Notes on Pegasus set-up

Please follow the Readme at <https://github.com/InsightDataScience/pegasus> but with these revisions:

### Clone the Ubuntu **branch** on the control machine (very important that you are cloning the feat/ubuntu16 branch):

git clone -b feat/ubuntu16 --single-branch <https://github.com/InsightDataScience/pegasus>

### Make sure these environment variables are set up in ~/.profile (Note: XXXX should be changed to your access key; if you are on the west coast, the region should be us-west-2; <path-to-pegasus> should be changed to the path to your pegasus directory)

export AWS\_ACCESS\_KEY\_ID=XXXX  
export AWS\_SECRET\_ACCESS\_KEY=XXXX  
export AWS\_DEFAULT\_REGION=us-west-1  
export REM\_USER=ubuntu  
export PEGASUS\_HOME=<path-to-pegasus>  
export PATH=$PEGASUS\_HOME:$PATH

### Execute profile when those variables have been set up:

. ~/.bash\_profile

### Verify pegasus is configured:

peg config

### Now check out a couple of sample yml files that you’ll use to configure your pegasus cluster:

cd ~/pegasus/examples/spark

### In this directory are two files you can modify for your uses:

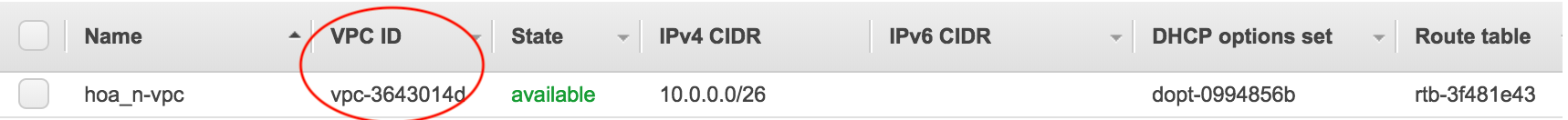
* master.yml
* workers.yml

### In each .yml file, you’ll want to change the following fields:

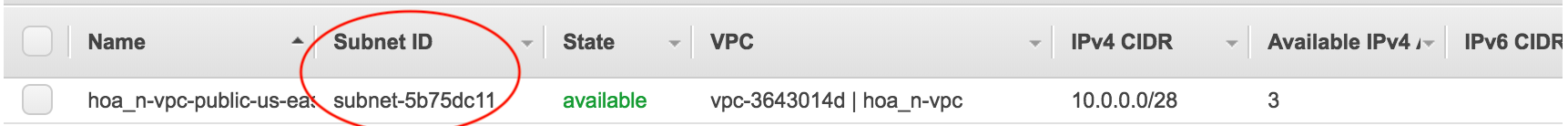
* subnet\_id:
* security\_group\_ids:
* key\_name: set this to the prefix of your pem file (e.g. set to hoa-nguyen and drop .pem)
* purchase\_type: make sure this field is always set to on\_demand

### Below are instructions on how to locate the right subnet\_id and security\_group\_ids for your cluster

### For the cluster created by Terraform for you, find the right VPC and subnet. Here’s the VPC that was created for me (taken from my AWS console). Hint: In AWS, search for “VPC” to find this window



### Here’s the subnet that was created for me and the above VPC



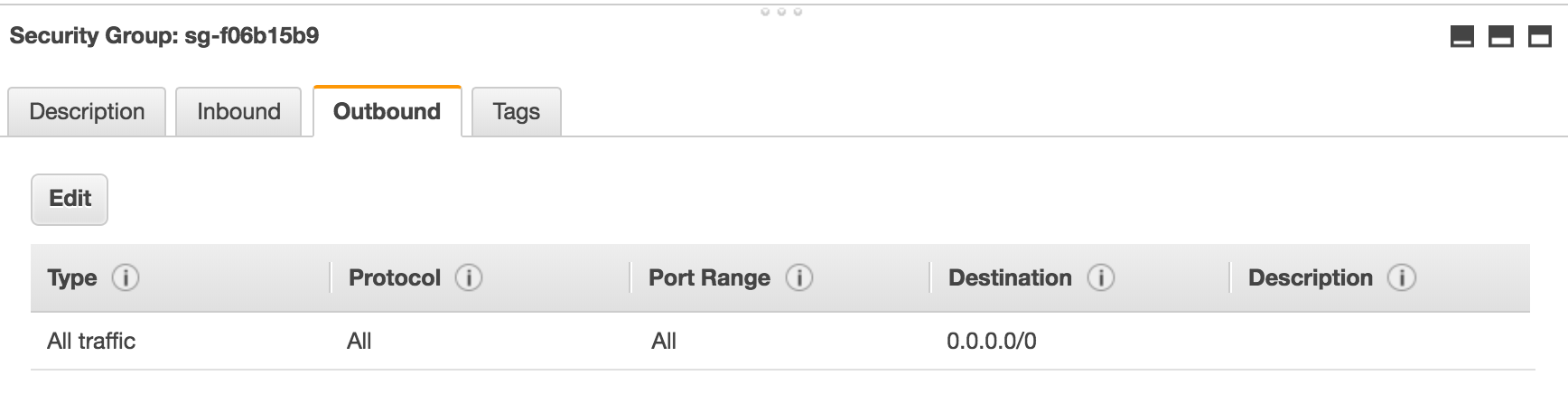
### Now find the security group (search for “EC2” in AWS) for the same VPC

### 

### In the future, you’ll want to make sure you lock down your security group so that you prevent hackers and unauthorized uses. However, for now, to make sure you can access your cluster, make sure they’re open to the public by checking the tabs for both inbound:

### 

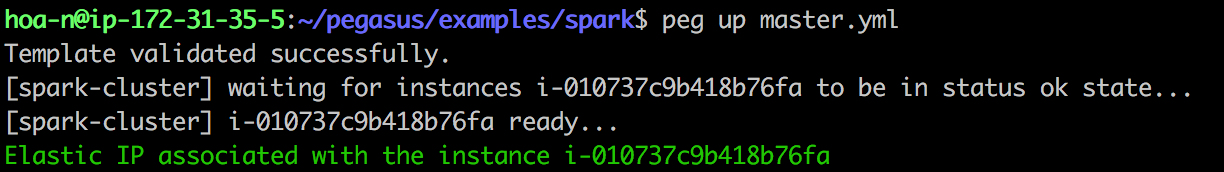
### and outbound traffic:



### Once your .yml files are configured, you’ll notice that the tag-name in the .yml file is spark-cluster. That is the name of your cluster. Now execute:

peg up master.yml

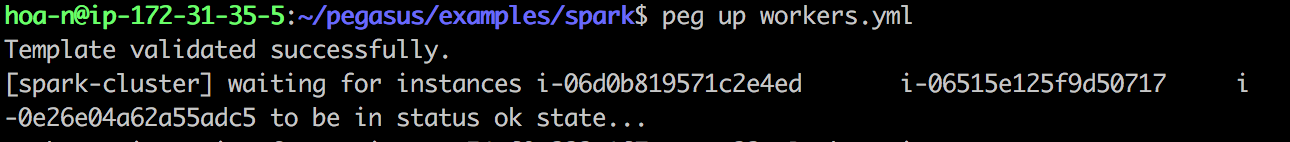
### Be patient and you should see something like this:



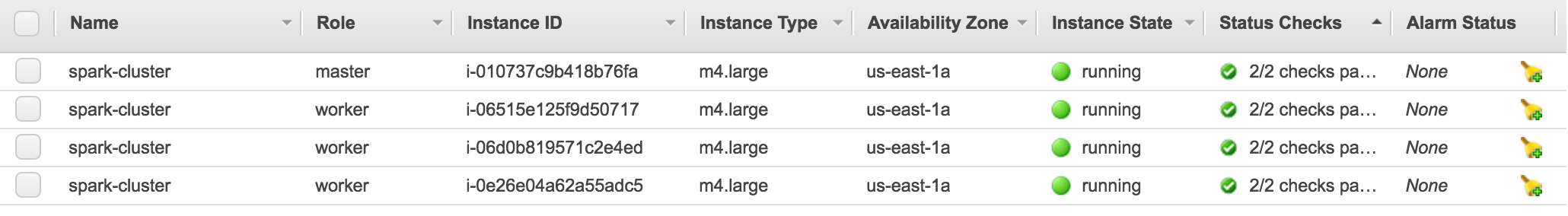
### Now execute this command:

peg up workers.yml

### And you should see something like this



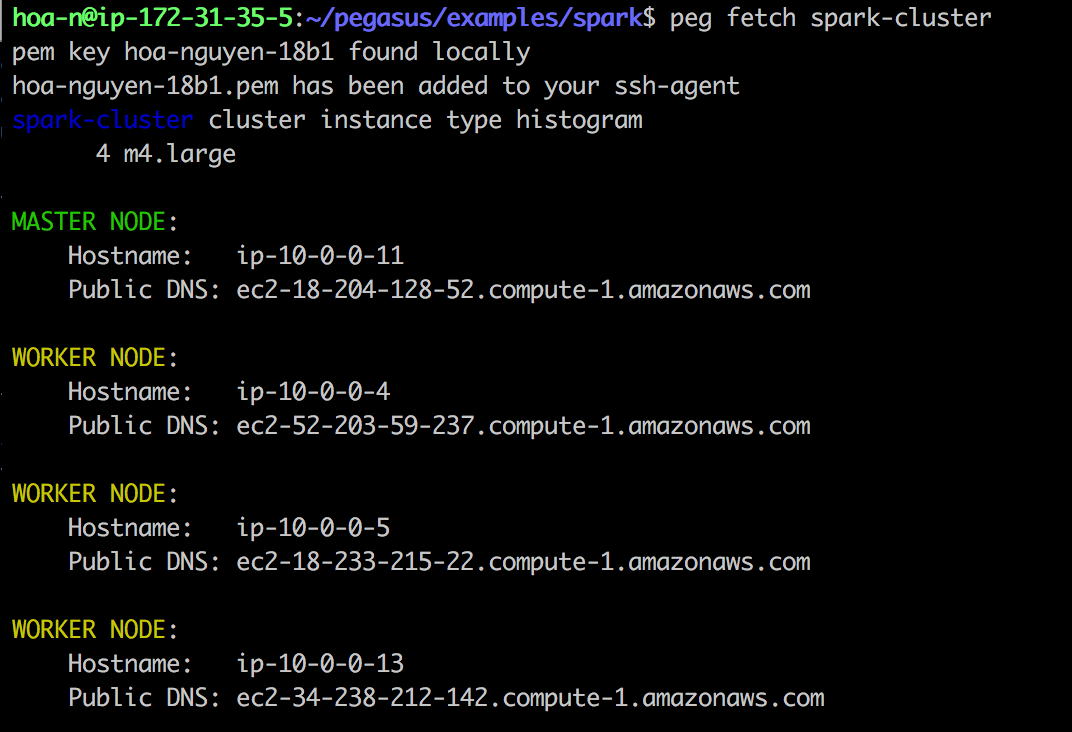
### The above output shouldn't be all that you see (I was hasty in my screen grab). It could take several minutes for the peg up command to complete. Once it does, you should be able to see the machines active on your AWS console:



### To verify your cluster is up, locally retrieve the details of your cluster and proceed with installing technologies, execute

peg fetch spark-cluster

### This is what you should see:



### Now you should start installing. The following three commands are crucial and need to be executed before you do anything else:

peg install spark-cluster ssh

peg install spark-cluster aws

peg install spark-cluster environment

### Only once those three commands are executed, now you can start installing technologies

### Install and start Hadoop/HDFS:

peg install spark-cluster hadoop

peg service spark-cluster hadoop start

### Install and start Spark:

peg install spark-cluster spark

peg service spark-cluster spark start

### Install and start Kafka

peg install spark-cluster zookeeper

peg service spark-cluster zookeeper start

peg install spark-cluster kafka

peg service spark-cluster kafka start

(note that there is a bug and the last command might just hang for no good reason. If so, just <ctrl-c> out of it)

Note that you can go through some of the [Kafka](https://github.com/InsightDataScience/data-engineering-ecosystem/wiki/Kafka) and [advanced Kafka](https://github.com/InsightDataScience/data-engineering-ecosystem/wiki/Kafka-advanced) exercises in the Insight wiki minus any installation instructions. I also believe the Kafka manager instructions are out of date. Beware that some of the instructions refer to older installations.