Answers (/community/s/group/0F90L0000001NDBSA2/) — elleg _ (/community/s/profile/0050L0000093B2OQAU) (Customer) asked a question.
July 3, 2014 at 12:20 PM (/community/s/question/0D50L00006BIubVSAT/mapr-cluster-instances-on-amazon-emr)

MapR cluster instances on Amazon EMR

Hello,

We are trying to estimate (a very rough ballpark) the cluster costs if we were to run mapr on amazon EMR. I see the following cluster recommendation in your docs:

Standard 100TB Rack Configuration

20 standard nodes

(20 x 12 x 2 TB storage; 3x replication, 25% margin)

48-port 1 Gb/s rack-top switch with 4 x 10Gb/s uplink

Add second switch if each node uses 4 network interfaces

To grow the cluster, just add more nodes and racks, adding additional service instances as needed. MapR rebalances the cluster automatically.

However, I am unable to correlate this back to the instances available on Amazon on [this][1] page.

1) They have listed Standard On-Demand instances, and in your docs, you mention "standard" nodes. Are these the same? In terms of actual instance types on EC2, is that an m1 or m3 or something else? There is also "High-Storage On-Demand Instances" in that list. Should we be using something like those for storage instead? The storage specs of standard and high-storage seem quite different, so we are not sure which one would be recommended.

2) When amazon talks about the MapR + EMR cost, does the EMR cost affect only things like instances running m/r jobs?

Thanks for your help!

[1]: http://aws.amazon.com/elasticmapreduce/mapr/pricing/ (http://aws.amazon.com/elasticmapreduce/mapr/pricing/) (https://community.mapr.com/external-link.jspa?url=http%3A%2F%2Faws.amazon.com%2Felasticmapreduce%2Fmapr%2Fpricing%2F (https://community.mapr.com/external-link.jspa?url=http%3A%2F%2Faws.amazon.com%2Felasticmapreduce%2Fmapr%2Fpricing%2F))

4 answers 43 views

^ Upvote Answer Share

Top Rated Answers

snelson _ (/community/s/profile/0050L0000093DAVQA2) (Customer)
5 years ago

We use [this][1] calculator to estimate cluster costs. Go to "Amazon Elastic MapReduce" on the left of the page. Choose a number of instances, choose MapR M7 (or whatever edition you want) as distribution and your instance type. The cost estimation it gives you includes the EC2 compute costs and the included ephemeral storage that is attached to each node.

At my company we are in a situation where our storage needs outweigh our compute needs. Instead of using the high storage instances (which cost a fortune!) we attach EBS volumes to the nodes in the cluster after the cluster has started. It's important to realize the EMR is just a handy way to provision and license MapR instances. It really just leverages EC2. You're able to do just about everything with the nodes as you can do with standard EC2 instances, which includes attaching EBS volumes. If you decide to add EBS volumes, incorporating them into the cluster is as easy as SSH-ing into each node, modifying /tmp/disks.txt to "/dev/xvdf" (replace xvdf with whatever you've attached the disk as) and then running `/opt/mapr/server/disksetup -F /tmp/disks.txt`. [This guide][2] documents that process.

[1]: http://calculator.s3.amazonaws.com/index.html (http://calculator.s3.amazonaws.com/index.html) (https://community.mapr.com/external-link.jspa?url=http%3A%2F%2Fcalculator.s3.amazonaws.com%2Findex.html (https://community.mapr.com/external-link.jspa?url=http%3A%2F%2Fcalculator.s3.amazonaws.com%2Findex.html))

[2]: http://doc.mapr.com/display/MapR/Setting+Up+Disks+for+MapR (http://doc.mapr.com/display/MapR/Setting+Up+Disks+for+MapR) (https://community.mapr.com/external-link.jspa?url=http%3A%2F%2Fdoc.mapr.com%2Fdisplay%2FMapR%2FSetting%2BUp%2BDisks%2Bfor%2BMapR (https://community.mapr.com/external-link.jspa?url=http%3A%2F%2Fdoc.mapr.com%2Fdisplay%2FMapR%2FSetting%2BUp%2BDisks%2Bfor%2BMapR))

✓ Selected as Best Upvote

All Answers

snelson _ (/community/s/profile/0050L0000093DAVQA2) (Customer)
5 years ago

We use [this][1] calculator to estimate cluster costs. Go to "Amazon Elastic MapReduce" on the left of the page. Choose a number of instances, choose MapR M7 (or whatever edition you want) as distribution and your instance type. The cost estimation it gives you includes the EC2 compute costs and the included ephemeral storage that is attached to each node.

Show More

✓ Selected as Best Upvote Reply

snelson _ (/community/s/profile/0050L0000093DAVQA2) (Customer)
5 years ago

You'll have to decide for yourself what the proper cluster size and node size is and how many disks per node. I think the only real way to decide this is if you try it and test. Or you could start by estimating how many concurrent readers and writers you plan to have to start. That could determine how many virtual CPU you'll need in your cluster. Once you determine that number, you can choose to have fewer instances with more CPU power, or more instances with less CPU power. I prefer fewer machines because it's easier to maintain. Then you plan how your cluster will grow. For example: how many volumes per machine before you add a new node?

We have done a bunch of performance tests, and we haven't found EBS to be slow. With EBS you can explicitly provision higher IOPS if your disks become a bottleneck (although if that happened I would suspect your table schema, not your disks), so it seems silly to me to make a blanket statement that EBS are slow. I would recommend using the new General Purpose SSD (gp2) EBS volumes because they are very fast 3000+ IOPS, and they are cheaper than the magnetic PIOPS volumes now.

As far as "If you are using MapR on Amazon EMR, you do not have to use this procedure" I believe it's referring to the ephemeral storage that comes attached to the nodes.

Upvote Reply

elleg _ (/community/s/profile/0050L0000093B2OQAU) (Customer)
5 years ago

Thanks, @snelson!! This is very helpful for our estimation. We started with the calculator as well, but didn't know what type of instances would work best. Like yourself, our product will have a higher storage need than compute need (at least initially). We intend to store raw data long term, and have random access based on a time range. So we might want to do something similar.

I looked into the guide you mentioned for attaching disks. This statement:

The following procedures are intended for use on physical clusters or Amazon EC2 instances. On EC2 instances, EBS volumes can be used as MapR storage, although performance will be slow.

If you are using MapR on Amazon EMR, you do not have to use this procedure; the disks are set up for you automatically.

Have you noticed the performance slowness they mention above? If so, what kind of workarounds did you need?

Thanks again!

Upvote Reply

elleg _ (/community/s/profile/0050L0000093B2OQAU) (Customer)
5 years ago

Weird, that I didn't get an email notification for this response from you. I only saw it when I came back to this page. Thanks, we will run some tests before we settle on the appropriate sizes for our cluster and nodes. Your information has been extremely helpful. Thanks a lot!!

Upvote Reply

Login to answer this question