DSC102

EC2 Set up

OS: Ubuntu Server 20.04 LTS (HVM), SSD Volume Type

Instance Type: t2.xlarge with 4cpus and 16 Gigabytes RAM

Configuration Detail: Use spot instance. Other goes as default

Storage: 32 Gigabytes of SSD.

EC2 packages installed:

1. python3
2. Aws command lines interface
3. Pandas
4. Pyarrows
5. S3fs
6. Dask --general
7. Dask -- Machine learning
8. Sklearn

Discussion on trade off and set-up: I open a spot instance in order to save money

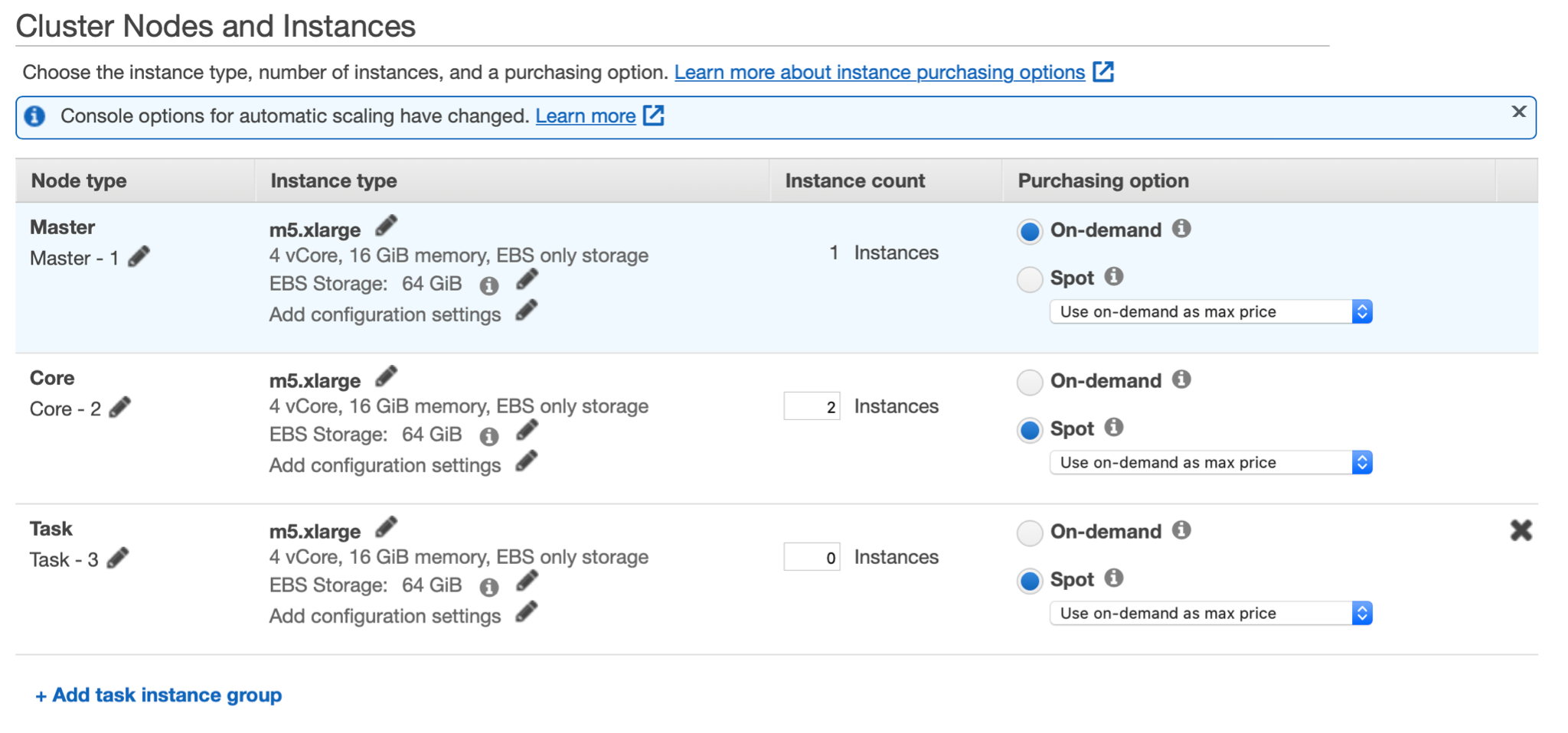
I also choose a 32 GBs SSD because the dataset we are dealing with is huge. To be more specific, we need to make sure we have a room for the data set itself and all packages needed for our feature engineering step and model deploying step.

16 GBs of ram ensure that we have sufficient memory to process the Origination dataset.

We can scale up the ram and storage if we are dealing with a larger data set.

**EMR Set Up (Not mentioned go by defaults):**

1. Software Configuration: emr-5.33.0 (default); Select Hadoop 2.10.1, Hive 2.3.7, Hue 4.9.0, Pig 0.17.0, and Spark 2.4.7. Although we don’t face any trade offs here, possible tradeoffs may be that the newer version might bring a lot of new changes that your previous code would not work.
2. Steps: Cluster enters waiting state
3. Cluster Composition: uniform instance group
4. Networking: as default
5. Cluster Nodes and Instances: shown on the screen shot.

****

We choose the master on-demand and set the cores on Spot. The current on-demand cost is $0.192 per instance/hr, while the on Spot cost is only $0.078 per instance/hr (more than $0.1 cheapter). **Tradeoff**: Although our work might be interrupted, the cost can be greatly reduced using Spot.

If we face a larger volume of data, we can scale up by changing the instance type and adding more instance count on the cluster nodes and instances part. Also, we may need to change the purchasing option from Spot to demand to on-demand to make sure that our project goes smoothly. :)