Answers (/community/s/group/0F90L0000001NDBSA2/) — akchang _ (/community/s/profile/0050L00000938xZQAQ) (Customer) asked a question.
December 29, 2012 at 4:53 PM (/community/s/question/0D50L00006BIuy3SAD/storage-pool-stripe-size)

storage pool stripe size

what is the default stripe size of MapR Storage Pools? Is it configurable?

4 answers 90 views

Upvote Answer Share

Top Rated Answers

Peter Conrad (/community/s/profile/005E0000000IDakIAG) (Employee)
6 years ago

The storage pool stripe size is configurable. A storage pool uses all its disks simultaneously, which increases data bandwidth--in a storage pool of three disks, the bandwidth for reads and writes is triple that of a single disk.

It is true that increasing the stripe width provides a performance benefit -- putting all disks on a node into the same storage pool would maximize the data throughput of that node. However, there is a potential reliability penalty. Because a storage pool is treated as a unit, failure in one disk in the storage pool will require that the contents of the entire storage pool be re-replicated from other nodes. With very large storage pools, this can take quite some time and if the wrong two disks fail during this replication time, you could lose data. Simulations show that having 10 disk storage pools made up of disks that fail at the rate of 30% per year on nodes with 1Gb/s networking will result in a mean time to data loss of less than a decade. Using 3 disk storage pools under this same scenario results in >1000 year mean time to data loss. Normal disks will fail much less than this, but when data loss is concerned, it is best to be somewhat conservative.

A default size of 3 disks per storage pool represents a sensible compromise between performance and safety.

Selected as Best Upvote

All Answers

MC Srivas (/community/s/profile/0050L0000093EJHQA2) (Customer)
6 years ago

MapR will assemble disks in groups of 3 or 2 drives to form storage pools. For example, if disksetup is invoked with 10 drives, it will create 4 storage pools of 3, 3, 2, and 2 drives respectively. If invoked with 9 drives, it will create 3 storage pools of 3 drives each.

Upvote Reply

Peter Conrad (/community/s/profile/005E0000000IDakIAG) (Employee)
6 years ago

The storage pool stripe size is configurable. A storage pool uses all its disks simultaneously, which increases data bandwidth--in a storage pool of three disks, the bandwidth for reads and writes is triple that of a single disk.

It is true that increasing the stripe width provides a performance benefit -- putting all disks on a node into the same storage pool would maximize the data throughput of that node. However, there is a potential reliability penalty. Because a storage pool is treated as a unit, failure in one disk in the storage pool will require that the contents of the entire storage pool be re-replicated from other nodes. With very large storage pools, this can take quite some time and if the wrong two disks fail during this replication time, you could lose data. Simulations show that having 10 disk storage pools made up of disks that fail at the rate of 30% per year on nodes with 1Gb/s networking will result in a mean time to data loss of less than a decade. Using 3 disk storage pools under this same scenario results in >1000 year mean time to data loss. Normal disks will fail much less than this, but when data loss is concerned, it is best to be somewhat conservative.

A default size of 3 disks per storage pool represents a sensible compromise between performance and safety.

Selected as Best Upvote Reply

Jon Strayer (/community/s/profile/0050L0000093AiHQAU) (Customer)
2 years ago

Hi Peter,
Would you happen to know the model that was used to make those estimates? It seems to me that I could live with a 100 year mean time to data loss. How do the mean time to data loss change with the number of drives? Or does anyone know how using AWS EBS effects this calculation?

Upvote Reply

Mufeed Usman (/community/s/profile/005E0000004p1nkIAA) (Employee)
2 years ago

I'd ascribe the meantime calculation based more on the probability of SPs getting affected when you have more disks under a fewer SP umbrella when the same is compared to fewer disks under a large number of SPs.
To my knowledge, this calculation should be platform agnostic and hence should not make any difference whether the setup is on-premise or cloud-based. Of course, you could have interfering parameter(s) like the reliability and durability of the so-called platform and components used that may affect the theoretical calculations.

Upvote Reply