describes, so files are self-describing. |
| Norbert | Service Programming | Created by Linkedin. Norbert is a library that provides easy cluster management and workload distribution. |
| Oozie | Scheduler | Workflow scheduler system for MR jobs using DAGs (Direct Acyclical Graphs). Oozie Coordinator can trigger jobs by time (frequency) and data availability. |
| Mahout | Machine Learning | Machine learning library and math library, on top of MapReduce. |
| H2O | Machine Learning | H2O is the world’s fastest in-memory platform for machine learning and predictive analytics on big data. It is designed to help users scale machine learning, math, and statistics over large datasets. |
| Nutch | Search Tool | Highly extensible and scalable open source web crawler software project. |
| Solr | Search Tool | Search engine built on top of Lucene. Standalone search engine. |
| Elasticsearch | Search Tool | It is a distributed, RESTful search and analytics engine capable of solving a growing number of use cases. Also based on Lucene. |

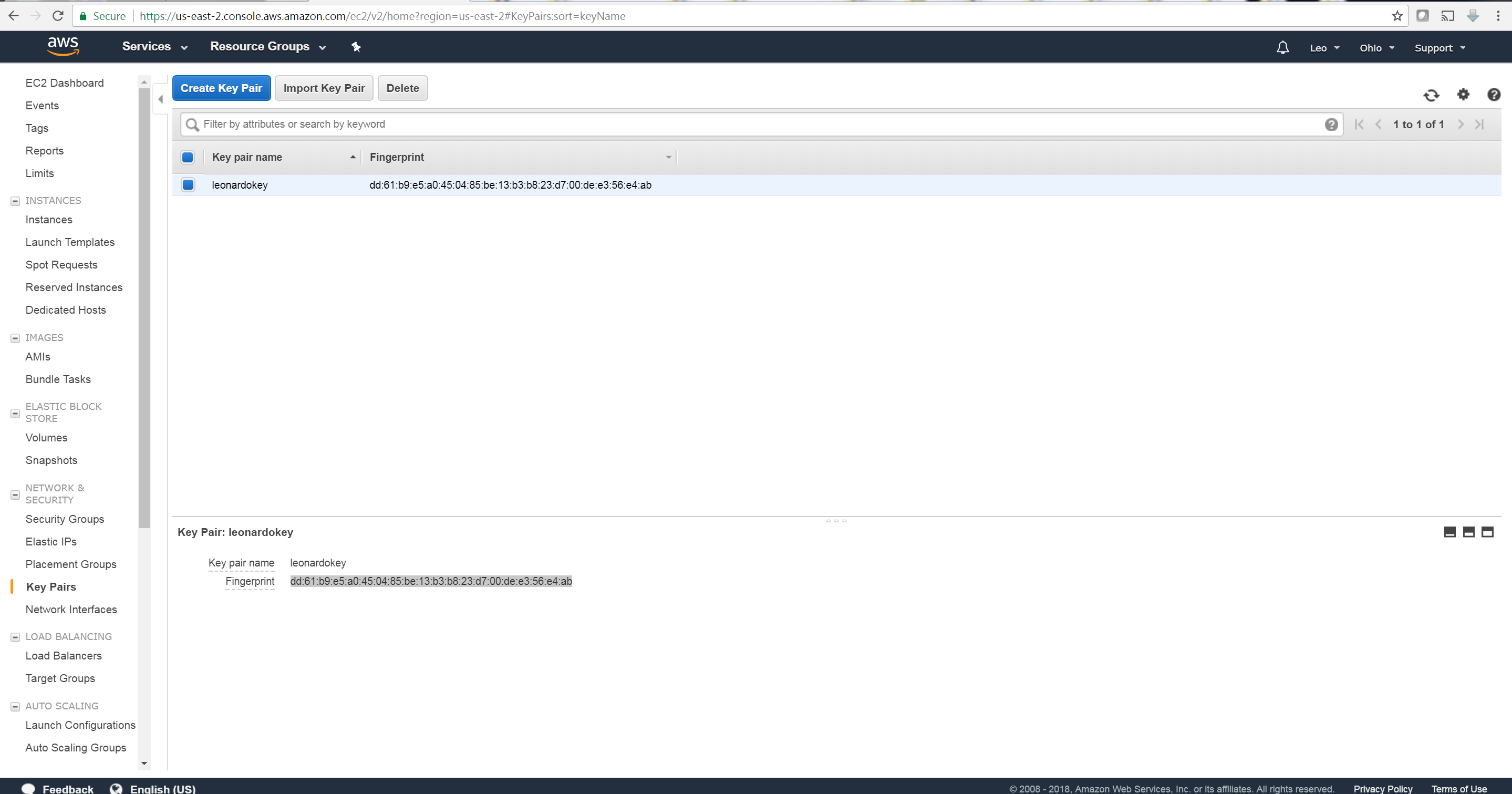
**Part 2**

Step 1

Set up S3 bucket

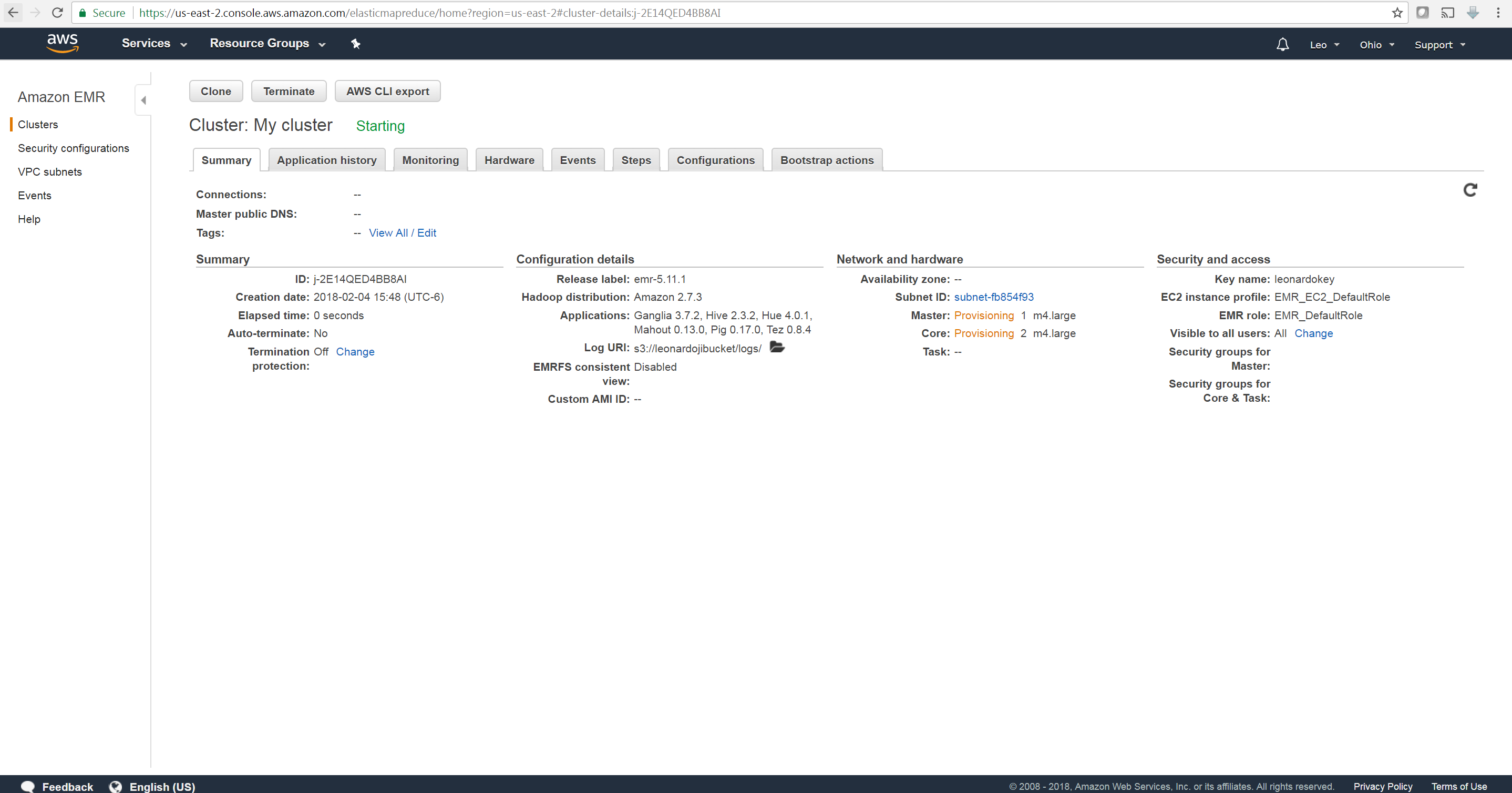


Set up SSH key

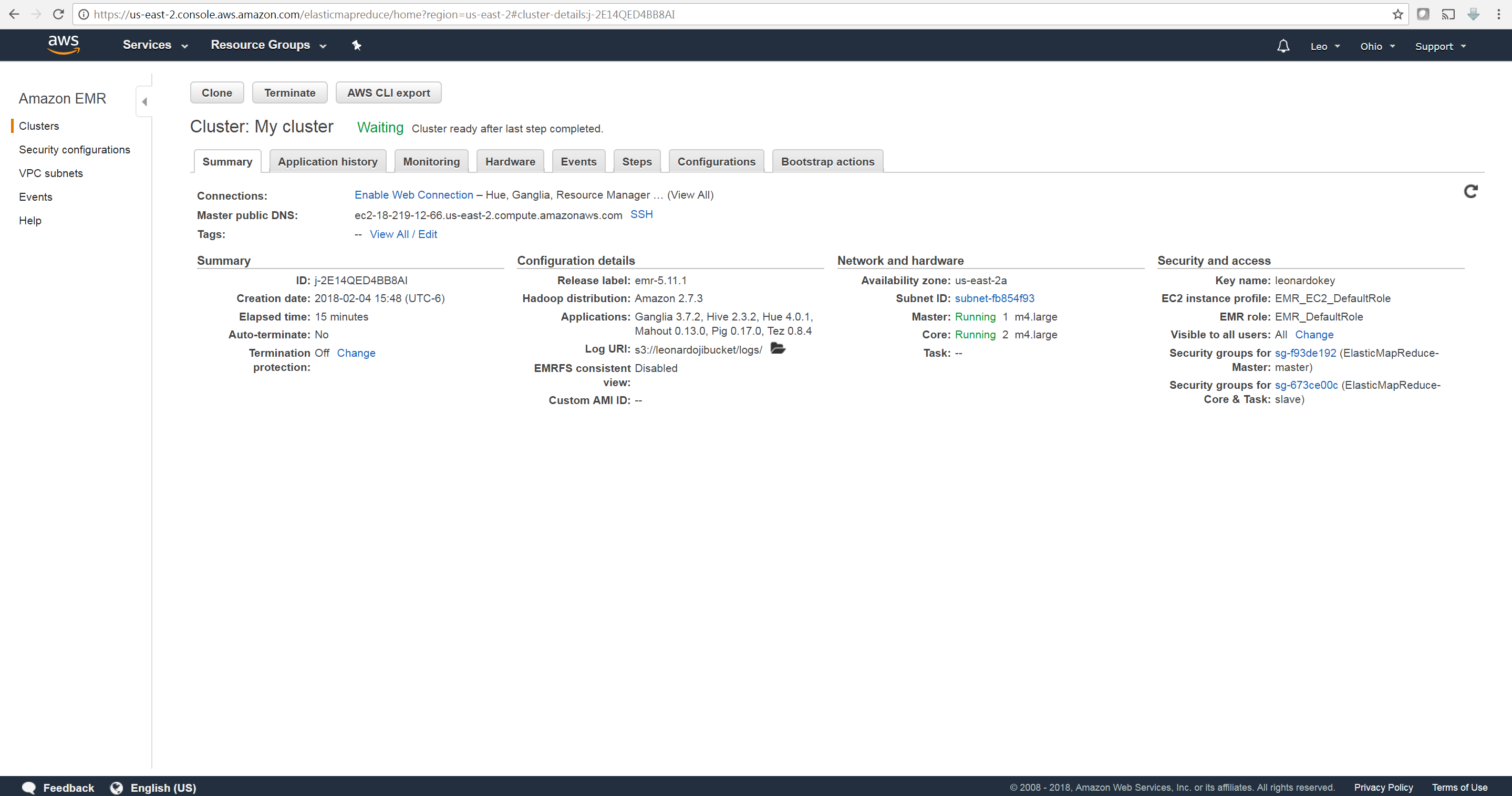


Step 2

Launch a Cluster



Step 3



cloudfront\_logs:

2014-07-05 20:00:00 LHR3 4260 10.0.0.15 GET eabcd12345678.cloudfront.net /test-image-1.jpeg 200 - Mozilla/5.0%20(MacOS;%20U;%20Windows%20NT%205.1;%20en-US;%20rv:1.9.0.9)%20Gecko/2009040821%20IE/3.0.9

Hive Code:

CREATE EXTERNAL TABLE IF NOT EXISTS cloudfront\_logs (

DateObject Date,

Time STRING,

Location STRING,

Bytes INT,

RequestIP STRING,

Method STRING,

Host STRING,

Uri STRING,

Status INT,

Referrer STRING,

OS String,

Browser String,

BrowserVersion String

)

Hive Code parse log:

ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.RegexSerDe'

WITH SERDEPROPERTIES (

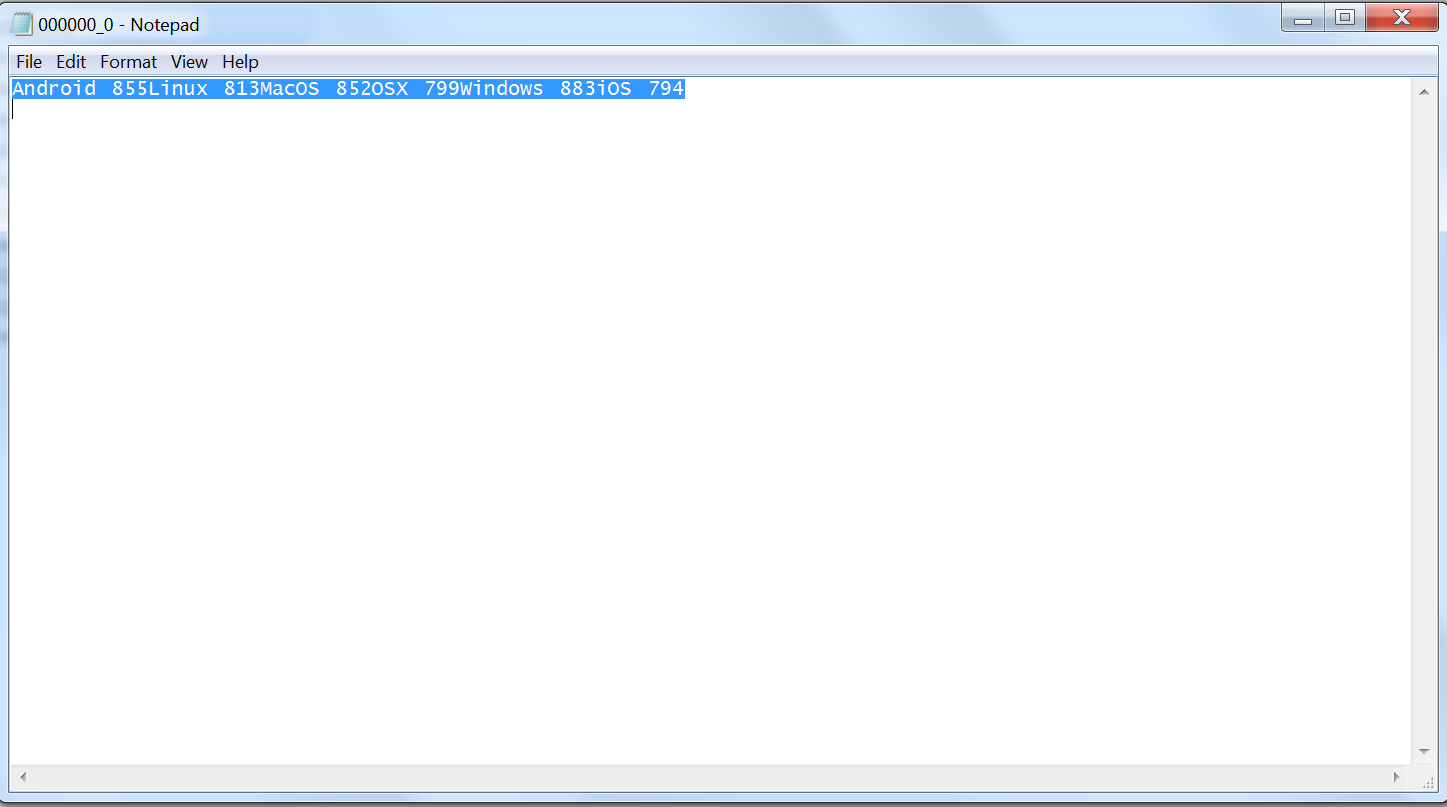
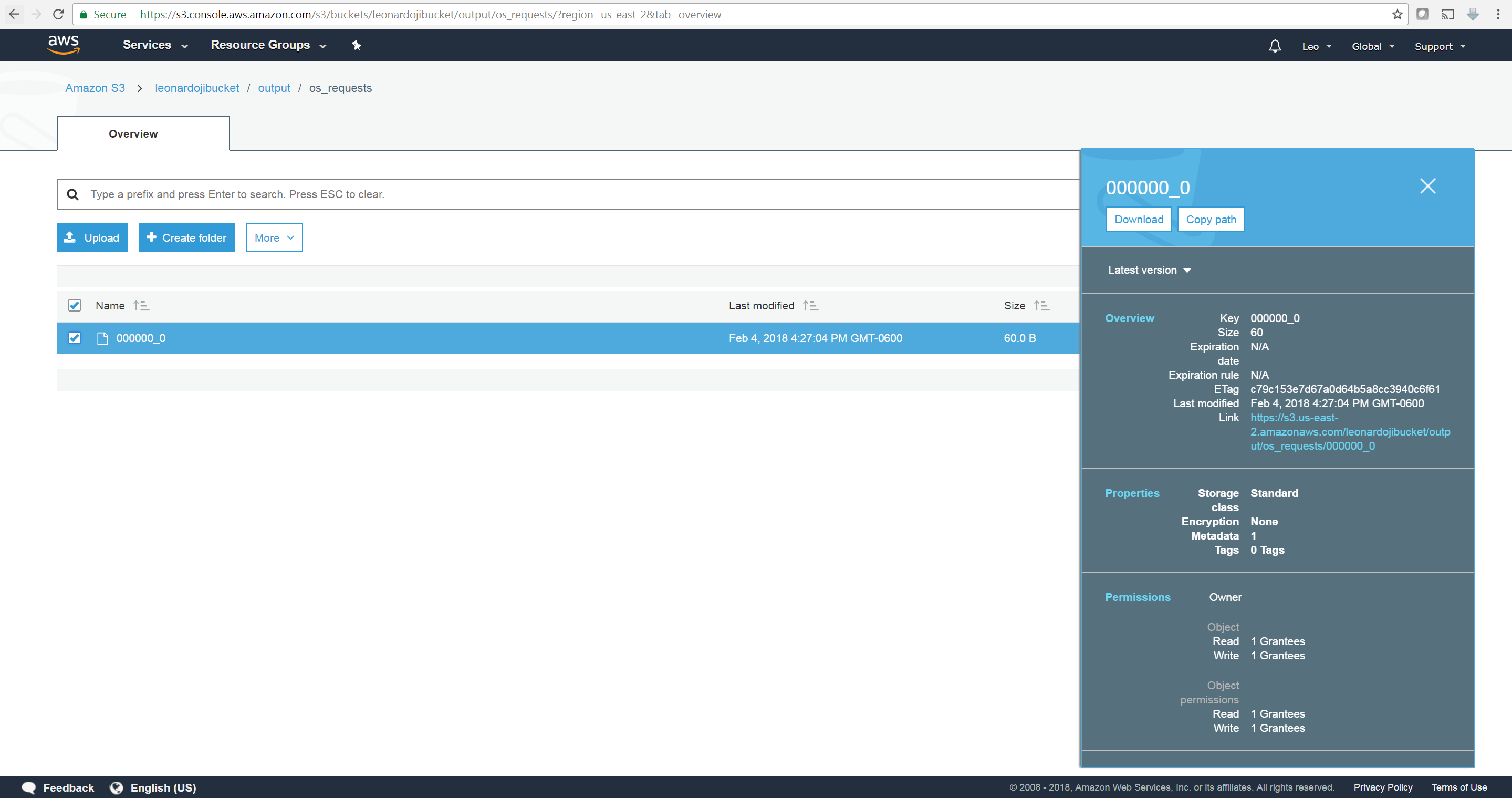
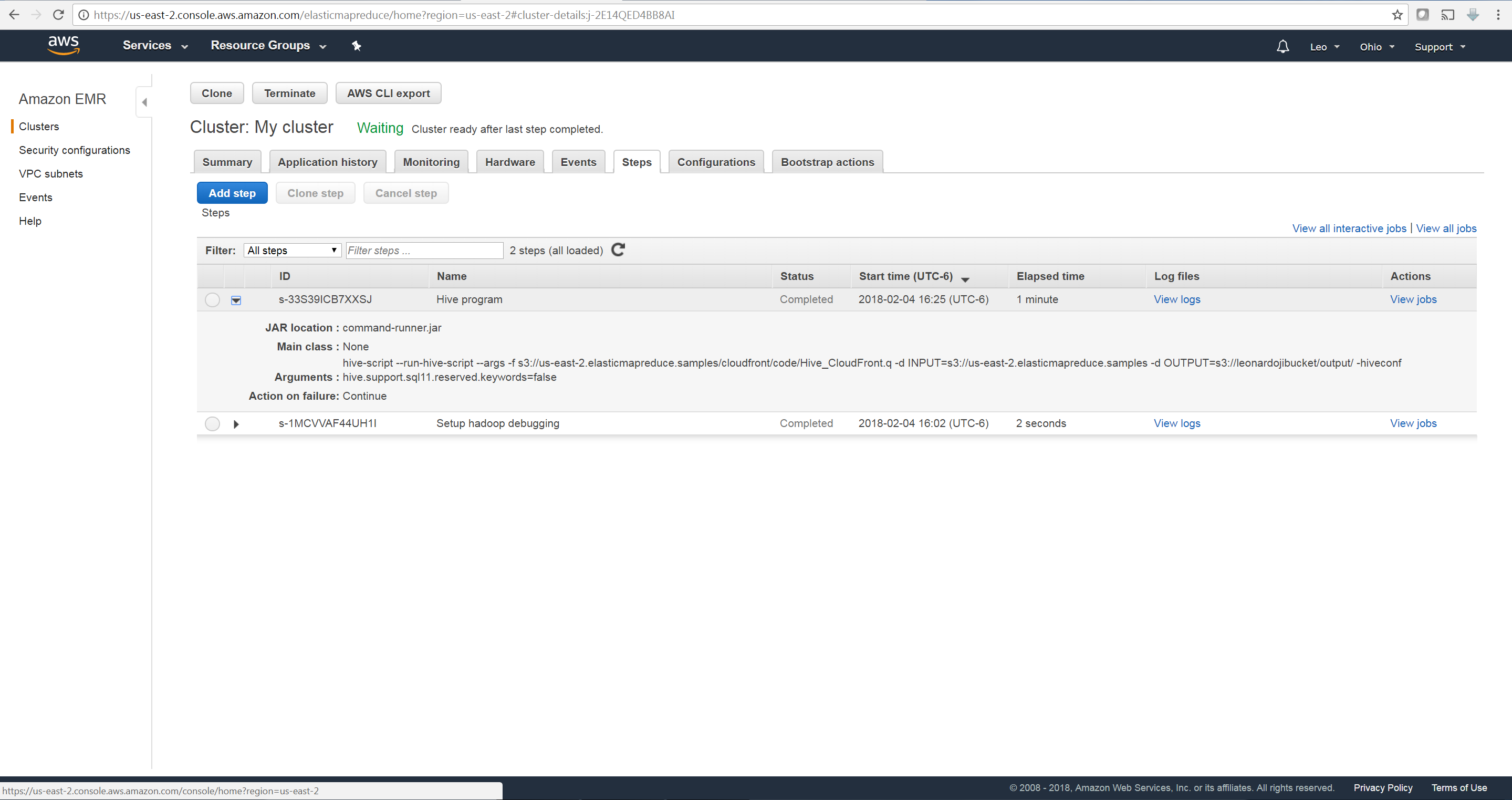
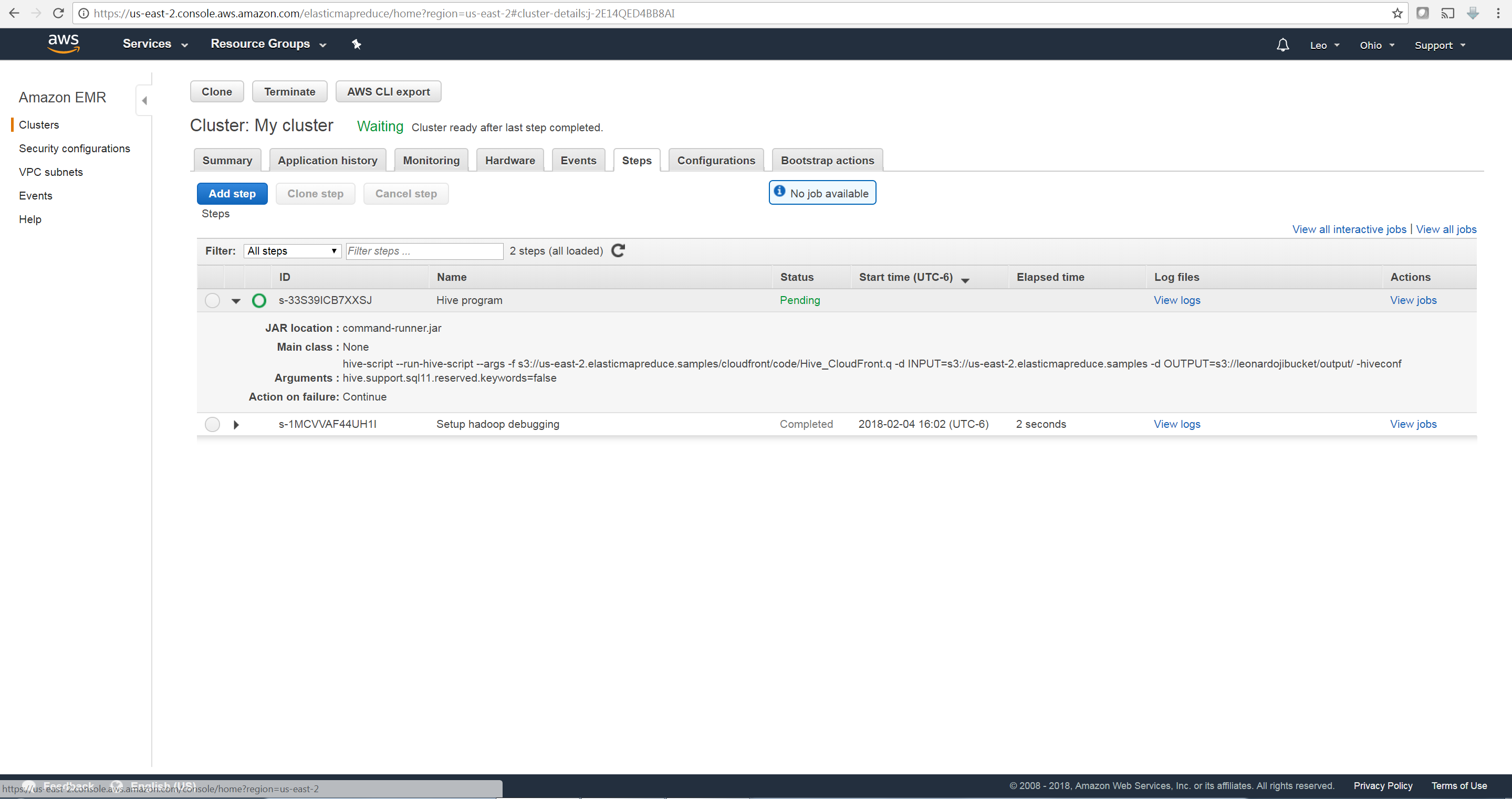
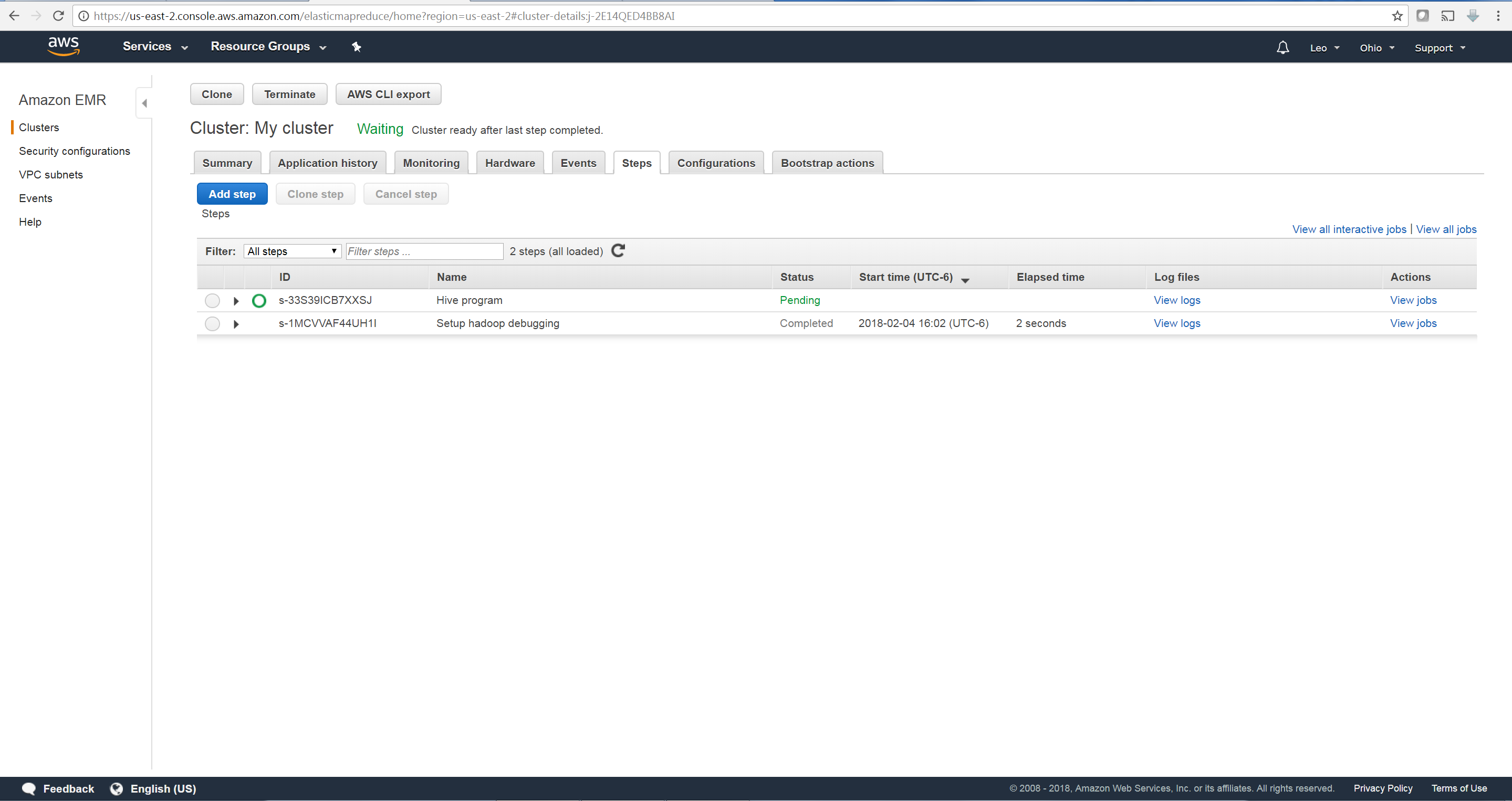
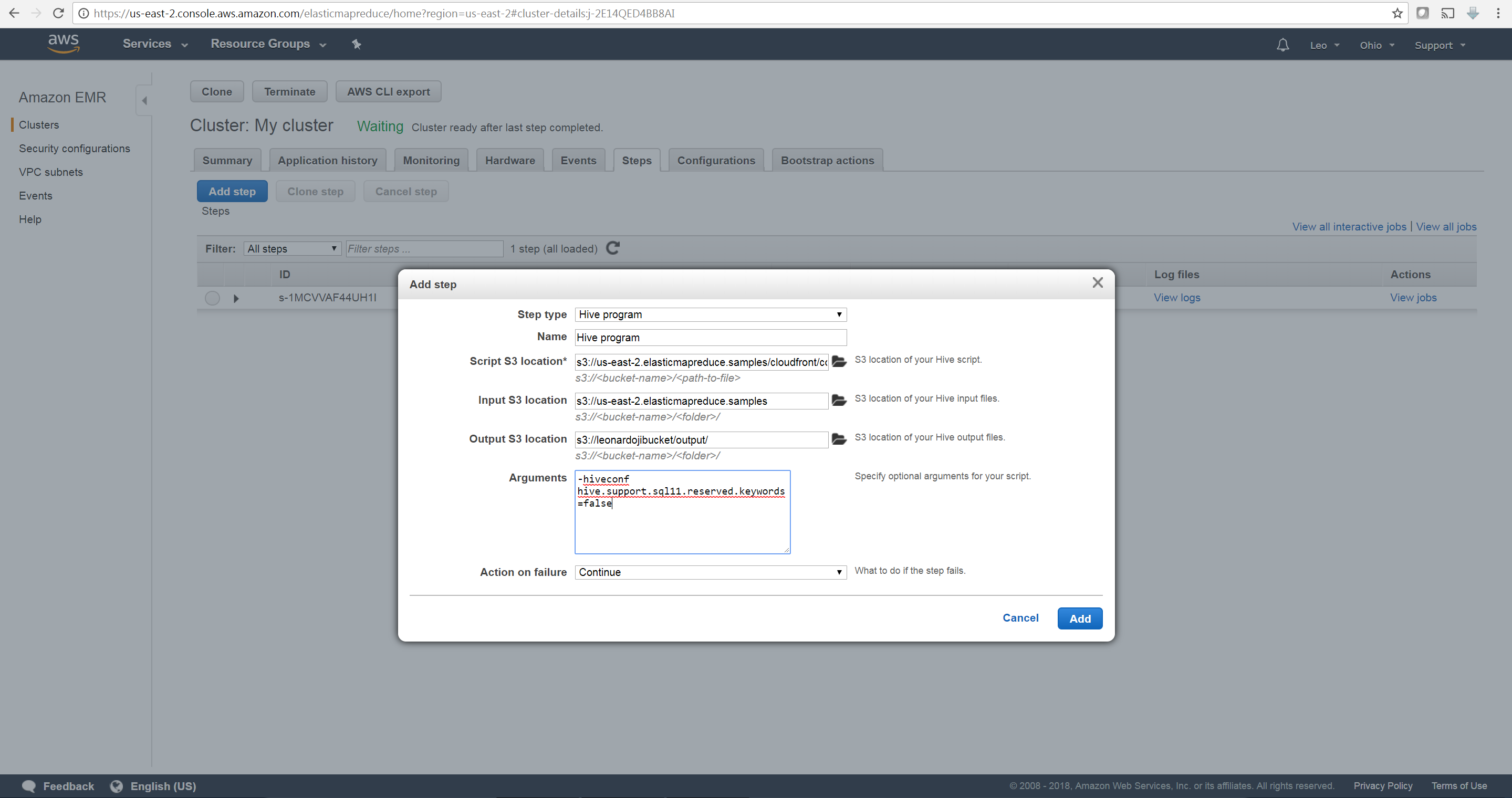
"input.regex" = "^(?!#)([^ ]+)\\s+([^ ]+)\\s+([^ ]+)\\s+([^ ]+)\\s+([^ ]+)\\s+([^ ]+)\\s+([^ ]+)\\s+([^ ]+)\\s+([^ ]+)\\s+([^ ]+)\\s+[^\(]+[\(]([^\;]+).\*\%20([^\/]+)[\/](.\*)$"

) LOCATION '${INPUT}/cloudfront/data/';

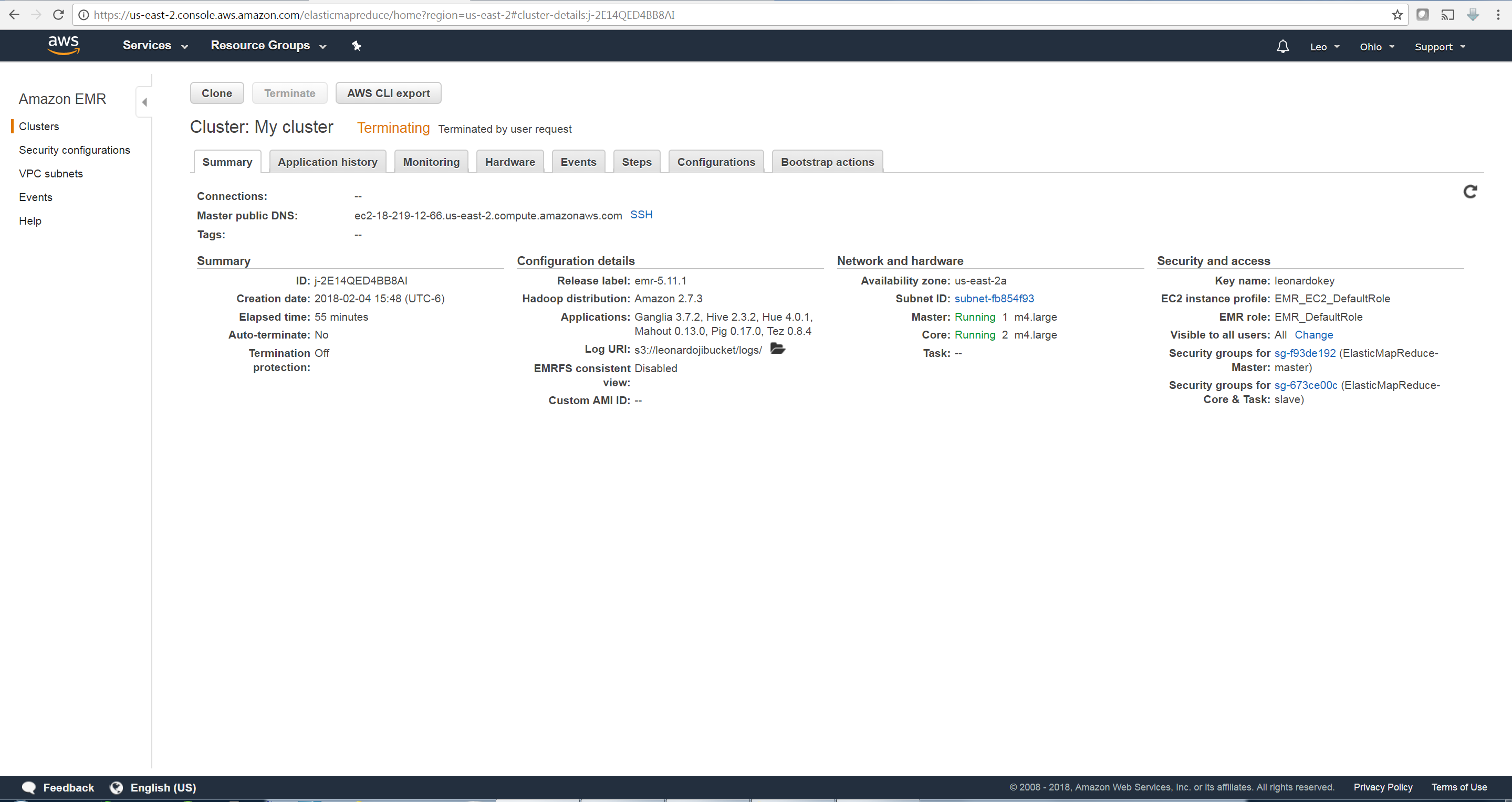
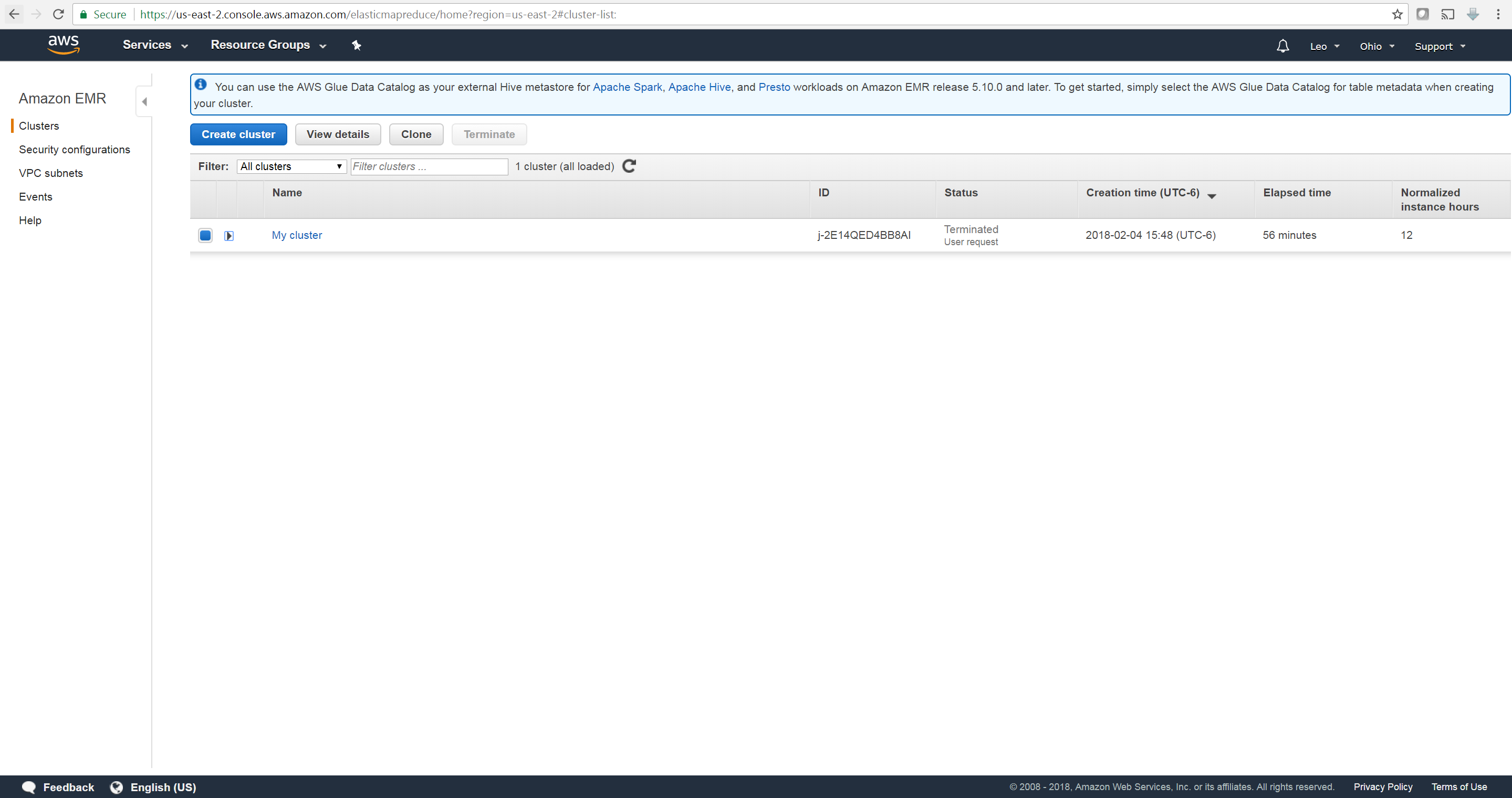
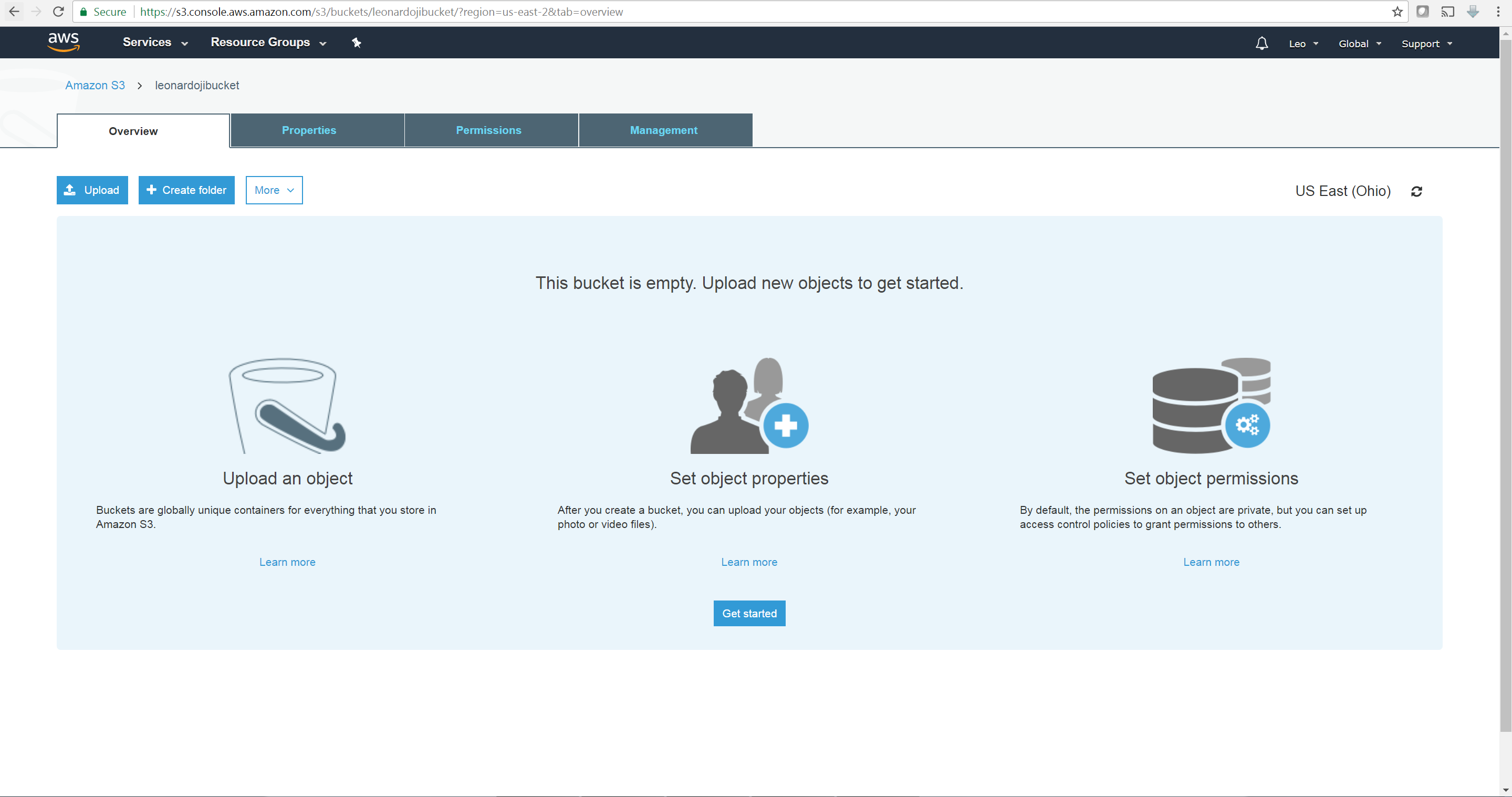
HiveQL query:

INSERT OVERWRITE DIRECTORY '${OUTPUT}/os\_requests/' SELECT os, COUNT(\*) count FROM cloudfront\_logs WHERE dateobject BETWEEN '2014-07-05' AND '2014-08-05' GROUP BY os;

Step 4



Step 5

**Part 3**

1. Define (A) Big Data Search and (B) Big Data Mining.

Big Data Search is finding a portion of data meets specific criteria uses index and associated search technologies.

Big Data Mining is extraction of useful insights from large datasets or streams of data.

1. What type of intermediate data does MapReduce store? Where does MapReduce store them?

Mappers first process the data. The maps of key-value pairs are returned from Mappers to be stored on the nodes’ local disk. Later the intermediate key-value pairs map will be available to Reducer. Reducer will aggregate intermediate key-value pairs maps into one output.

1. List ideas to improve/tune existing text processing and mining approaches to support Big Data scale.

Text processing and mining needs to support the large scale of streaming data in the form of sequences of posts on social networks. The sentiment analysis need to be done in real time and update the sentiment over time. User feedback on the post can be incorporated to be part of sentiment evaluation system.

Consumer sentiment from web such as reviews, comments, tweets can be used to gather people’s views on product or service. From unstructured web documents, we can deploy a workflow of web crawl, clean data, sentiment extraction, dashboard to assign positive, negative, neutral sentiments on a product or service.

1. A) Explain the significance of social networks and their roles in the context of Big Data?

Finding communities in social network data according to specific business question is a very challenging task. The problem is finding clusters or identifying groups are more like to each other. Then we can market a product or service to a group that are more likely to purchase it. Or we can find key influencer or critical person in a social network group for marketing.

B) List challenges of Big Data in supporting social network analytics and discuss approaches to handle them with justifications.

* Social network community detection: the challenge is use clustering algorithms to find groups share common characteristics.
* Social network topology discovery: the challenge is to find key influencer in a star topology, or to find a critical node in a ring topology, or to categorize active communities and fully connected components in a mesh topology.
* Social network condensation: social network is very large, condensing or merging nodes with common traits would make analysis much easier. We need to use an algorithm to merge or cluster nodes.