

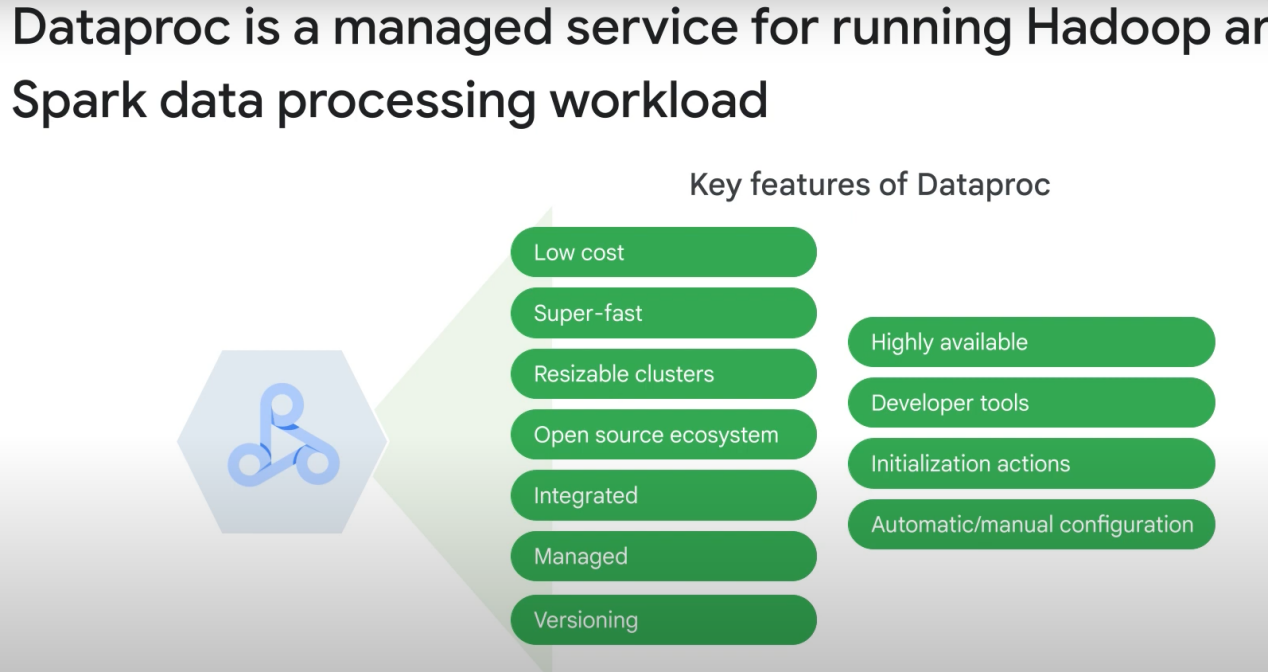
**Apache spark** is a declarative programming module: you just tell the system what you want, and it figures out how to implement it.

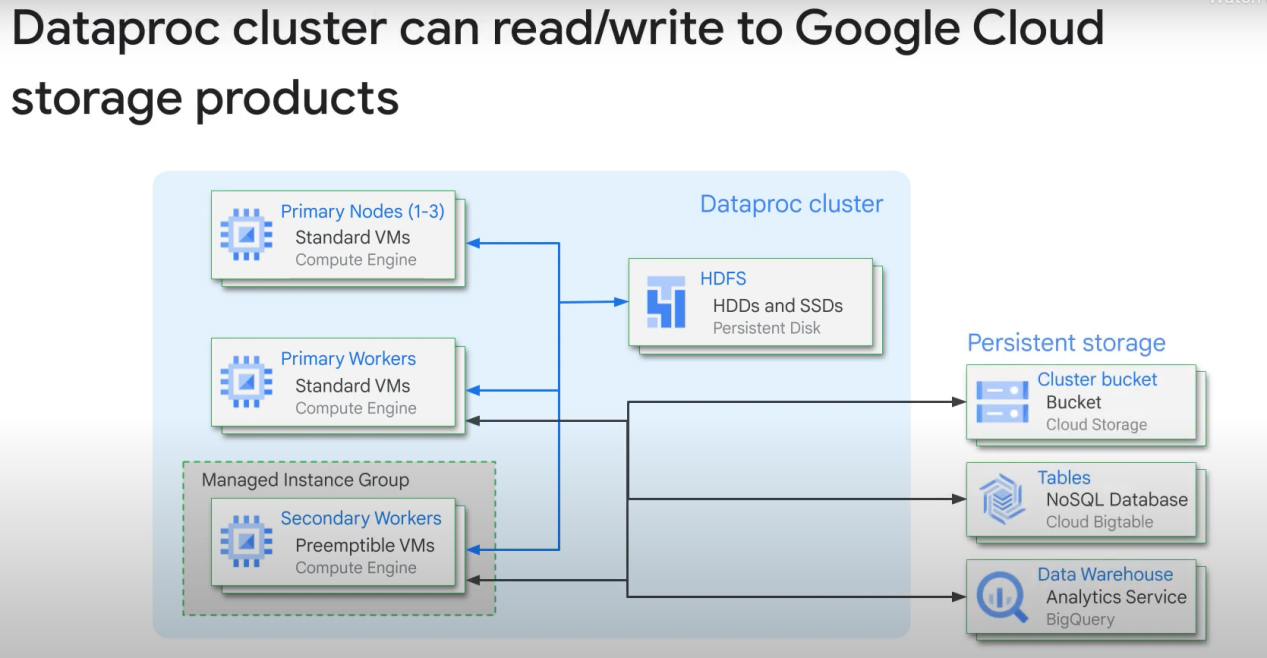
Why Use hadoop on google cloud with dataproc?

\* You can delete/Turn off clusters when not used (cheaper)

\* No need to learn anything. Just requires hadoop expertise.

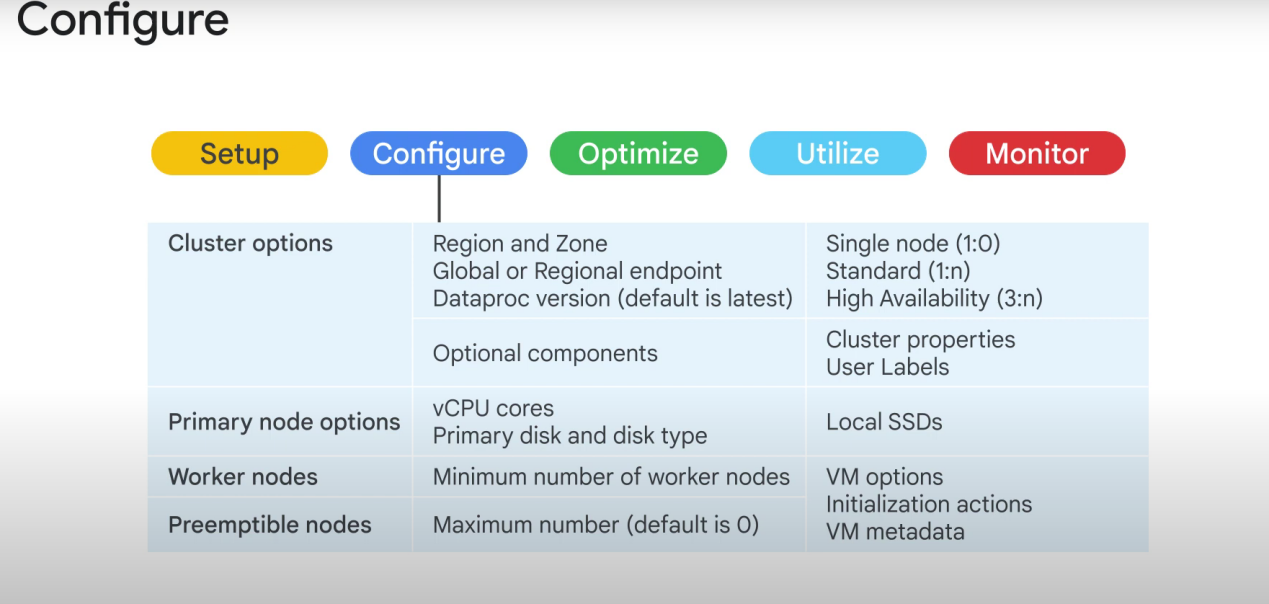
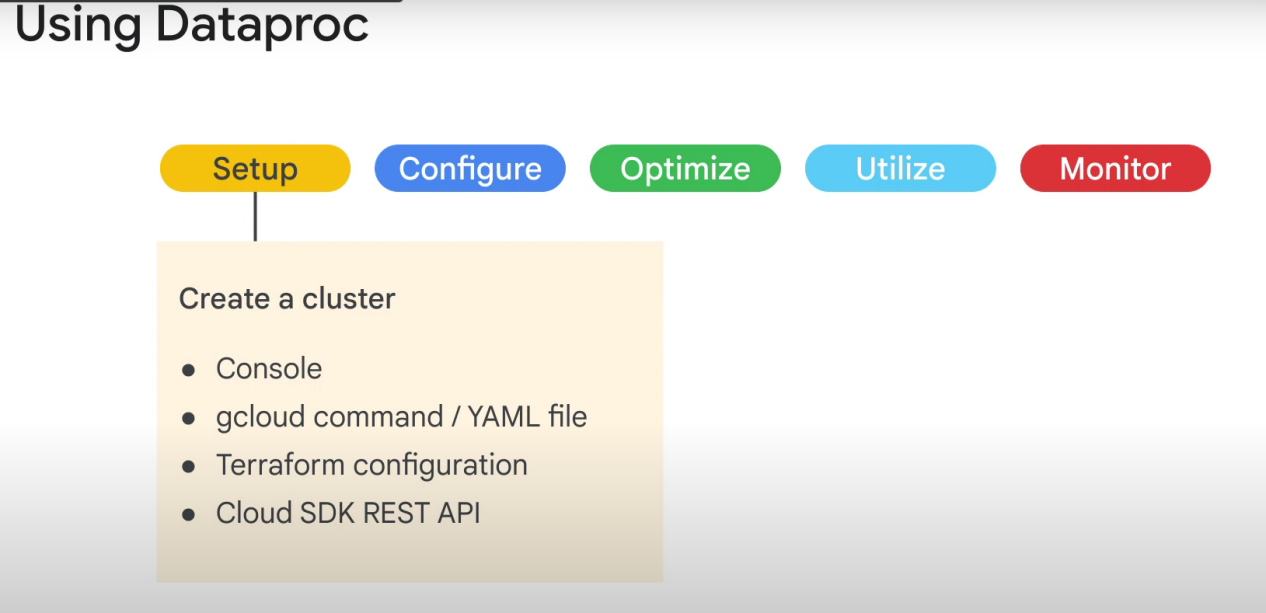
+

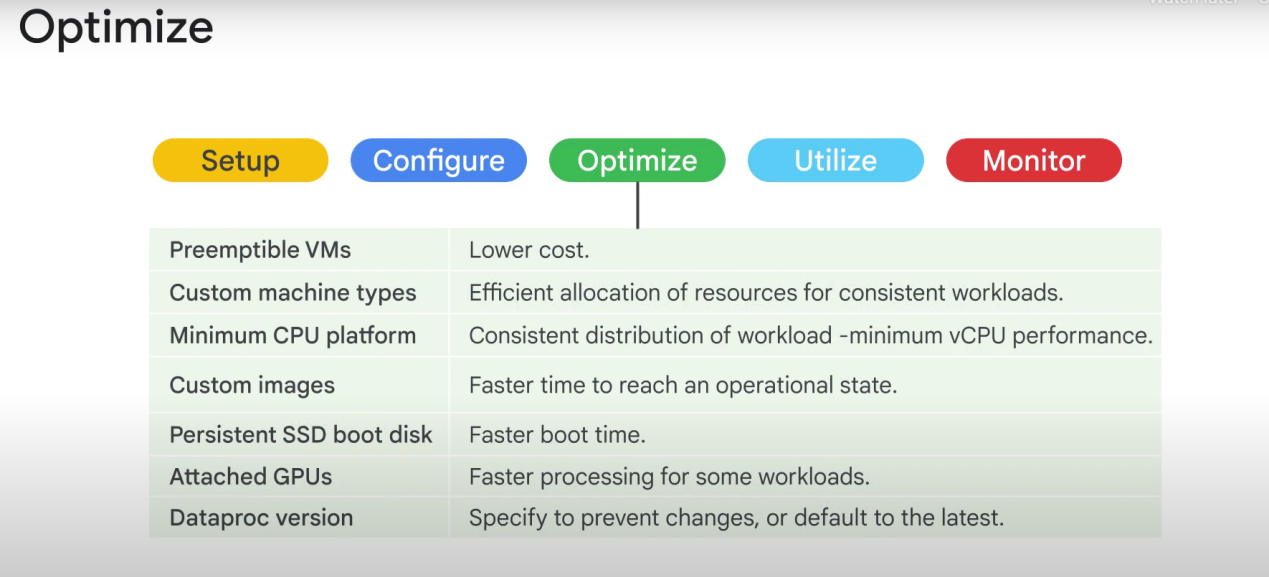


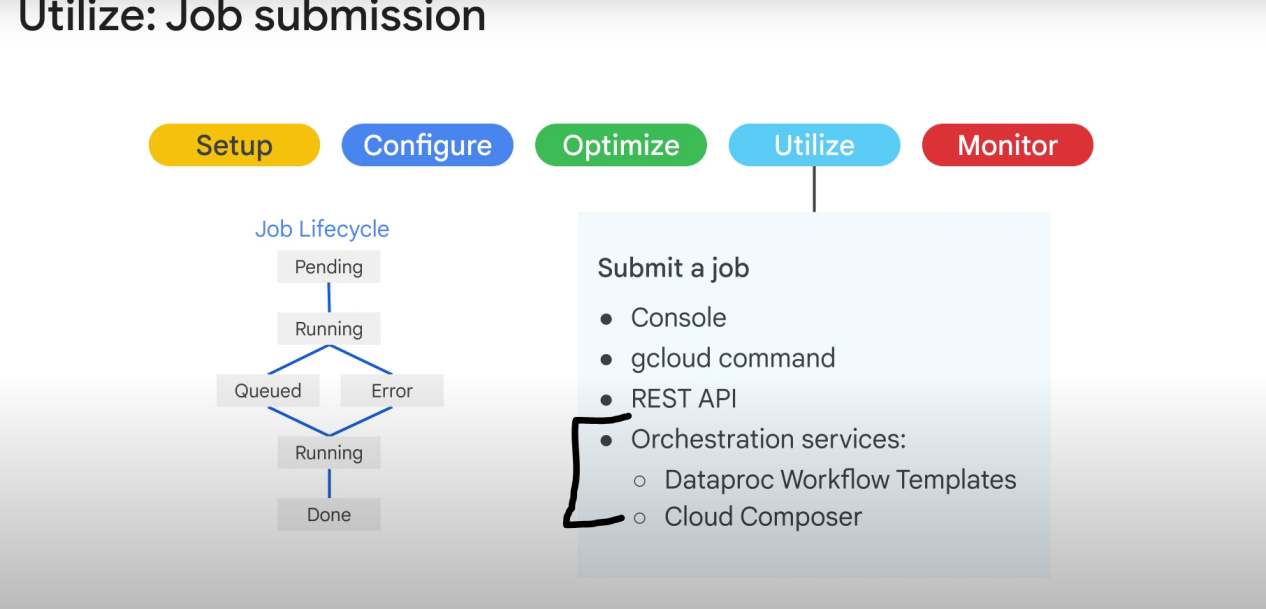


Standard Cluster mode:

|  |
| --- |
| C. Provides 1 primary and N workers |



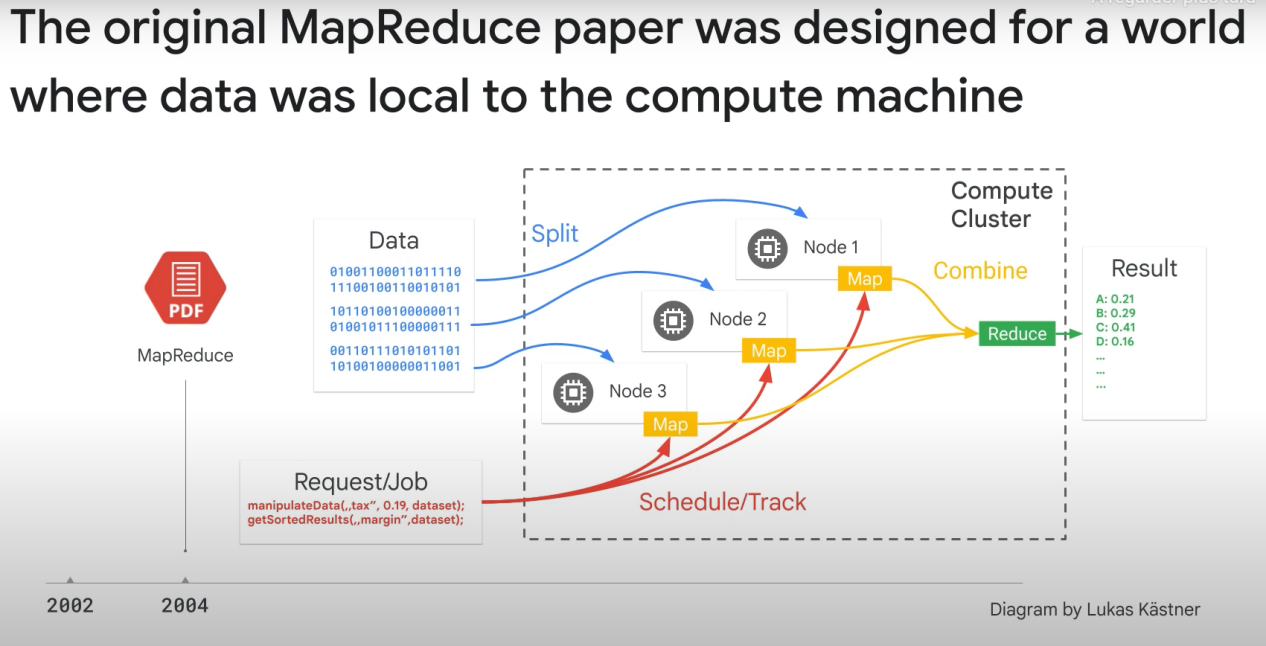




Note: don’t start a job directly from dataproc, because it will not save any metada for job and data management

Note 2: A job cannot be rerun. But you can create restartable jobs via terminal or the REST API.

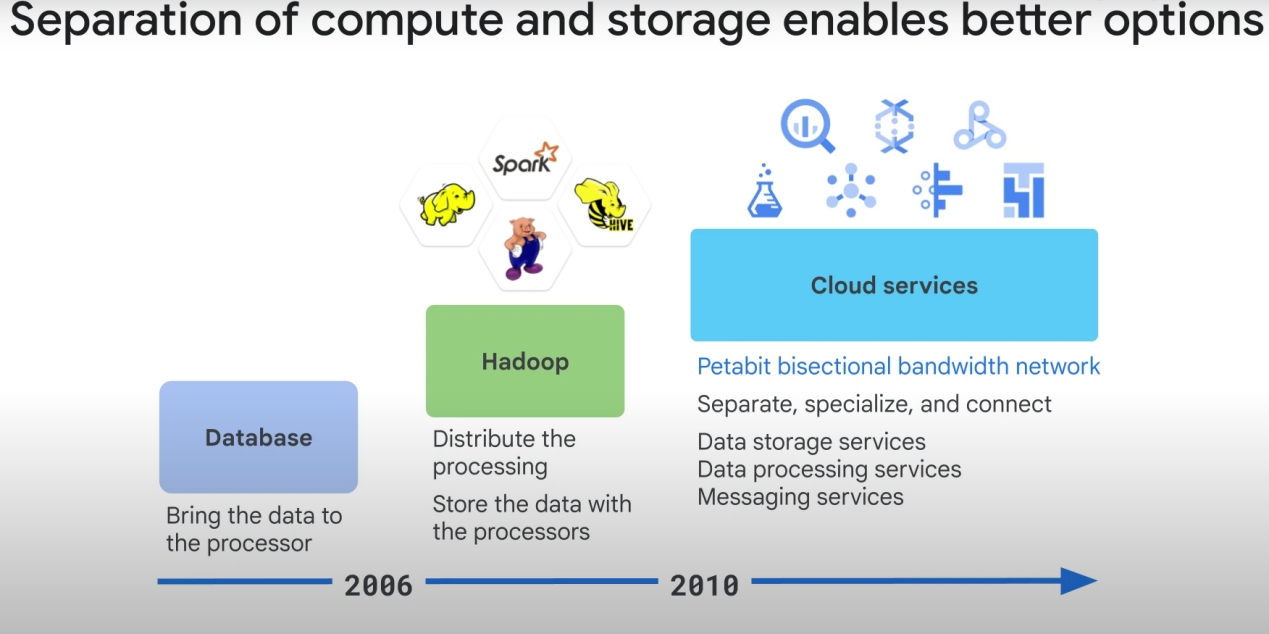
Cloud storage insteaf of HDFS (hadoop data file system)

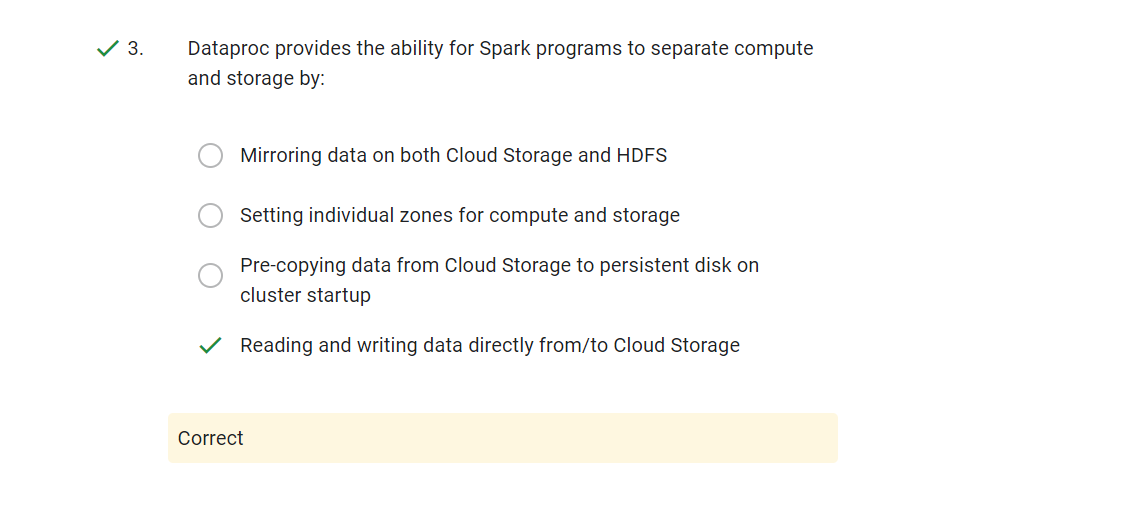


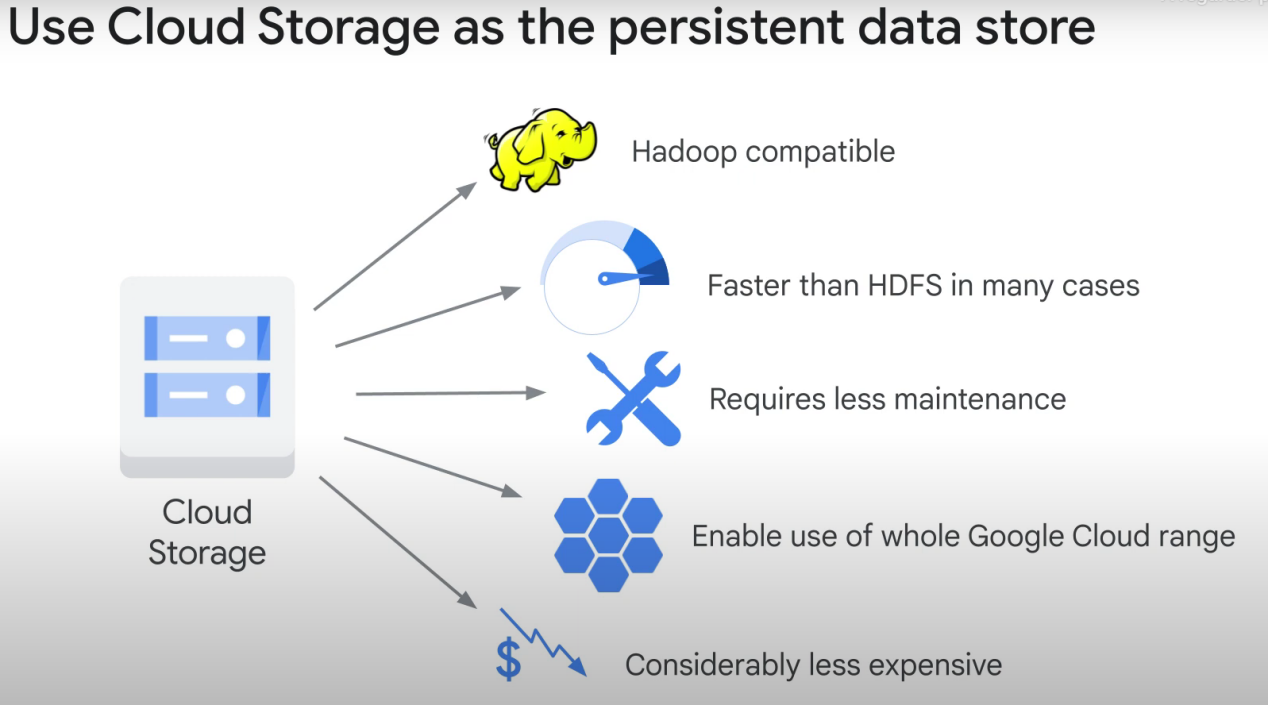
Data used to be stored in the same machine that does the computing to gain time ( the process was still slow).

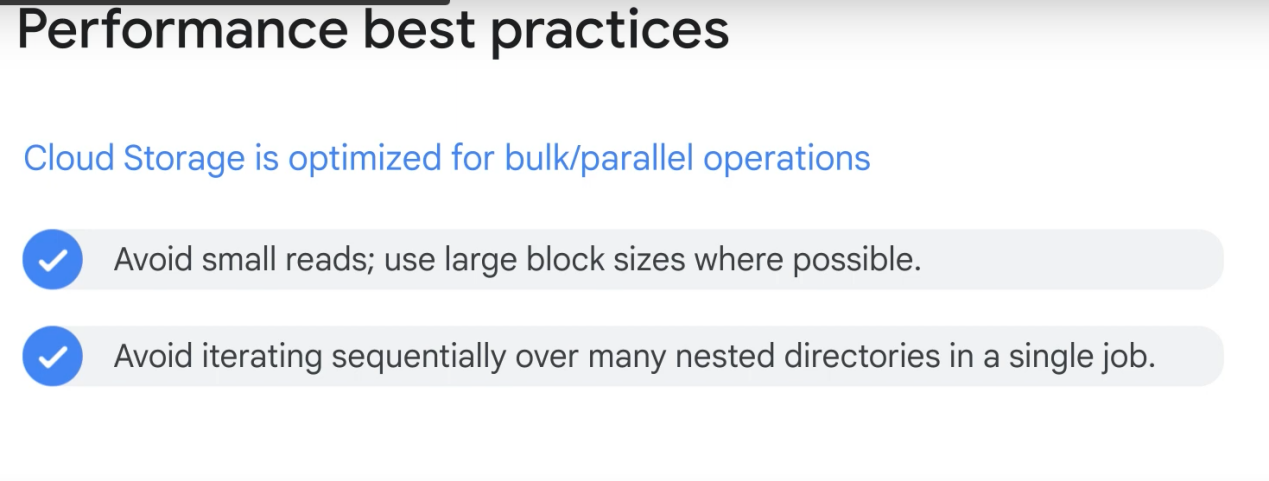
With cloud storage, we separated them. Google network’s provides Over **1petabit/sec** of bandwith. This is so fast that it makes no sense to copy the data to the computing machine, we just transfer it every time we need to process it.

When we separated storage and processing, we are now able to turn off computing clusters when they are not used == saves money.

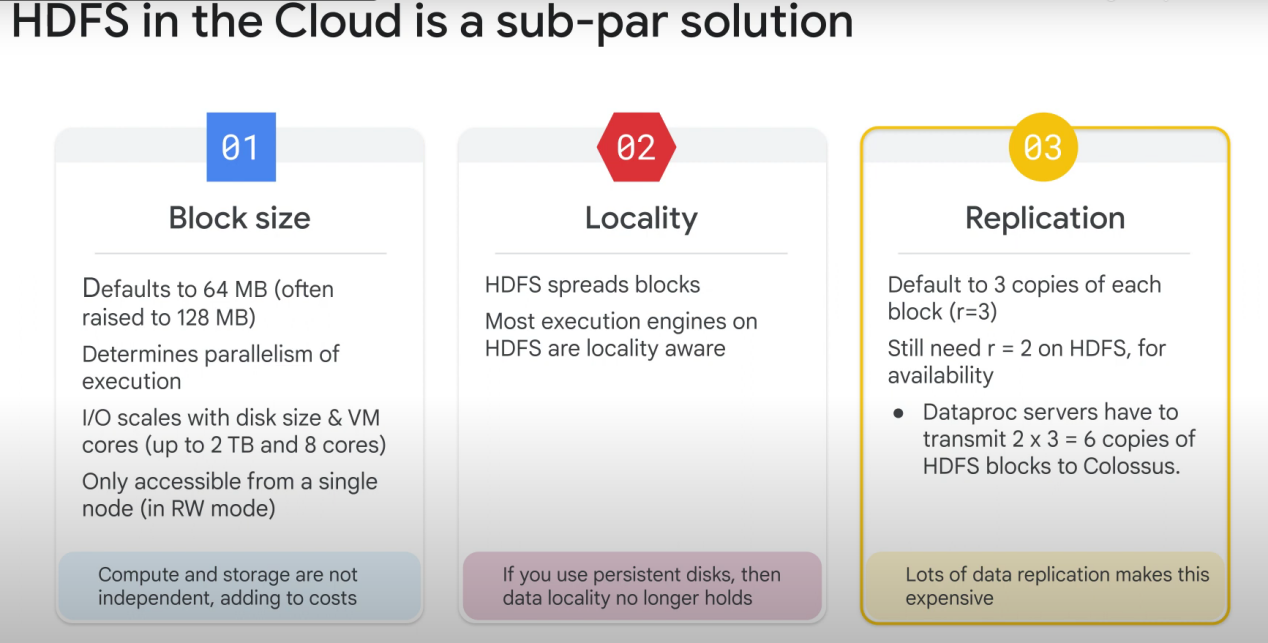








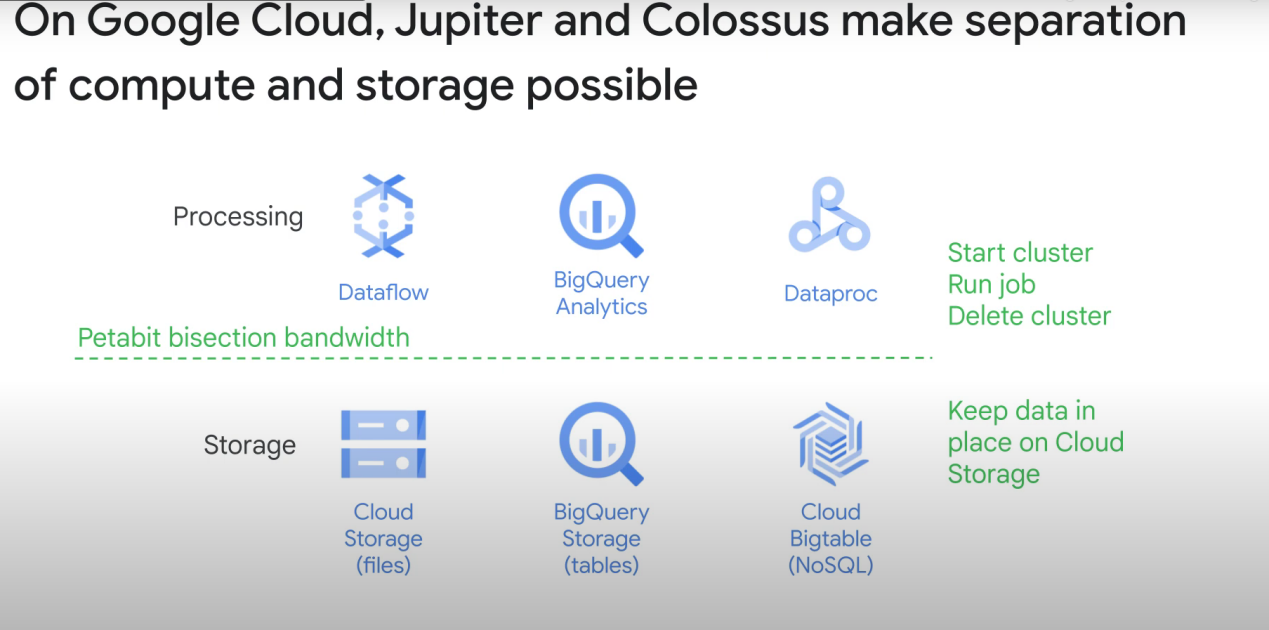
**If you have many very small blocks, it may be better to just use HDFS**



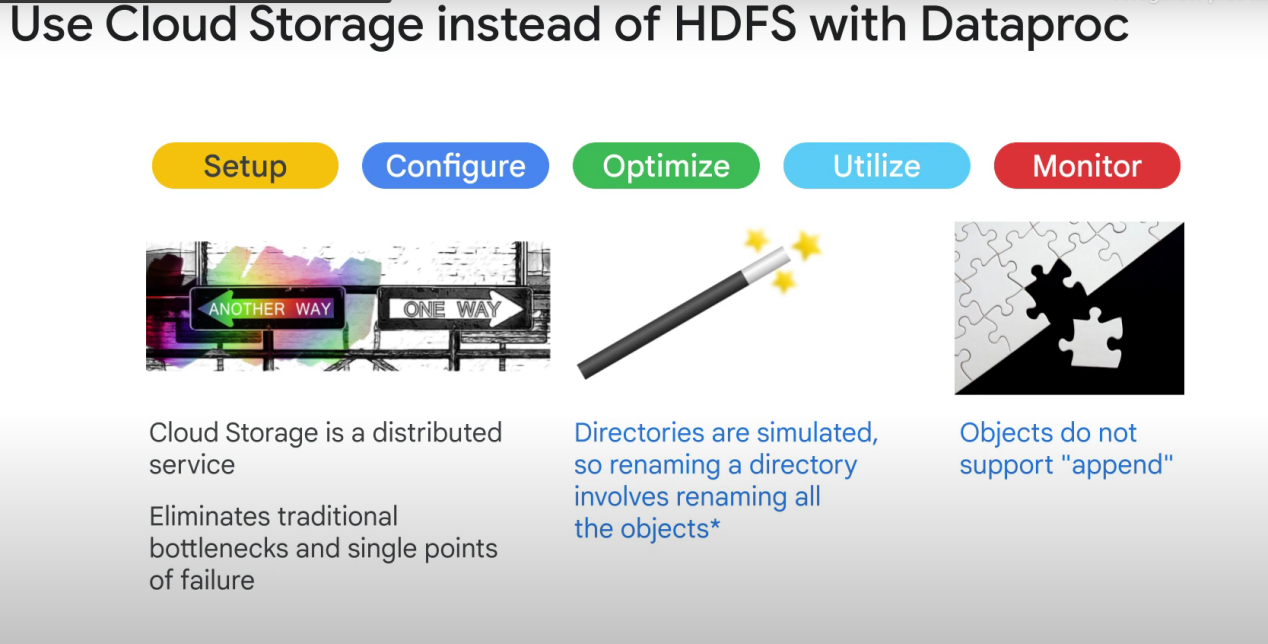
Data locality = edge computing (computing close to storage)

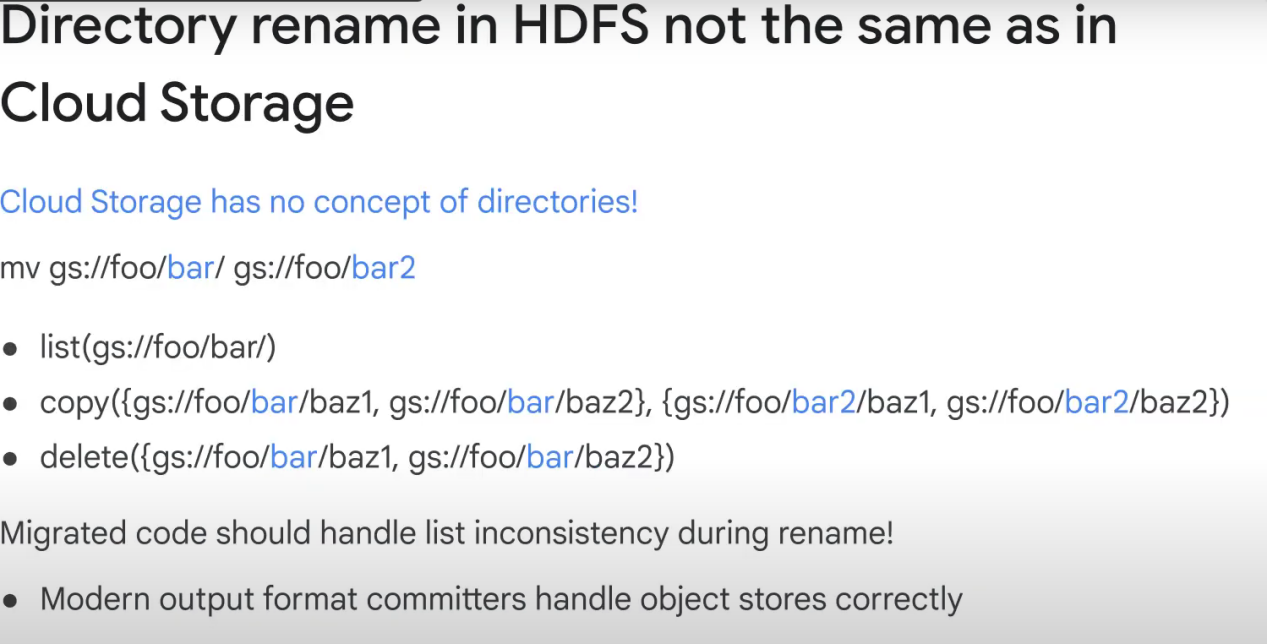
All the storage layer network = **colossus**

All the processing layer network = **Jupiter**



The bad things about dataproc:





Cloud storage isn’t a real file system. You’ll need to rename every object of a directory if you want to change the name of a directoy, changing its «path».