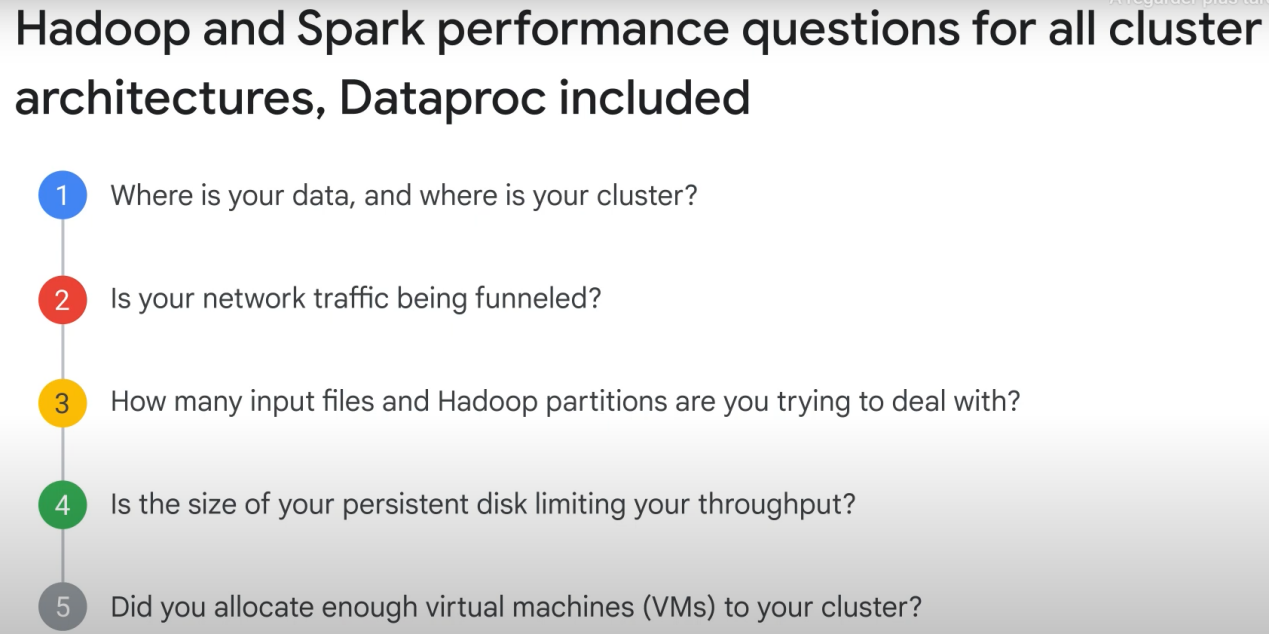
Make sure the cloud storage bucket is in the same location/region than your dataproc region.

\* No more than 10K input files. If you have more, try to join the data.

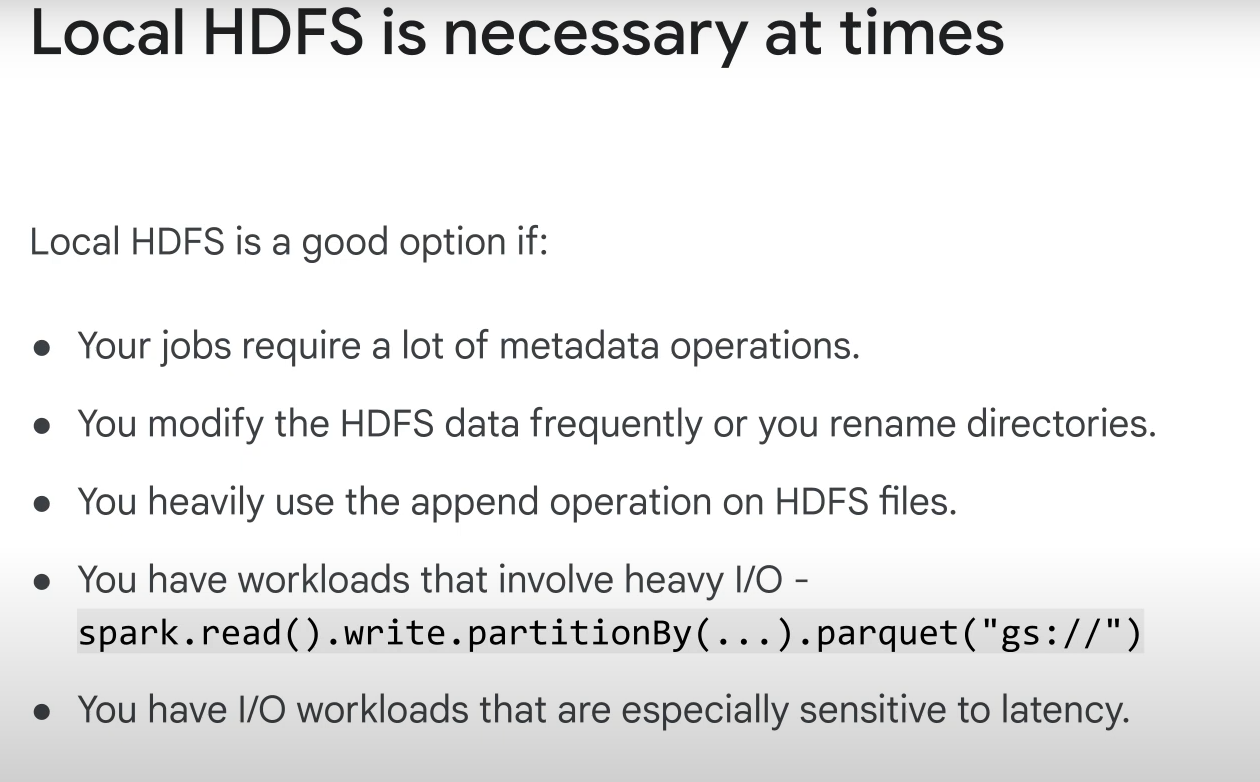
\* if you have 50K+ hadoop partions, increase the size of the blocks and reduce number of partitions

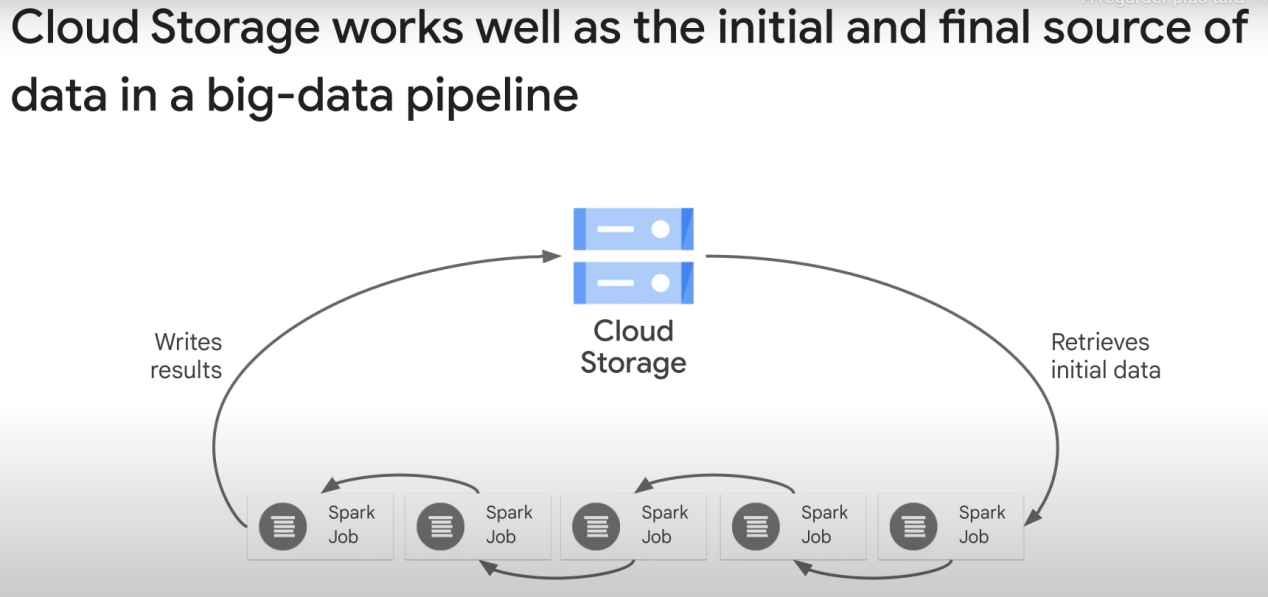
\* Check that persistent disk is not limiting your throughput

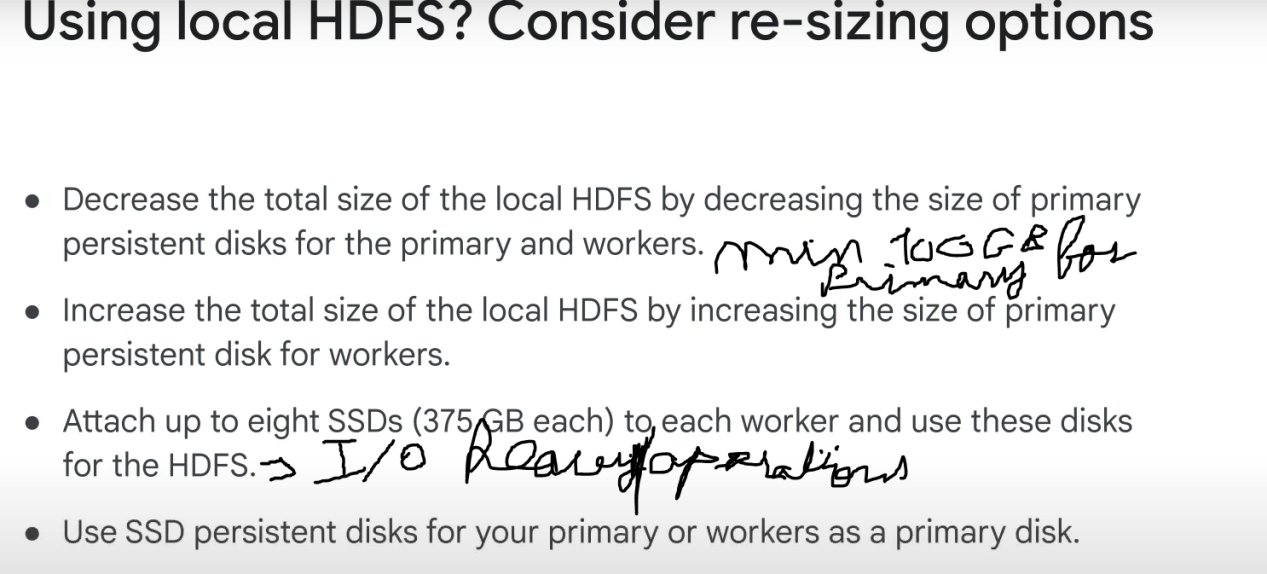
\* Do you have enough VMs allocated to your cluster. You can test and resize your cluster as needed.

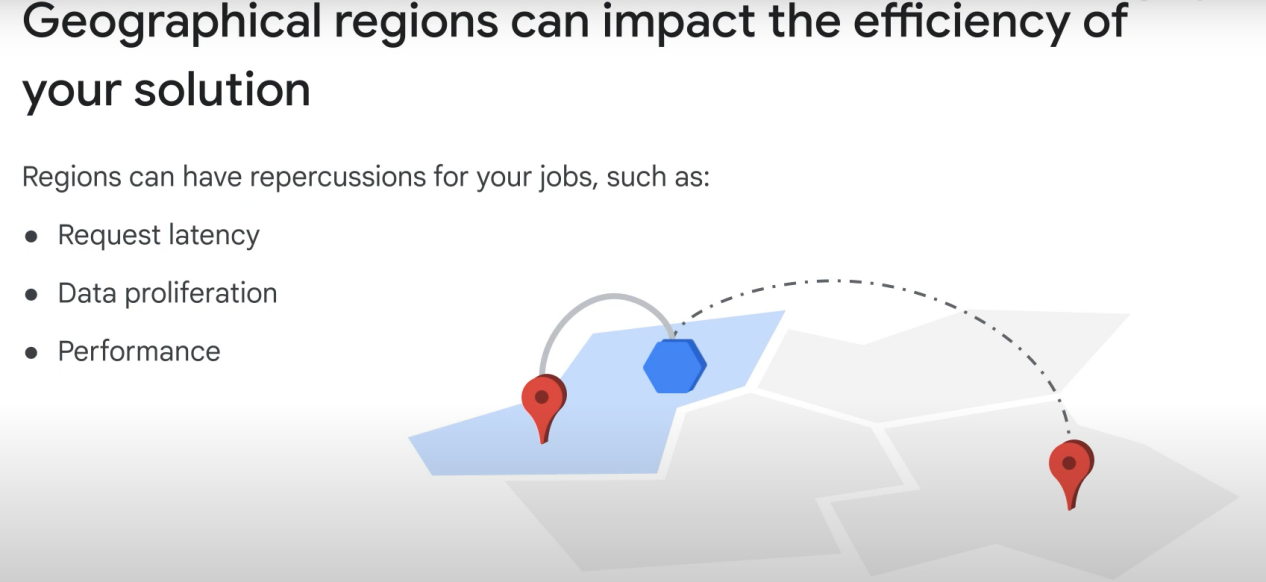


Optimizing Dataproc storage:

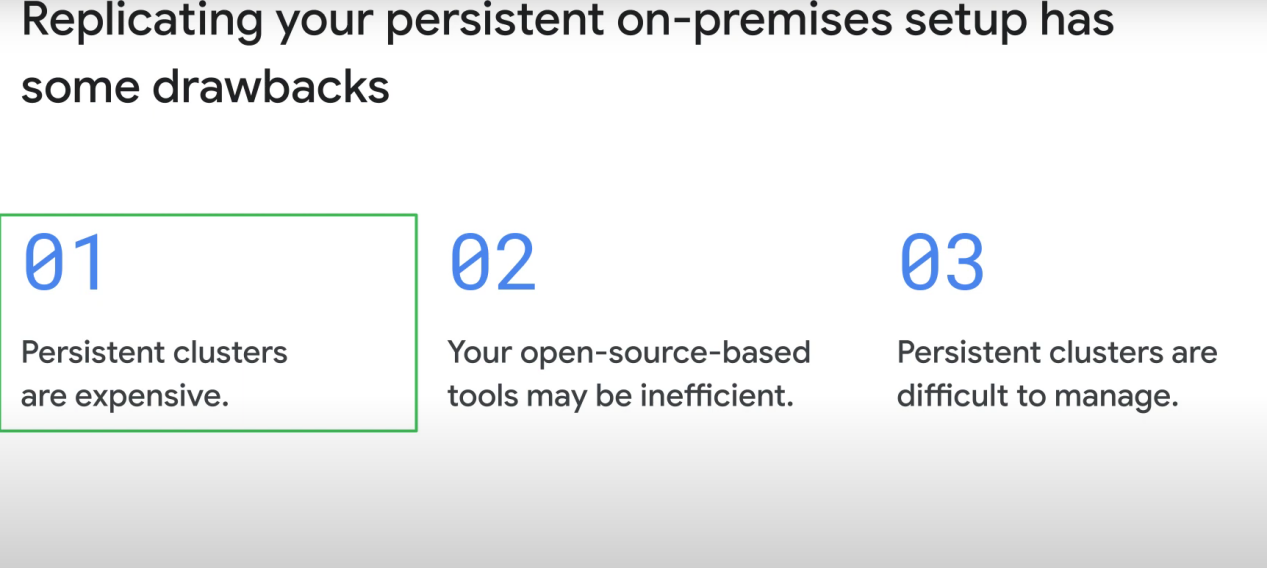




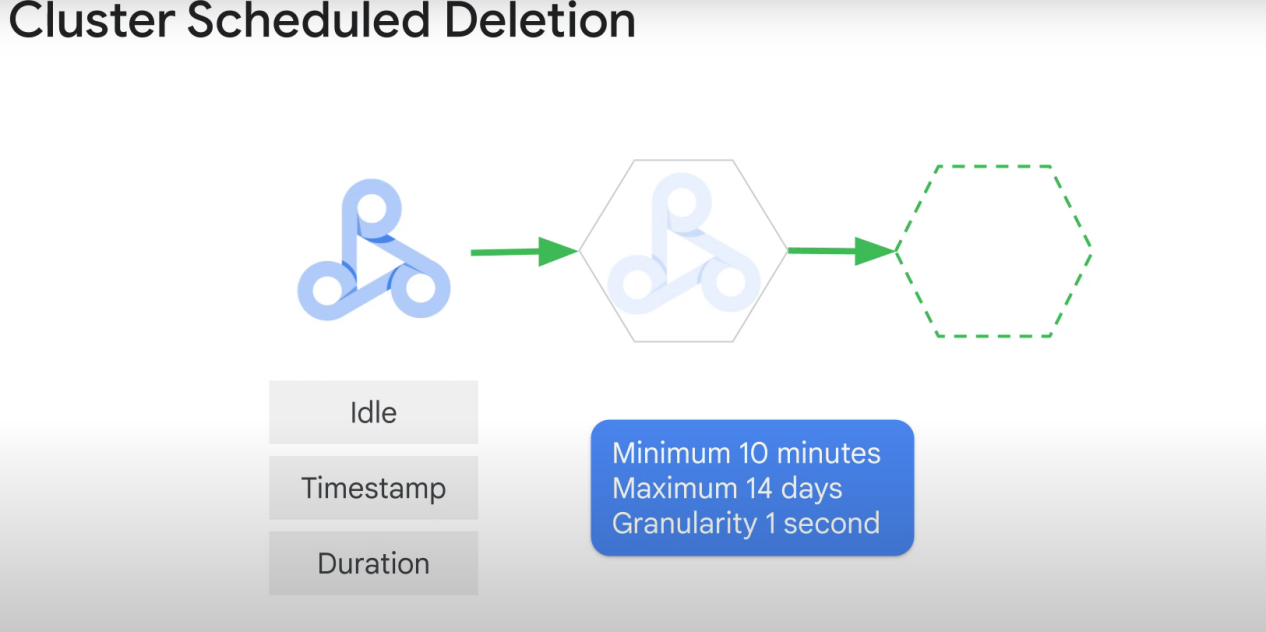


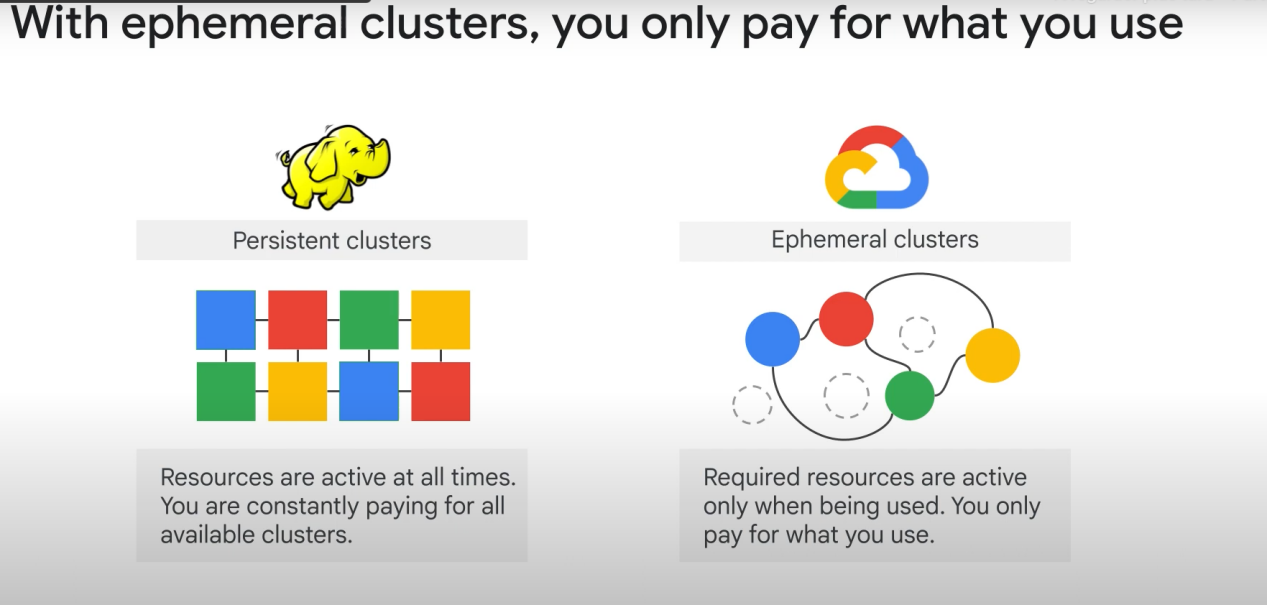


It’s dumb to keep all your data in a persistent (persistent cluster), as the point of cloud storage is to save you money by only running clusters when you need processing. Just store your data on google cloud, and transfer it to the computing clusters when you need to do computations.



You can automatically delete clusters after a certain time of IDLE:





Jobs’ scoped clusters: separate dev, staging, and production jobs, each in one cluster

