[Course Overview](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=4c10c956-67e1-4a00-9eeb-6a8928cc0bd0&clip=0&mode=live)

[Course Overview](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=4c10c956-67e1-4a00-9eeb-6a8928cc0bd0&clip=0&mode=live)

[Hi. My name is Janani Ravi, and welcome to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=4c10c956-67e1-4a00-9eeb-6a8928cc0bd0&clip=0&mode=live&start=0) [this course on Architecting Big Data](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=4c10c956-67e1-4a00-9eeb-6a8928cc0bd0&clip=0&mode=live&start=6.994500000000001" \t "psplayer) [Solutions Using Google Dataproc. A little](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=4c10c956-67e1-4a00-9eeb-6a8928cc0bd0&clip=0&mode=live&start=11.186428571428573) [about myself. I have a Master's degree in](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=4c10c956-67e1-4a00-9eeb-6a8928cc0bd0&clip=0&mode=live&start=14.246818181818183) [electrical engineering from Stanford and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=4c10c956-67e1-4a00-9eeb-6a8928cc0bd0&clip=0&mode=live&start=16.566090909090914) [have worked at companies such as](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=4c10c956-67e1-4a00-9eeb-6a8928cc0bd0&clip=0&mode=live&start=18.213600000000007) [Microsoft, Google, and Flipkart. At](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=4c10c956-67e1-4a00-9eeb-6a8928cc0bd0&clip=0&mode=live&start=20.25) [Google, I was one of the first engineers](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=4c10c956-67e1-4a00-9eeb-6a8928cc0bd0&clip=0&mode=live&start=21.83291666666667) [working on real-time collaborative editing](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=4c10c956-67e1-4a00-9eeb-6a8928cc0bd0&clip=0&mode=live&start=24.517333333333333) [in Google Docs, and I hold four patents](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=4c10c956-67e1-4a00-9eeb-6a8928cc0bd0&clip=0&mode=live&start=26.859) [for its underlying technologies. I](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=4c10c956-67e1-4a00-9eeb-6a8928cc0bd0&clip=0&mode=live&start=30.225111111111115)[currently work on my own startup,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=4c10c956-67e1-4a00-9eeb-6a8928cc0bd0&clip=0&mode=live&start=32.134) [Loonycorn, a studio for high-quality video](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=4c10c956-67e1-4a00-9eeb-6a8928cc0bd0&clip=0&mode=live&start=34.17400000000002) [content. In this course, you'll learn to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=4c10c956-67e1-4a00-9eeb-6a8928cc0bd0&clip=0&mode=live&start=37.00933333333334) [work with managed Hadoop on the Google](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=4c10c956-67e1-4a00-9eeb-6a8928cc0bd0&clip=0&mode=live&start=39.34115384615383) [Cloud and the best practices to follow for](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=4c10c956-67e1-4a00-9eeb-6a8928cc0bd0&clip=0&mode=live&start=41.201) [migrating your on-premises jobs to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=4c10c956-67e1-4a00-9eeb-6a8928cc0bd0&clip=0&mode=live&start=44.592272727272736) [Dataproc clusters. We'll study in some](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=4c10c956-67e1-4a00-9eeb-6a8928cc0bd0&clip=0&mode=live&start=46.9475) [depth how separation of storage and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=4c10c956-67e1-4a00-9eeb-6a8928cc0bd0&clip=0&mode=live&start=49.33299999999999) [compute allows you to utilize clusters](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=4c10c956-67e1-4a00-9eeb-6a8928cc0bd0&clip=0&mode=live&start=51.28599999999998) [more efficiently purely for processing](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=4c10c956-67e1-4a00-9eeb-6a8928cc0bd0&clip=0&mode=live&start=53.9722) [data and not for storage. We start off by](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=4c10c956-67e1-4a00-9eeb-6a8928cc0bd0&clip=0&mode=live&start=56.516200000000005)[creating a Dataproc cluster and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=4c10c956-67e1-4a00-9eeb-6a8928cc0bd0&clip=0&mode=live&start=59.771461538461516) [configuring firewall rules to enable us to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=4c10c956-67e1-4a00-9eeb-6a8928cc0bd0&clip=0&mode=live&start=61.488) [access the cluster manager UI for our](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=4c10c956-67e1-4a00-9eeb-6a8928cc0bd0&clip=0&mode=live&start=64.39427272727275) [local machine. We'll execute map reduce](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=4c10c956-67e1-4a00-9eeb-6a8928cc0bd0&clip=0&mode=live&start=67.04223076923078) [jobs in the cloud using the web console,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=4c10c956-67e1-4a00-9eeb-6a8928cc0bd0&clip=0&mode=live&start=69.01669230769234) [as well as the command line. We'll add](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=4c10c956-67e1-4a00-9eeb-6a8928cc0bd0&clip=0&mode=live&start=71.5694) [additional compute capacity to our cluster](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=4c10c956-67e1-4a00-9eeb-6a8928cc0bd0&clip=0&mode=live&start=74.232) [using preemptible VMs and monitor our](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=4c10c956-67e1-4a00-9eeb-6a8928cc0bd0&clip=0&mode=live&start=76.65599999999998) [cluster using Stackdriver. We'll then](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=4c10c956-67e1-4a00-9eeb-6a8928cc0bd0&clip=0&mode=live&start=80.01542857142857) [study how we can use the Spark distributed](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=4c10c956-67e1-4a00-9eeb-6a8928cc0bd0&clip=0&mode=live&start=82.63050000000001) [analytics engine on our Dataproc cluster.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=4c10c956-67e1-4a00-9eeb-6a8928cc0bd0&clip=0&mode=live&start=85.43650000000005) [We'll work with the PySpark shell on our](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=4c10c956-67e1-4a00-9eeb-6a8928cc0bd0&clip=0&mode=live&start=87.37199999999999) [cluster, as well as submit Spark jobs](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=4c10c956-67e1-4a00-9eeb-6a8928cc0bd0&clip=0&mode=live&start=89.84) [using the web console. We'll also see how](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=4c10c956-67e1-4a00-9eeb-6a8928cc0bd0&clip=0&mode=live&start=92.23909090909092) [we can write code to integrate our Spark](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=4c10c956-67e1-4a00-9eeb-6a8928cc0bd0&clip=0&mode=live&start=94.74799999999998) [jobs for BigQuery and cloud storage](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=4c10c956-67e1-4a00-9eeb-6a8928cc0bd0&clip=0&mode=live&start=97.02399999999993) [buckets using connectors. We'll then use](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=4c10c956-67e1-4a00-9eeb-6a8928cc0bd0&clip=0&mode=live&start=100.08899999999998) [our Dataproc cluster to perform extract,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=4c10c956-67e1-4a00-9eeb-6a8928cc0bd0&clip=0&mode=live&start=103.17299999999999) [transform, and load operations using Pig](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=4c10c956-67e1-4a00-9eeb-6a8928cc0bd0&clip=0&mode=live&start=105.59699999999997) [as a scripting language and work with Hive](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=4c10c956-67e1-4a00-9eeb-6a8928cc0bd0&clip=0&mode=live&start=108.24921428571432) [tables. At the end of this course, you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=4c10c956-67e1-4a00-9eeb-6a8928cc0bd0&clip=0&mode=live&start=111.84635714285722) [should be comfortable working with](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=4c10c956-67e1-4a00-9eeb-6a8928cc0bd0&clip=0&mode=live&start=113.96666666666663) [Google's managed Hadoop offering and have](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=4c10c956-67e1-4a00-9eeb-6a8928cc0bd0&clip=0&mode=live&start=115.16) [a sound idea of how to migrate jobs and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=4c10c956-67e1-4a00-9eeb-6a8928cc0bd0&clip=0&mode=live&start=117.23600000000002) [data on your on-premises Hadoop cluster to the Google Cloud.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=4c10c956-67e1-4a00-9eeb-6a8928cc0bd0&clip=0&mode=live&start=122.312)

[Introducing Google Dataproc for Big Data on the Cloud](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=0&mode=live)

[Module Overview](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=0&mode=live)

[Hi, and welcome to this course where we'll](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=0&mode=live&start=0) [study how we can architect big data](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=0&mode=live&start=3.672615384615384) [solutions using Google's Dataproc. We'll](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=0&mode=live&start=7.240666666666666) [start off by understanding what exactly](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=0&mode=live&start=10.107333333333333) [Dataproc is and how you can move your big](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=0&mode=live&start=12.035333333333334) [data computations to the cloud using GCP's](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=0&mode=live&start=15.365999999999996) [managed Hadoop offering. Now Hadoop is a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=0&mode=live&start=18.383) [ubiquitous big data processing technology](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=0&mode=live&start=21.159272727272732) [that has been around for over a decade and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=0&mode=live&start=23.01336363636365) [is very mature and is widely used. Hadoop](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=0&mode=live&start=26.80749999999999) [can be thought of as a software framework](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=0&mode=live&start=30.074846153846153) [that allows you to run computations on a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=0&mode=live&start=31.79361538461538) [number of different nodes where the nodes](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=0&mode=live&start=34.06100000000001) [belong to a cluster. Hadoop clusters](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=0&mode=live&start=36.525000000000034) [tightly couple storage, as well as](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=0&mode=live&start=39.076166666666666) [compute. You store data on different](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=0&mode=live&start=41.71516666666666) [machines in the cluster, and you run](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=0&mode=live&start=43.582083333333316) [processes on them in parallel. The](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=0&mode=live&start=45.58) [original concepts behind Hadoop were](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=0&mode=live&start=48.3412) [developed at Google in order to process](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=0&mode=live&start=50.55257142857143) [and index its web search results. This was](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=0&mode=live&start=53.14657142857145)[breakthrough technology at the time, and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=0&mode=live&start=55.8830909090909) [its open source version that is Hadoop is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=0&mode=live&start=57.42236363636362) [now widely available and used by almost](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=0&mode=live&start=60.125928571428574) [every organization. However, this tight](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=0&mode=live&start=62.83342857142858) [coupling of storage and compute leads to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=0&mode=live&start=64.66629999999999) [utilization and cluster sizing](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=0&mode=live&start=67.138) [inefficiencies. You end up paying for more](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=0&mode=live&start=70.11479999999997) [than what you need. Time has brought about](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=0&mode=live&start=72.83059999999996) [improvements in the products that are](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=0&mode=live&start=75.26161538461538) [available as well. Managed Hadoop allows](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=0&mode=live&start=76.93653846153846) [us to solve for these inefficiencies by](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=0&mode=live&start=78.9131) [splitting storage and compute. Where you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=0&mode=live&start=81.571) [store the data and where you perform](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=0&mode=live&start=83.77199999999999) [computations are completely separated.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=0&mode=live&start=86.58539999999999) [This allows us to efficiently use our](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=0&mode=live&start=89.035) [cluster. Dataproc is Google's managed](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=0&mode=live&start=92.36699999999999) [cloud Hadoop offering, and that's what](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=0&mode=live&start=94.40180000000002) [we're going to study here. Dataproc allows](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=0&mode=live&start=96.74) [you to set up clusters and perform](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=0&mode=live&start=98.85130769230771) [processing on huge amounts of data in a pay-as-you-go manner.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=0&mode=live&start=101.661)

[Prerequisites, Course Outline, and Spikey Sales Scenarios](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live)

[Before we move on to understanding and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=2.627) [working with Cloud Dataproc, let's see](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=4.54690909090909) [some of the prerequisites that you need to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=6.065571428571429) [have in order to make the most of your](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=7.751857142857146) [learning. Now this course assumes that](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=10.05916666666667) [you're familiar with cloud computing.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=11.816166666666675) [You've worked with cloud computing](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=13.623666666666663) [platforms before, either the GCP or AWS or](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=15.517) [Azure. If you haven't, then this course](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=19.917000000000005) [here, Choosing and Implementing Google](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=22.343199999999996) [Cloud Compute Engine Solutions, would be](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=23.93909090909091) [the right course for you to take before](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=26.00963636363637) [you move on to this course. This course](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=28.314400000000006) [also assumes that you're familiar with the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=30.162)[basics of big data processing in the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=31.775499999999994) [Hadoop ecosystem. If not, here is a course](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=34.54966666666668) [that you can take on Pluralsight, The](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=36.91300000000002) [Building Blocks of Hadoop -- HDFS,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=38.613) [MapReduce, and YARN. In order to be](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=40.881) [completely comfortable understanding the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=43.29180000000001) [concepts and working with the demos in](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=44.56860000000002) [this course, you need to have a basic](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=46.19969230769231) [understanding of cloud computing. It need](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=47.788) [not be on the GCP though. If you work with](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=49.51850000000002) [Azure or AWS, that's fine too. You need to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=52.66180000000001) [understand how VMs work. GCP uses its](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=55.98584615384615) [virtual machines in order to run managed](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=58.797153846153826) [Hadoop on the cloud. You need to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=61.404500000000006) [understand the basics of Hadoop, Spark,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=64.00318181818182) [Hive, Pig, and other ecosystem](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=66.39554545454544) [technologies. All of these technologies](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=68.4953) [run on top of the basic Hadoop framework.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=70.6658) [You don't need to have in-depth knowledge,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=73.216) [but you have to have some kind of](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=75.1805) [understanding of how they work. We start](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=77.44049999999999) [this course off by understanding what](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=79.4639090909091) [exactly is Google's Dataproc for managed](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=81.29663636363641) [Hadoop. We'll see why it's better than](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=83.21049999999998) [on-premises datacenters and why you might](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=85.538)[want to run your big data processing jobs](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=87.45700000000001) [on managed Hadoop. Once we've understood](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=90.23533333333333) [the concepts, in the module after that,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=92.8133333333333) [we'll focus on hands-on demos. We'll see](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=95.38990909090913) [how we can run MapReduce jobs on Dataproc.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=97.8194) [We'll see how we can monitor our jobs](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=100.381) [using Stackdriver monitoring and use](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=103.01633333333339) [initialization actions to set up our VMs.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=104.81925) [If your organization is working with big](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=108.216) [data, chances are your developers use](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=110.02009090909095) [Spark as a data analytics engine. We'll](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=111.86025) [see how we can work with Spark on](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=114.7093) [Dataproc. And, finally, we'll see how we](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=118.07970000000002) [can use Dataproc in order to perform](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=120.1426923076923) [extract, transform, and load operations](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=122.058) [using Pig and how we can use Hive on](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=124.02424999999998) [Dataproc as our data warehouse. Through](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=127.99614285714286) [all of the demos in this course, we'll](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=131.26992307692308) [assume that your part of the engineering](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=133.0773076923076) [team at SpikySales. com. Spikey Sales is a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=135.38242857142853) [hypothetical online retailer. They are an](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=139.13785714285714) [e-commerce site, which focuses on flash](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=141.48150000000004) [sales of trending products, which means](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=143.33850000000012) [they tend to have very large spikes in](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=145.251) [user traffic. At other times, traffic to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=147.87500000000003) [their website is fairly calm and quiet.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=150.3555833333333) [Their engineering team wants to move from](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=152.91) [an on-premises data center to the GCP.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=155.10927272727275) [They're exploring if GCP as cloud](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=157.62646153846157) [computing is a perfect fit for their use](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=160.14738461538468) [case. It's pay-as-you-go, and they have no](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=162.92366666666672) [idle capacity during the off-sale periods.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=164.84108333333347) [They want to move the analytics and other](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=167.20454545454547) [data that they have from their on-premises](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=170.31416666666667) [data center to cloud storage buckets.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=172.42175000000003) [Cloud storage allows them elastic storage,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=174.79399999999998) [pay as you go, as well as global access.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=177.732) [The engineering team currently has a large](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=180.583) [on-premises Hadoop cluster, and they have](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=183.40190000000007) [utilization concerns. As they get more](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=185.70200000000003) [popular, their Spikey Sales tend to have](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=187.88600000000005) [large utilization. If they buy additional](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=191.63324999999998) [machines to add capacity, those machines](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=193.95409090909087) [tend to be largely un-utilized when there](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=195.74372727272717) [are no sales. Administering this](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=199.30333333333326) [on-premises cluster is also proving to be](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=201.62949999999998) [a heavy burden. They want to scale the big](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=204.0637499999999) [data processing requirements that they](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=207.07000000000002) [have. The on-premises Hadoop cluster is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=208.70583333333337) [also expensive to scale, and they have](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=211.00633333333332) [many solutions which use Spark, Hive, and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=213.68545454545458) [Pig. So they'll use the entire suite of](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=216.2531818181819) [technologies available in the Hadoop](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=219.25576923076932) [ecosystem. They're especially excited](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=221.13846153846168) [about Cloud Dataproc on the GCP because](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=222.67690000000002) [they want a way to deal with spikes in](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=225.388) [workloads. The cluster is mostly idle for](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=228.74230000000009) [them when they use their on-premises](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=230.99761538461536) [solution. They want their cluster to be](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=232.8802307692307) [utilized completely. They also need to be](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=235.53174999999996) [able to quickly scale up their cluster processing on sale days.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=1&mode=live&start=237.7344615384615)

[Distributed Processing](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live)

[Let's first talk about what exactly Hadoop](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=3.036) [is and why it's useful for big data](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=4.770384615384616) [processing. Now back in the late 90s, most](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=7.054777777777778) [of the systems that we built were](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=10.169333333333332) [monolithic. There are two ways to build a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=12.425666666666663) [system, and in the 90s, they built large](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=14.660399999999994) [monolithic systems with components that](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=16.846) [call into each other. But as the 2000s](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=19.05655555555555) [rolled about and Google started web search](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=21.996999999999996) [at a massive scale, it started looking for](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=24.04799999999999) [distributed solutions. A distributed](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=28.03300000000001) [system basically involves a cluster of](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=29.970222222222223) [machines. There are many off-the-shelf](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=32.382888888888886) [components which make up this distributed](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=34.493727272727284) [cluster. A single machine in this cluster](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=37.056818181818215) [is called a node, and a group of machines](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=38.910714285714285) [working together to achieve a single](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=41.24916666666667) [objective is called a cluster. When you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=43.86616666666666)[have a number of machines working together](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=46.52483333333333) [in this manner, you require software that](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=48.19725000000001) [will coordinate processes across all of](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=50.70899999999999) [these machines. And this software was what](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=53.290181818181814) [Google developed back in the early 2000s.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=56.04181818181817) [Google's original software was made up of](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=58.68549999999999) [two basic components. It involved a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=61.286)[distributed file system called Google File](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=63.66080000000001) [System, and it involved a computation](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=66.04981818181818) [framework that ran processes on this](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=68.46727272727273) [distributed file system. This was called](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=70.964) [MapReduce. Google engineers wrote a paper](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=73.28836363636363) [on this breakthrough framework, and HDFS](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=75.36109090909089) [is the equivalent of the Google File](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=77.89163636363638) [System that was developed in the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=81.111) [open-source world. And MapReduce is the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=83.64825000000003) [open-source equivalent of the MapReduce](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=85.8812) [framework originally from Google. These](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=87.61520000000002) [technologies came together to form Hadoop,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=90.15057142857144) [HDFS, that is, a distributed file system.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=93.16390000000001) [This is what we use for storage of our big](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=96.365) [data. MapReduce is the compute, the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=99.0904545454545) [software that runs on top of the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=101.09716666666668) [distributed file system to process your](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=103.525) [data in parallel. And YARN is the cluster](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=106.33300000000003)[manager that coordinates the running of](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=109.79700000000003) [this software. YARN is responsible for](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=112.26900000000006) [resource allocation across the cluster for](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=114.641) [multiple jobs. It runs the MapReduce](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=116.96300000000001) [parallel programming framework so that](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=119.09409090909091) [data that is stored on HDFS, the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=120.79045454545455) [distributed file system, is operated on in](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=123.88938461538463) [parallel. This is how Hadoop achieves its](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=127.27792307692312) [massive big data processing capabilities.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=130.45218181818183) [A developer working on the Hadoop platform](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=132.694) [simply writes code in terms of map and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=134.93983333333327) [reduce operations, and these operations](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=137.84681818181815) [are then passed in to the cluster. YARN,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=140.01818181818172) [the cluster manager, is responsible for](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=143.03736363636366) [triggering this job on the cluster,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=145.097) [running these map and reduce operations in](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=147.3197272727273) [parallel across all of the nodes in the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=149.91309090909093) [cluster. YARN figures out where and how to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=152.87745454545464) [run the job and stores the result in HDFS.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=155.88393750000003) [Hadoop is the basic underlying framework](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=159.273) [over which all of these other technologies](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=161.16309090909093) [run, Hive, HBase, Pig, Flume and Sqoop,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=164.42572727272739) [Spark, and Oozie. If you've worked with](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=167.21076923076924) [any of these technologies, you're](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=169.6376153846154) [basically using Hadoop behind the scenes for your big data processing.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=2&mode=live&start=171.55466666666666)

[Storage in Traditional Hadoop](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live)

[Before we understand why exactly Google's](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=3.353) [Cloud Dataproc is such an improvement over](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=5.235363636363637) [the on-premises Hadoop that we worked on](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=7.561272727272728) [so far, we need to understand how storage](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=10.211727272727275) [and compute work in traditional Hadoop.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=13.141) [Here are the three components that make up](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=15.313) [Hadoop. HDFS, which stands for Hadoop](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=17.76544444444444) [distributed file system, is the storage](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=20.392833333333336) [component. MapReduce is the parallel](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=23.17783333333334) [programming framework. And YARN is for](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=24.977333333333338)[coordination. Let's talk about storage](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=26.979181818181818) [first and figure out where exactly data is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=28.64690909090909) [stored in traditional Hadoop. Now the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=31.880800000000008) [Hadoop distributed file system is a file](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=34.28333333333333) [system that spans multiple machines in](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=36.702999999999996) [your distributed cluster. The actual file](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=39.2805) [system resides on different nodes within](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=42.17100000000001) [your cluster, and all of these nodes are](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=44.87100000000002) [built on commodity hardware. These](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=47.850600000000014) [clusters are highly fault-tolerant. Nodes](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=49.819) [fail really often; data is often](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=52.494)[replicated on multiple nodes in order to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=54.57600000000001) [enable recovery. HDFS has been built for](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=57.29416666666666) [batch processing operations. The data](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=60.73524999999998) [access pattern tends to favor high](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=63.024461538461544) [throughput rather than low latency. As you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=65.46784615384615) [add more machines to your cluster, you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=67.98) [increase the size of data that you can](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=69.646) [store, which means HDFS is capable of](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=71.95699999999998) [supporting very, very large data sets.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=74.33524999999995) [Assume that you have a very large data](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=76.25181818181821) [set. It's made up of 100 million records.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=78.7946)[Now HDFS will manage the storage of this](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=81.397) [data across multiple disks. These disks](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=84.83033333333329) [will belong to different machines. Each of](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=87.38585714285713) [these disks will be associated with a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=90.20214285714286) [different node in your cluster. You have a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=92.71914285714286) [cluster of machines for your big data](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=95.38754545454546) [processing. These disks are associated](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=97.41881818181821) [with the different machines in that](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=99.16245454545452) [cluster. Now one machine here represents](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=101.3426363636363) [one node in your cluster, and a single](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=103.71600000000001) [node in your cluster is denoted as the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=106.932)[master node. The master node in your](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=110.33680000000003) [cluster is also often called the name node](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=112.72335714285717) [in your cluster, and the remaining nodes](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=115.17992857142863) [in your cluster are called data nodes,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=117.72349999999997) [also referred to as worker nodes. Now if](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=120.40642857142856) [you have your large data set comprised of](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=123.65638461538462) [100 million records, the name node does](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=125.94192307692315) [not actually store the data. It's](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=128.02433333333332) [responsible for managing the overall file](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=129.95212500000002) [system. It stores the directory structure](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=132.32887499999993) [and the metadata of all of the files that](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=134.43961538461542)[make up your data set. The name node is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=137.5000769230769) [responsible for knowing which part of your](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=140.3537692307691) [data is stored on which machine. Let's say](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=142.7258) [you've split your data into five. There](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=145.482) [are 20 million records on each of your](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=147.66718181818183) [file nodes. The name node will know which](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=150.14863636363648) [record is stored where. The data node is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=152.81020000000007)[where your data is actually stored. This](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=155.3154) [is comprised of the physical disks where](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=157.356) [the records, each of your 100 million](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=159.62400000000008) [records, will be stored. This is what a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=161.95088888888895) [traditional Hadoop cluster looks like. You](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=164.49222222222218) [have multiple data nodes within the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=166.26625) [cluster, and each of these data nodes have](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=167.92375000000004) [blocks of storage where your data will be](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=170.42075) [stored. And the name node will have a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=173.3960833333333) [mapping knowing where exactly a particular](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=176.23141666666675) [file is stored, in which block and which node in your cluster.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=3&mode=live&start=178.991)

[Compute in Traditional Hadoop](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live)

[Now that we know how data is stored in a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=2.234) [distributed manner, let's talk about how](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=4.608166666666667) [the MapReduce parallel programming](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=6.337285714285715) [framework allows us to operate on our data](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=7.718636363636363) [in parallel. MapReduce is just a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=11.24372727272727) [programming paradigm that allows us to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=13.31336363636364) [take advantage of the inherent parallelism](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=15.0959090909091) [that is present in data processing. Let's](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=17.548909090909095) [say you have a data set with 100 million](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=20.479) [records. Such data sets are actually the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=23.566) [norm nowadays, they're not the exception.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=26.186499999999995) [Now the MapReduce framework processes this](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=28.464) [data in two stages. The first of these](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=30.84660000000001) [stages is called the map operation, and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=33.6913076923077) [the second stage is called the reduce](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=35.88984615384618) [operation. The map operation runs in](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=38.91216666666667) [parallel across multiple machines on your](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=41.00736363636365) [distributed cluster, so the map process](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=42.947)[runs on multiple machines, and it works on](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=45.59660000000001) [that subset of data that is present on](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=48.83328571428572) [that machine. Once the map operation is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=51.77385714285717) [completed, the output of the map operation](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=54.16858333333332) [is passed on to the reduce operation. The](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=56.492) [reduce process then aggregates the output](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=59.198909090909105) [of the mappers together to give some](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=61.54)[meaningful output. Map and reduce are the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=64.48725) [only two functions that a programmer needs](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=67.23883333333336) [to define. And the rest of the detail,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=70.05285714285718) [making sure the data is processed on](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=72.57492307692308) [correctly, making sure the output of the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=74.571) [mapper is passed to the reducer, all of](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=76.68324999999999) [this is taken care of by Hadoop. The map](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=79.1739) [process looks at every record in your data](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=82.06736363636364) [set. Multiple map processes run on](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=84.47681818181817) [different nodes of your cluster, and they](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=86.65400000000002) [process only those records that are](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=89.2355)[present on that node. The map process](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=91.352) [looks at one record and produces one](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=93.99799999999999) [key-value pair as the output. The reduce](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=96.87966666666662) [process can be thought of as some kind of](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=99.42533333333334) [aggregation operation which takes the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=101.61833333333337) [output of several mappers together and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=103.65040000000002) [combines them together in some meaningful](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=106.333) [way. So the output of several mappers is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=109.1452857142857) [passed to the reduce phase, and then you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=111.88724999999997) [get the final output as the result. So if](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=114.37229999999998) [you want to summarize map and reduce, map](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=116.8528) [is a step that is performed in parallel on](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=118.73599999999999) [multiple machines where your data is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=121.53990909090906) [stored, and reduce is a step which](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=123.98027272727266) [combines the output of the mapper that is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=126.20028571428571) [the intermediate results to give you the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=128.618) [final result. As an example, let's assume](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=131.12866666666665) [that you have a very large file in](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=132.901)[petabytes, and this has been stored across](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=135.7003846153845) [multiple machines on your cluster. It's](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=138.2297142857143) [managed by HDFS. How do you go from this](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=140.49757142857143) [very large file to counting the word](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=143.3647142857143) [frequencies in this file? You want to know](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=145.5047142857143) [how many times each word occurs in this](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=147.99192307692306) [file. This file that we start off with,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=150.6577692307692) [which makes up our data set, is typically](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=152.42039999999992) [very, very large in petabytes. Now this](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=155.11500000000004) [file will be split up, and different](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=157.7425714285714) [subsets of the file will be stored on](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=159.65207142857133)[different machines in your cluster. So you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=162.022125) [have all of these machines which hold](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=164.33069230769235) [subsets of your data. Each partition or](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=166.5416153846155) [each subset of your data is given to a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=168.66761538461543) [different mapper process, which operates](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=171.2554615384617) [on that data. So you can imagine a mapper](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=173.1527142857143) [process is operating on each individual](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=176.98200000000008) [subset here. The mappers work in parallel.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=179.60400000000016) [They perform the same operation but on](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=182.13883333333334) [that subset of the data that is present on](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=184.464) [that machine. So within each mapper, the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=187.36445454545446) [rows in your files are processed one after](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=189.90866666666662) [the other. But all the mappers work in](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=192.94813333333323) [parallel. Each mapper will process just](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=195.74650000000005) [one line at a time. It'll split the line](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=197.785) [into words, and output the word itself and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=200.89) [a count of 1. So for every word, a count](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=203.5046875) [of 1 is associated with that word. So each](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=206.47031249999992) [row or line will be split into as many](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=209.4848571428571) [key-value pairs as there are words in that](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=212.6282142857143) [line, and that is the output of every](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=215.10992857142867) [mapper. The output of every mapper is then](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=218.45950000000002) [passed on to the reducer. The reducer is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=221.89925000000008) [where the actual counting happens. Behind](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=225.50487499999997) [the scenes, Hadoop will group all of the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=228.39023076923078)[key outputs of the mapper along with the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=230.44807692307697) [associated values, and the reducer will](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=233.01881818181815) [simply sum up the values associated with](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=235.33154545454536) [each key. And it'll get the final word](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=238.43950000000007) [frequencies. The Hadoop framework is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=242.04528571428577) [responsible for ensuring that all of the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=243.8360909090909) [values associated with the same key are](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=246.183) [passed on to one reducer. This combination](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=249.75183333333342) [of mappers and reducers together form the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=252.9706) [basic building blocks of big data processing.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=4&mode=live&start=255.3142)

[Separating Storage and Compute with Dataproc](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live)

[We'll now move from the world of](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=1.487) [traditional Hadoop to Cloud Dataproc and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=4.1738333333333335) [see why Cloud Dataproc is so much more](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=6.519333333333333) [powerful and flexible. The first](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=9.930000000000001) [significant advantage that you'll](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=11.9025) [encounter when you move your big data](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=13.164363636363635) [processing systems to the cloud is the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=15.153636363636359) [fact that you can create and take down](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=17.13266666666667) [your Hadoop clusters on the fly. If you're](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=19.389733333333343) [working with the traditional on-premises](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=21.867399999999996) [data center, you're probably already aware](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=23.6275) [that provisioning Hadoop machines and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=26.534500000000005) [setting up your cluster is an onerous,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=28.181000000000004) [time-consuming process. This is no longer](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=30.3383) [true when you're using managed Hadoop on](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=32.8241) [the cloud. This is the most obvious reason](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=35.72766666666666) [to choose cloud computing platforms. But](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=38.626363636363656) [before that, let's talk about the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=40.705400000000004)[drawbacks of tight coupling between](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=42.83780000000002) [storage and compute that you have to live](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=44.511166666666675) [with if you're using traditional Hadoop](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=46.93983333333336) [clusters. Now we know that each node](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=49.17259999999999) [contains data blocks, and the node also](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=51.609769230769224) [runs processes, mappers and reducers that](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=53.979) [work on these data blocks. Storage and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=56.58681818181819)[compute are tightly coupled. They run on](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=59.50299999999999) [the same node. The first problem that you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=62.103499999999976) [have with this setup is that you have to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=64.35692857142858) [keep your cluster up and running even when](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=66.407) [there is no processing happening. The](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=69.38976923076923) [cluster is required to store your data.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=71.665625) [The cluster has to be provisioned. Your](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=74.547)[machines have to be up and running in](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=76.44292307692305) [order for data to be persistent. Now if](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=78.87275000000001) [you encounter a situation where you have](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=81.7537142857143) [to use this cluster to store even more](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=83.38721428571431) [data, the size of your data has expanded](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=85.40245454545456) [significantly, then you need to resize](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=87.665) [this cluster by adding more machines. And](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=90.37399999999998) [resizing the cluster requires all of your](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=93.41458333333333) [data to be re-sharded and re-distributed](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=95.73566666666662) [across the nodes in your cluster. For a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=97.96275) [rapidly growing organization, this is a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=101.39009090909092)[major pain point. Another issue is that of](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=103.56536363636364) [utilization. For an organization such as](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=106.10390909090911) [Spikey Sales where the utilization of](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=107.845) [their cluster varies significantly from](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=110.49966666666668) [day to day, sizing the cluster, figuring](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=112.61372727272726) [out the right number of machines in the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=115.36727272727268) [cluster, and figuring out the best](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=117.87050000000002) [possible utilization of these machines](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=119.587) [become very difficult. If you try to over](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=121.828875) [provision your cluster in order to meet](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=124.63007142857145) [the demands of traffic on sale days, then](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=126.66857142857148) [the cluster becomes too big. The](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=129.18245454545453) [utilization of your hardware is far too](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=131.11866666666666) [low. At the other end of the spectrum, if](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=133.40533333333323) [you try to increase utilization and cut](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=136.17569230769234) [down on the fixed cost of your machines,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=138.19313333333335) [you end up with a cluster that is far too](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=140.15366666666674) [small. Then the latency and the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=142.91233333333335) [performance of that cluster might be](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=144.75233333333338) [unacceptable. And for many organizations,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=146.7195555555556) [this happens to be the major driving force](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=148.4522) [behind moving their big data processing to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=151.68580000000003)[the cloud. Cloud-based Hadoop offerings](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=154.49299999999997) [for managed Hadoop allow you to control](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=156.67370000000003) [the size of your clusters on the fly. You](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=159.893) [can dynamically create and use the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=162.71692307692302) [clusters whenever you require them. You](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=164.85514285714285) [can scale the clusters very, very fast.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=167.35299999999998) [You can add more virtual machines to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=169.383) [increase the size of your clusters on sale](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=171.34416666666672) [days. On other days where your](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=173.90518181818186) [organization might experience low traffic,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=176.22554545454557) [you can simply decommission those bits of](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=178.32690909090906)[hardware. You can simply turn down those](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=181.42536363636353) [machines. When you use GCP's managed](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=184.35899999999998) [Hadoop offering, the administrative](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=186.85) [overhead of performing these operations is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=188.67749999999998) [completely abstracted away from you. All](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=191.3812857142857) [you have to do is run scripts on a command](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=193.869) [line or click on the web console. Here is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=196.83899999999997)[a visual representation of the tight](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=199.59545454545457) [coupling between storage and compute in a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=201.62781818181827) [traditional Hadoop cluster. The data nodes](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=204.3068571428571) [that perform processing also store the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=206.85150000000002) [data. In the case of a Dataproc cluster,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=209.44650000000004) [though, you'll have VM instances, which](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=212.1119999999999) [are responsible for running processes.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=214.16)[You'll store your data in cloud storage](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=216.425) [buckets. And this is a significant reason](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=219.4411250000001) [why Cloud Dataproc is so powerful. You can](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=221.95050000000003) [store your data on cloud storage buckets](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=224.58176923076923) [where it can be accessed by other GCP](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=226.51646153846153) [services. You'll perform all of your](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=229.18309090909094) [processing on your Hadoop cluster. Low](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=231.52472727272732) [latency, high throughput connections](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=234.655) [between cloud storage and your cluster](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=236.847) [machines make this the preferred way to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=239.03640000000004) [operate. The nodes that make up your](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=241.6235) [Hadoop cluster on the cloud are GCE VM](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=244.25433333333334) [instances, Google Compute Engine virtual](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=248.28107692307694) [machine instances. The master node, as](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=251.01030769230772) [well as all of the worker nodes which run](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=253.174) [your big data processing operations are](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=255.586) [GCE VM instances, and all compute is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=258.41127272727283) [performed on these instances. The map and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=261.85045454545457) [reduce operations that we spoke of earlier](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=264.4436363636364) [are processes on these VMs. You can](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=267.3390000000001) [configure one or more VM instances as the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=270.2968333333334) [master node. On the GCP, it's a simple](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=273.20816666666684) [configuration setting to run your cluster](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=275.6976923076924) [in the high-availability mode with](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=277.45938461538486) [multiple masters. So if a master fails,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=280.04299999999995) [another one can take over easily. All of](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=282.8848181818183) [the data on which you want to run big data](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=285.4783333333333) [processing won't be on the actual nodes](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=288.2649999999998) [themselves. Instead, you'll use cloud](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=290.401) [storage buckets for storage. Cloud storage](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=293.07099999999997) [buckets on the GCP allow object storage](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=295.8759090909091) [just like AWS's S3 or Azure's blob](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=298.2170909090909) [storage. This is elastic storage. The](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=301.5915000000001) [storage expands as you add more data, and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=304.6169375) [you only pay for what you use. You don't](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=307.17643749999985) [provision storage upfront and pay fixed](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=310.23133333333334)[costs. You only pay for the data that you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=312.67733333333337) [actually store in a bucket. This is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=315.26715384615403) [infinitely scalable. You just add more](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=317.27099999999996) [data as you need to. You don't need to add](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=319.1909999999999) [additional machines. And more](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=322.606) [significantly, storage is completely](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=324.058) [decoupled from compute. It's scalable](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=325.798) [independent of your compute option, so you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=327.8376666666667) [can have a very small cluster and a huge](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=330.1569999999998) [data set, or you can have a large cluster](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=332.9548571428572) [and a small data set. Here are the basic](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=335.7538333333333)[components that make up a traditional](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=339.1036666666667) [Hadoop framework. Here we've switched](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=341.1746666666667) [storage over to Google Cloud Storage. Now](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=343.0636666666666) [this is preferable. It's not that you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=345.77507692307694) [can't use HDFS when you're using Cloud](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=347.9736153846154) [Dataproc. However, you would prefer to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=350.2785) [store your data on Google Cloud Storage](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=352.4820000000001) [thus effectively separating compute from](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=355.6986666666666) [storage. Instead of provisioning real](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=358.61144444444426) [machines with software installed to run](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=360.29839999999996) [MapReduce, you'll use Google Compute](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=361.955) [Engine VM instances. And instead of YARN,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=364.58125000000007) [you'll use something called a Dataproc](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=367.80863636363625) [service. YARN is just one component within](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=370.28609090909066) [Dataproc service. You can think of the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=372.75475) [Dataproc service, what you use for](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=375.1187500000001) [coordination within your managed Hadoop cluster, as YARN++.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=5&mode=live&start=377.1332500000002)

[Hadoop vs. Dataproc](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live)

[If your organization is already using an](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=2.374) [on-premises Hadoop cluster, then the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=5.141800000000001) [chances are you're storing your data on](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=7.0548) [HDFS. If you're using Cloud Dataproc,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=9.598600000000003) [though, the preferred solution is to use](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=11.608363636363634) [Google Cloud Storage to store your data.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=14.048874999999999) [Now HDFS is still available on all of your](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=17.303) [cluster nodes. It still runs on the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=20.65018181818183) [persistent disks of your cluster VMs. We](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=22.86127272727272) [should avoid using it, though, so that we](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=24.88775) [don't tightly couple storage and compute.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=27.013750000000005) [This basically means that when here using](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=29.86) [Dataproc, you don't have to instantiate](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=31.712666666666667) [your cluster machines if you're not](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=33.48763636363636) [actually running any processing. You don't](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=36.19745454545453) [have to instantiate those clusters only to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=38.63963636363636) [store data. Store data on cloud storage](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=40.86436363636364) [buckets. This is what Google prefers, and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=43.73727272727274) [this is what is recommended in order to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=46.761583333333334) [reduce your costs. You should not pay for](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=49.47825000000003) [VM instances if they don't need to be](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=51.9845) [active. Let's take a look at some of the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=54.36449999999999)[differences between traditional](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=56.6590909090909) [on-premises Hadoop versus Cloud Dataproc.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=57.408) [In the traditional Hadoop setup, clusters](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=60.86) [stay in existence forever. You add](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=63.084800000000016) [machines, you install the software, you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=65.30018181818183) [provision your cluster. In the case of](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=67.49672727272731) [managed Hadoop with Dataproc, you can](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=69.68533333333332) [create your clusters dynamically when you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=71.3208333333333) [need them. Google's managed Hadoop](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=74.17977777777777) [offering does not do away with HDFS. HDFS](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=77.072) [runs on the persistent disks of VMs, and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=80.1053333333333) [you can use it if you want to. You'll just](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=82.67166666666667) [have to make sure that your cluster is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=85.22084615384615) [instantiated if you want persistent](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=87.02023076923076) [storage. When you're working with](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=88.64122222222221) [traditional Hadoop on-premises, the data](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=91.00677777777776) [for jobs that you run will be stored in](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=93.15253846153847) [HDFS. In the case of Dataproc, though,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=96.75115384615387) [ideally you want your data stored in](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=99.04099999999998) [Google Cloud Storage buckets, which offers](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=101.246) [elastic storage and fast access. When you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=104.4167272727273) [store data in your cluster, your cluster](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=107.93118181818181)[encapsulates state. It holds state. The](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=110.98381818181817) [cluster with Dataproc is stateless. It's](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=113.66755555555557) [only used for processing. In addition to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=116.376125) [loosely coupling storage and compute, when](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=119.485) [you store data in cloud storage buckets,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=121.38645454545455) [there are other advantages to using bucket](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=123.69300000000001) [storage versus persistent disk storage.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=126.22000000000004)[Persistent disk is the storage that you're](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=128.2894) [familiar with. It's block storage of data.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=131.214875) [Buckets offer object storage. The data](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=133.755) [itself is treated differently. When your](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=136.64820000000003) [data is in a bucket, you can configure](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=139.17583333333334) [fine-grain policies for version](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=141.57116666666673) [management, as well as lifecycle](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=143.05581818181818) [management for this data. A persistent](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=145.27036363636367) [disk associated with a Compute Engine VM](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=147.77553846153845) [on the GCP can be a maximum of 64 TB in](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=150.342923076923) [size. Buckets, though, are infinitely](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=154.60623076923076) [scalable. You can add as much data as you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=156.60353846153845) [want in a bucket. Persistent disks are pay](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=159.48233333333343) [what you allocate, so if you have a 3 TB](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=161.87171428571426) [disk, you pay for the entire 3 TB no](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=164.746)[matter if you use just 100 GB from within](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=167.5421875) [it. Buckets are pay what you use. You only](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=170.393) [pay for the data that you've stored.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=173.40677777777776) [Persistent disks are closely tied to GCE](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=175.674) [VMs and can be only accessed by processes](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=178.01253846153855) [on those VMs. Buckets are completely](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=180.83263636363637) [independent of GCE VMs, and they can be](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=183.27954545454548)[accessed by other Google services as well.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=186.54924999999992) [Persistent disks can be accessed from](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=189.414) [within a particular zone. A zone can be](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=191.67899999999997) [thought of as a single data center or a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=194.1095333333333) [failure domain within GCE or can be](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=196.18073333333325) [accessed from within a region in certain](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=198.66984615384618) [cases. In the case of buckets, it allows](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=201.51023076923084) [regional, as well as global access. A](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=204.49146153846164) [region on the GCP can be thought of as a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=207.01153846153846) [geographical area, and global basically means across the world.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=6&mode=live&start=210.13692307692298)

[Using the Cloud Shell, Enabling the Dataproc API](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live)

[Let's take a little break from concepts](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=2.862) [and log on to our GCP account and enable](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=4.779) [the APIs that we require to work with](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=7.4852727272727275) [Dataproc. You can create an account on the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=10.642363636363639) [Google Cloud Platform using a Gmail](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=12.861846153846155) [account or a G Suite account that is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=14.426) [associated with your organization. If](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=17.321333333333328) [you're logging in for the very first time,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=19.011307692307692) [you probably don't have a project yet. You](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=20.949769230769245) [can click on the drop-down on top and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=22.911000000000005) [create a new project or choose an existing](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=24.895000000000014) [project. A project on the GCP can be](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=27.598230769230764) [thought of as a logical unit that holds](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=30.733) [your resources. Every resource on the GCP](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=33.3458) [belongs to a particular project. A project](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=35.93200000000001) [is an independent billing unit and is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=38.62809090909091) [typically associated with a particular](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=41.29890909090908) [team in your organization. Spikey-dataproc](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=43.18742857142856) [is the project that the Spikey Sales](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=45.146) [organization engineers are using in order](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=48.35199999999999) [to test out Dataproc. This should take you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=51.12166666666667) [to the dashboard for the spikey-dataproc](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=54.35681818181819) [project. Make sure that you have billing](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=56.43554545454548) [enabled for this particular project so](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=58.36499999999998) [that you can instantiate VMs that make up](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=59.948) [your Dataproc cluster. So long as you're](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=63.48618181818184) [good about deleting the resources that you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=65.92166666666665) [create, the charges that you'll incur when](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=67.86183333333327) [you perform the demos in this course](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=69.95800000000001) [should be under $5. If you're working on](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=72.25118181818182) [the GCP, the cloud shell is a very handy](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=76.04463636363637) [feature. The cloud shell is basically a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=79.31174999999998) [terminal window on an ephemeral VM that](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=81.56530769230767) [Google creates for you on the cloud. This](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=83.808) [VM comes preinstalled with all of the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=86.87599999999999) [utilities that you need to work with](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=89.29246153846154) [Google services such as the gcloud SDK,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=91.35153846153847) [gsutil to work with cloud storage buckets,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=93.72090909090907) [and any other tools that you need. You can](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=96.323) [explore this dashboard and other options](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=98.73292307692309) [that are available here. On the top right,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=100.26400000000001) [you can see that I've signed in with a G](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=102.10400000000004) [Suite account, Spikey Sales. Use the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=105.76709090909087)[hamburger icon on the top left in order to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=107.97133333333333) [access all of the products and services](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=110.24533333333336) [that GCP has to offer. If you click on](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=112.24300000000002) [this, it'll open up a navigation menu, and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=114.88000000000002) [you can use this to navigate to the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=116.86400000000006) [service that you're interested in. We're](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=119.36400000000003) [going to start off with the APIs and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=121.4053076923077) [Services where we'll enable the APIs that](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=123.68776923076928) [we need to work with Dataproc. We've come](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=126.03233333333331) [to the API library here. This gives us a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=128.96976923076923) [search box, which allows us to search for](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=131.3582307692308) [the API that we're interested in. We are](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=133.4530000000001) [interested in the Dataproc API. This gives](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=135.74053846153845) [us one search result. Click through here,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=138.27292307692304) [and click on the Enable button. Once this](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=140.88833333333335) [API is enabled, you can set up Dataproc](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=144.00300000000001) [clusters on the GCP. You can click on the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=147.1150000000001) [Back button here at the top in order to go](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=149.57412499999995) [back to the main page. Here you'll see](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=151.38037499999987) [links to manage this API and another](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=154.01692307692315) [button that will take you to a tutorial](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=156.4558125) [that allows you to try this API. And at](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=159.12631249999995) [any point in time, clicking on the Google](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=161.95223076923074) [Cloud Platform link at the very top will](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=163.9091538461537) [take you to the main dashboard for your](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=166.2194285714286) [project. Activate your cloud shell by](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=168.74171428571438) [clicking on the cloud shell icon on the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=170.59299999999996) [top right. Here you have your cloud shell](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=173.05059999999992) [terminal window. You can see that the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=175.99825000000007)[current project is spikey--dataproc. You](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=178.72044444444444) [can ensure that the current session of](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=180.48992307692308) [your cloud shell is always set to this](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=181.98738461538454) [project by calling gcloud config set](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=184.68627272727275) [project spikey-dataproc. The gcloud is a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=187.94481818181828) [command line utility, which is the SDK](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=190.2254615384615) [that Google offers to work with GCP](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=192.31576923076912) [services. Gcloud comes preinstalled on the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=195.16838461538464) [cloud shell. If you're working on a local](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=197.8748461538462) [machine, you need to download the gcloud](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=200.88171428571422) [SDK. Google also offers a nifty code](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=203.33721428571417) [editor on your browser. You can click on](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=206.24399999999991) [the edit icon there and launch code editor](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=209.63569230769227) [beta. The three-dot menu will show you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=213.01107692307679) [additional actions that you can perform on](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=214.861) [your cloud shell such as upload a file to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=216.58084615384615) [cloud shell, download a file from that VM,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=218.9651538461539) [and so on. You can use the cross button at](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=221.20620000000002) [the very right in order to close this terminal window.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=7&mode=live&start=224.773)

[Dataproc Features](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live)

[So far, we've mainly spoken of the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=2.121) [decoupling of storage and compute when we](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=4.087416666666667) [use Dataproc as opposed to traditional](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=6.211555555555555) [Hadoop. Let's see what other features](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=8.370222222222221) [Dataproc has to offer. Now there are](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=10.645555555555555) [different types of clusters that you can](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=12.942800000000002) [create when you use Dataproc. You can](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=14.727333333333338) [create a regular single-master cluster,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=16.710545454545457) [you can have a single-node cluster as well](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=18.481909090909102) [with the master and the worker in the same](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=20.929000000000002) [node. This is useful when you're](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=23.246500000000005) [prototyping and developing your big data](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=25.364) [processes. Or you can set up your cluster](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=27.596) [in high-availability mode with more than](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=30.25233333333334) [one master. Once you've created your](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=32.30775) [cluster, accessing your cluster in order](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=34.85024999999999) [to run jobs on it or to administer it is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=36.81042857142856) [very straightforward. You have a variety](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=39.84369230769231) [of ways in which you can connect to this](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=41.44984615384615) [cluster. You can use the web console, or](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=43.838538461538455) [you can script using the gcloud command](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=45.925307692307676) [line utility. You can also use REST APIs](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=48.90414285714286) [in order to administer your cluster, or](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=52.04799999999999) [you can use client libraries that give you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=54.656363636363636) [programmatic access. The advantage of](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=58.2352727272727) [using a cloud platform such as Google is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=59.82300000000001) [that cluster actions such as resizing are](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=62.33266666666668) [very, very fast. You can reset your](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=64.876) [cluster even when you have processes](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=67.384)[running. You don't need to pause or stop](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=69.53371428571428) [your jobs just because you need to add or](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=71.5104) [remove a few nodes. You can increase and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=73.63107692307692) [decrease the number of workers that you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=75.79969230769227) [have enabled on the fly. GCP allows you to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=77.92200000000001) [gracefully decommission the downsized](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=81.036) [cluster nodes. If you plan to decommission](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=83.57933333333335) [a node, processes that are running on that](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=86.1252307692308) [node will first be completed. Only then](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=88.164) [will that node be removed. You might want](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=91.08008333333335) [to scale your cluster for a variety of](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=93.85115384615385) [different reasons. For example, the Spikey](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=95.68623076923079) [Sales organization has found that on sale](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=97.32536363636366) [days, they need analytics reports quickly.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=99.391) [This allows management to analyze how the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=102.767) [sales are doing and to tweak their product](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=104.83469230769234) [offerings. Now the batch processes that](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=107.2780909090909)[are run to generate these reports can be](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=109.29136363636363) [run much faster with additional compute](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=112.68091666666669) [nodes. You might also want to scale if](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=115.54041666666672) [you're storing your data on HDFS and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=117.773) [require additional storage. Because](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=119.529) [clusters can be tweaked so easily, large](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=122.9374) [clusters are completely fine when working](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=124.8582) [with Dataproc. Large clusters can be](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=127.38099999999999) [scaled up or down based on your](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=129.48999999999995) [requirements. In fact, there is brand-new](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=131.45349999999988) [functionality available from Google to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=133.707)[enable you to auto-scale your clusters.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=135.68) [Based on YARN cluster metrics, nodes will](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=138.22400000000005) [automatically be added to or removed from](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=140.96481818181817) [your cluster within bounds that you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=143.40272727272728) [specify. You can also configure your](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=146.67683333333338) [cluster very explicitly by specifying what](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=149.09881818181825) [kind of OS images you want to run on your](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=151.303) [cluster VMs. You can switch between](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=154.07383333333337) [different versions of the various](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=156.098) [technologies, a different version for](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=157.93550000000005) [Spark, Hadoop, Hive, Pig, and so on.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=160.13463636363633) [Google is responsible for the version](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=163.34) [management of these technologies and will](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=165.5024) [upgrade them separately. All of this is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=167.74327272727274) [managed completely by the GCP. If you want](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=170.54072727272737) [your cluster VMs configured with](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=172.9694285714286) [additional software, you can specify them](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=174.67459999999997) [in the form of initialization actions. You](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=176.825) [can specify separate initialization](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=179.48780000000005) [actions for the master node, as well as](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=181.02718181818182) [worker nodes. We've already discussed the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=184.21263636363642) [auto-scaling of clusters that is currently](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=186.5086666666667) [in alpha. Another cool feature is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=188.75816666666677) [auto-zone placement. If you want your](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=191.25066666666666) [cluster nodes to be within a particular](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=193.34266666666662) [region, you can specify auto-zone](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=195.47883333333317) [placement, and GCP will choose the right](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=197.3203636363636) [zone for you for the different nodes.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=200.009) [Scheduled deletion is a feature that you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=202.389) [can enable on your cluster nodes in order](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=204.55971428571428) [to ensure that you don't inadvertently](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=206.4425714285714) [incur costs for a cluster that you no](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=208.29216666666665) [longer use. Jobs running on your cluster](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=211.92483333333328) [can also be configured to be restartable.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=215.1009166666667) [Scheduled deletion of cluster nodes is a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=218.276) [very cool feature that allows you to get a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=220.31066666666658) [tight grip on utilization and costs. You](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=223.10800000000006) [can delete the cluster after a specified](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=225.4992) [idle period, at a specified future time,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=228.12560000000005) [or after a specified interval after you've created the cluster.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=8&mode=live&start=230.70599999999996)

[Migrating to Dataproc](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live)

[Dataproc's advantages have convinced your](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=2.724) [organization that moving to the cloud is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=4.897333333333334) [the right thing to do. Dataproc is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=7.496307692307693) [basically Hadoop on the cloud, so you'll](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=9.503692307692308) [find that migrating to Dataproc is very](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=11.77266666666667) [easy, but using it well in order to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=14.23783333333334) [extract all of the efficiencies that it](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=16.528999999999996) [offers requires a big change in mindset,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=18.468) [and that's where you might trip up. Let's](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=21.11950000000001) [take a look at some of the design](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=23.951) [decisions that you might need to make so](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=25.391) [that when you move to the cloud, you're](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=27.026799999999998) [set up to utilize all of the benefits.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=28.988933333333325) [Dataproc supports almost all of the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=31.263090909090906) [technologies in the Hadoop ecosystem, so](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=33.20272727272728)[migrating should be relatively](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=35.46149999999999) [straightforward. There is built-in support](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=37.00989999999997) [for MapReduce, Spark applications written](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=39.050333333333334) [either in Scala or in Python, Pig for](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=41.117) [extract/transform/load operations, and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=45.07699999999998) [Hive to use as your data warehouse. Here](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=46.38911111111112) [are some of the mindset changes that you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=49.454928571428574) [might need to make when you're migrating](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=51.10235714285715) [to Cloud Dataproc. Now when you're on the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=52.843272727272726) [cloud, clusters are no big deal. You can](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=55.68836363636364) [create clusters on the fly. You can make](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=58.172583333333314) [the clusters as large as you want them to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=60.44415384615385) [be. When you're working with on-premises](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=62.69761538461542) [Hadoop, clusters are really big deal. You](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=64.42027272727272) [have to specially order machines,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=66.59984615384616) [provision them for Hadoop, and so on. When](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=68.66407692307696) [you move to the cloud, you need to start](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=71.78471428571429) [thinking of clusters as something that is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=73.86114285714285) [a lightweight compute option. You create](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=75.80914285714287) [and tear down clusters on the fly. When](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=77.94733333333333) [you have the on-premises mindset, though,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=80.96254545454546) [your cluster size and setup is set in](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=82.85581818181821) [stone. When you're working on the cloud,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=85.54927272727275) [don't worry about the size of the cluster](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=88.054) [and the number of nodes that you need to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=89.82371428571426) [provision when you first create them. You](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=92.16433333333332) [can resize and tweak your cluster once](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=94.48133333333334) [you've seen the traffic patterns and your](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=96.5136666666667) [processing requirements. With the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=99.4048) [on-premises mindset, cluster properties](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=101.3182) [are typically set in stone and cannot be](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=102.8386) [tweaked. When you're using Dataproc on the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=105.90566666666668) [cloud, you need to trust GCP. You need to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=108.61233333333337) [trust that GCP will handle all of the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=111.44806666666668) [orchestration, administration, and setting](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=113.9694) [up and scaling of the cluster for you. If](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=115.48299999999999) [you have the on-premises mindset, you'll](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=118.50299999999999) [try to tightly configure and control](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=120.19499999999996) [everything simply adding to your burden.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=122.903) [When you're working with the cloud](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=126.119) [platform, you won't keep your cluster](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=127.45863636363637) [running at all times and incur unnecessary](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=129.0212222222222) [costs. You'll bring your cluster up only](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=132.14477777777773) [when you want to explicitly run some kind](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=134.24146153846146) [of job on the cluster. If you have the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=136.48633333333333) [on-premises mindset, you'll end up keeping](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=139.36033333333333) [your cluster up at all times and paying](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=141.05500000000004) [for the VM instances that you have](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=143.16700000000012) [running. Remember that clusters on the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=145.55363636363634) [cloud can be brought up very easily. If](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=147.784) [you require multiple clusters, and you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=150.59636363636366) [have jobs that need to be isolated, go](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=152.61746153846153) [ahead and create multiple clusters. If you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=155.19469230769224) [have the on-premises mindset, you'll](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=157.97233333333338)[always use the single cluster that has](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=160.00764285714286) [been set up and provisioned right. This](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=162.30814285714283) [makes it much harder for you to control](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=164.51684615384616) [the resources that are allocated to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=166.41161538461543) [specific jobs on that cluster.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=168.333) [Administering of clusters in the cloud is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=170.768) [very simple using the web console. You can](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=173.2524615384616) [submit jobs using this console, monitor](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=176.022) [them, and even cancel them if you want to.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=177.94199999999995) [If you have the on-premises mindset,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=180.77290909090902) [you'll continue to use the on-premises](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=182.634)[procedures such as logging into the master](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=184.61345454545457) [machine to submit and manage jobs. Based](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=187.01975000000002) [on the kind of workloads you want to run,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=189.51366666666667) [you can configure your cluster exactly the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=191.8276923076923) [way you want it to be. You can configure](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=193.92015384615382) [the machine types that make up the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=196.61100000000005) [individual nodes. If you're running a job](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=198.7040000000001) [that requires high compute capacity, you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=201.002) [can configure your cluster with high CPU](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=203.0851111111111) [machines. If your job has large memory](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=205.76688888888884) [requirements, your cluster can be made up](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=207.87430769230767) [of high memory machines. Do not try to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=209.886) [match your cluster with what you currently](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=212.38753846153853) [have on your on-premises data center.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=214.92600000000002) [Tweak your cluster based on your](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=217.851) [requirements. When you're working on the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=220.56900000000002) [cloud, don't be afraid to provision very](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=222.11783333333332) [large clusters to have your job run](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=223.652) [quickly. You can shut down the cluster](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=226.26299999999992) [once your job has completed. If you're](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=228.61000000000004) [working with the on-premises mindset,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=230.76463636363638) [you'll try to match your cluster size on](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=232.03872727272736) [the cloud to what you have on-premises.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=234.29909090909086) [And that doesn't really work that well.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=236.471) [Think of the number of nodes you have in](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=238.34861538461539) [your cluster kind of like a configuration](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=240.89624999999998) [parameter that you have to tweak on the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=243.08549999999994) [fly. Scale your clusters up and down as](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=245.83820000000003) [needed. If you continue with your](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=248.67446666666672) [on-premises mindset, you'll try to keep](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=250.1765) [the cluster size constant at all times to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=251.5535) [avoid provisioning new machines. And here](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=254.9953636363636) [is a really cool feature that you can use](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=257.37842857142857) [on the GCP. You can use preemptible VMs in](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=259.84185714285707) [order to add compute capacity to your](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=263.1452857142856) [cluster. Preemptible VMs are kind of like](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=266.04128571428544) [Amazon's spot instances, but you don't](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=268.8403333333335) [require reservations upfront. Preemptible](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=271.496) [VMs should be used as additional compute](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=273.611) [capacity when you really need it. This is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=275.95800000000014) [an instance that you can create and run at](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=278.15086666666673) [a much lower cost than normal instances.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=280.3912666666669) [This is because the GCE might terminate](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=282.6153636363636) [these instances at any point in time, and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=285.48154545454537) [your preemptible instance will be](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=287.75300000000016) [terminated at least once in a 24-hour](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=289.083) [period. Preemptible instances are a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=293.2287499999999) [low-cost way to get additional compute](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=295.6857777777778) [capacity. They're not used for data](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=297.85111111111115) [storage. If the preemptible instances that](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=299.9638181818181) [you've added as worker nodes to your](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=302.066) [cluster are reclaimed by Google Compute](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=304.6093333333334) [Engine, then the jobs that you're running](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=306.74754545454545) [on those preemptible instances will slow](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=308.99836363636365) [down. But those jobs will not be stopped.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=311.1474545454545) [They'll continue running, albeit more](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=314.306) [slowly, on your permanent instances. If](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=316.93000000000006) [you're working with Dataproc on the Google](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=319.82161538461537) [Cloud, make sure you use the default](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=321.69492307692315) [Hadoop distribution. If you try to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=323.679) [shoehorn a third-party distribution to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=325.61699999999985) [GCP, you might have to deal with issues](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=327.1501538461539) [that are very specific to you. If you want](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=329.40676923076944) [to tailor your VMs in the Hadoop framework](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=332.6257692307691) [specific to your use case, you should use](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=335.799) [initialization actions and cluster](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=339.03899999999976) [metadata to do so. Previously when you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=340.41049999999996) [were on on-premises Hadoop, you might have](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=342.6656666666665) [done this using config files or third-party configuration tools.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=9&mode=live&start=344.90227272727265)

[Dataproc Pricing](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live)

[If you're moving all of your data and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=2.497) [compute to the cloud, how much is Dataproc](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=4.384999999999998) [going to cost you? Now the cost for](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=5.968) [Dataproc is primarily the price you pay](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=9.131692307692306) [for the VMs that run your compute](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=11.961533333333334) [operations plus a small additional charge.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=14.945400000000001) [There is a small additional charge for the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=17.503) [Hadoop software or Spark, Hive, or Pig, or](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=19.759) [any of the other technologies that you use](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=22.02825) [from the Hadoop ecosystem. They're all the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=24.31075) [open-source versions. The main cost](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=26.7102) [component in your Dataproc cluster is the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=28.50675) [Compute Engine instances that you need to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=30.741500000000002) [keep up and running. GCE VMs come in a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=33.18166666666666)[variety of flavors. There are standard,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=36.06618181818182) [predefined machine types, and there are](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=37.798) [custom machine types as well. Over and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=40.355090909090904) [above the VM prices, you get sustained](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=43.13981818181818) [discounts for continuous usage of these](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=45.42818181818181) [VMs. And if you commit upfront to a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=47.336153846153856) [certain period of usage, committed](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=49.844461538461566) [discounts apply as well. If you choose to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=52.230000000000004) [add additional virtual CPUs for more](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=55.38709090909092) [compute capacity, there is an additional](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=57.014727272727306) [fee per vCPU per hour. This additional fee](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=60.070000000000014) [for each vCPU can range from 1 cent to 64](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=64.20678571428572) [cents. Remember, this is per vCPU per](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=68.71607142857148) [hour. Now if you really want to understand](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=71.69627272727271) [Dataproc pricing, you need to understand](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=75.1049090909091) [GCE VM pricing, and this is discussed](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=77.7067272727273) [extensively in the prereqs for this](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=80.3558181818182) [particular course, which talks about](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=82.494) [Google Compute Engine solutions. If you're](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=84.71677777777776) [storing your data on cloud storage buckets](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=87.03366666666668) [as opposed to HDFS, the charges for this](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=88.91900000000003) [data are completely separate and are based](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=91.69658333333335) [on the kind of storage that you're using,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=94.454) [whether it's a multi-regional bucket, a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=96.26054545454546) [regional bucket, and so on. If you use](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=98.25687500000001) [other GCP services from Dataproc, you'll](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=101.36425) [be billed for those separately, for](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=103.18275) [example, Stackdriver monitoring for your](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=105.6922) [Dataproc jobs. If your Dataproc jobs](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=108.3592) [connect to BigQuery or BigTable, which are](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=110.87966666666667) [other cloud storage technologies on the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=113.42358333333333) [Google Cloud, you'll be billed for those](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=115.46927272727272) [separately as well. And on this note, we](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=117.79728571428572) [come to the end of this introductory](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=120.27157142857149) [module on Cloud Dataproc. We studied how](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=122.89399999999999) [Hadoop works first, which is a ubiquitous](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=126.16735714285716) [big data processing technology. One](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=128.78385714285713) [drawback of Hadoop is that it tightly](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=130.6615) [couples storage and compute. Data is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=133.33899999999994) [stored on the Node on which it's](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=135.569)[processed. This can lead to utilization](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=138.01899999999995) [inefficiencies and cluster sizing](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=140.31955555555555) [inefficiencies because if you have to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=141.88) [scale storage, you end up scaling compute](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=143.78363636363633) [as well, and vice versa. Managed Hadoop](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=146.1909090909091) [offerings on cloud platforms allow you to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=149.06409090909088) [decouple storage from compute and, thus,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=151.8254545454546) [allow you to independently scale either of](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=154.154) [these, and Dataproc is Google's managed](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=156.45572727272724) [cloud Hadoop offering. In addition, we saw](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=158.47) [that Dataproc has other features such as](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=160.91650000000004) [automatic cluster resizing, scheduled](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=164.05066666666673) [deletion of your cluster, which gives it](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=165.841) [powerful advantages over on-premises](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=168.62699999999998) [Hadoop. In the next module, we'll get](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=170.15316666666666) [really hands on. We'll create our Dataproc cluster and use it to run MapReduce jobs.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=d287cf96-4694-4d9d-a283-8bce9af091c1&clip=10&mode=live&start=172.70874999999998)

[Running Hadoop MapReduce Jobs on Google Dataproc](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=0&mode=live)

[Module Overview](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=0&mode=live)

[Hi, and welcome to this module where we'll](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=0&mode=live&start=0) [see how we can set up our first Dataproc](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=0&mode=live&start=2.8912000000000004) [cluster, and then use it to run Hadoop](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=0&mode=live&start=6.3660000000000005) [MapReduce jobs. We've already seen that](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=0&mode=live&start=10.146) [Dataproc is Google's managed Hadoop](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=0&mode=live&start=12.456999999999999) [offering on the cloud, which means that](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=0&mode=live&start=14.164499999999999) [the MapReduce jobs that you have running](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=0&mode=live&start=16.308500000000002) [on your on-premises cluster can be](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=0&mode=live&start=18.421916666666668) [migrated to Dataproc on the cloud fairly](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=0&mode=live&start=20.287416666666672) [seamlessly. In this module, we'll see how](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=0&mode=live&start=22.642666666666663) [we can create a Dataproc cluster, and once](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=0&mode=live&start=25.44499999999999) [the cluster has been created, we'll see](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=0&mode=live&start=28.156499999999998) [how we can connect to it in a variety of](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=0&mode=live&start=30.227769230769233) [ways. We'll use the web console, as well](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=0&mode=live&start=33.03161538461538)[as work from the command line using the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=0&mode=live&start=35.083) [gcloud command line utility. We'll also](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=0&mode=live&start=37.22875) [see how we can use Stackdriver monitoring](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=0&mode=live&start=39.47724999999999) [in order to monitor the jobs that we have](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=0&mode=live&start=41.985181818181815) [running on our cluster including cluster](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=0&mode=live&start=44.839) [metrics themselves. We'll also see how we](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=0&mode=live&start=48.302499999999995) [can use initialization actions in order to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=0&mode=live&start=51.07741666666667)[configure the nodes on our cluster.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=0&mode=live&start=53.346) [Initialization actions on the GCP allow us](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=0&mode=live&start=55.801090909090924) [to configure and initialize a software that we need to customize our VM.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=0&mode=live&start=58.54833333333334)

[Creating a Dataproc Cluster Using the Web Console](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live)

[In this demo, we'll use the web console in](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=2.612) [order to create our very first Dataproc](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=4.643428571428571) [cluster, and then we'll see how we can log](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=6.367142857142857) [in to the master node using SSH. Log on to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=9.045285714285715) [your GCP account, and we'll start off in](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=13.609400000000003) [the main console of our spikey-dataproc](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=15.34113333333334) [project. We are in the dashboard right](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=17.014000000000003) [now. Use the hamburger icon on the top](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=19.83850000000001) [left in order to view the products and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=22.17149999999999) [features that GCP has to offer. You can](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=24.38372727272727) [scroll down, and under Big Data, you'll](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=26.64554545454544) [find the Dataproc option. GCP gives you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=29.737666666666662) [the option to pin commonly used services](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=32.65192857142856) [to the very top of this menu. So you can](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=34.73542857142854) [click on the pin button and have Dataproc](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=37.620142857142866) [show up at the top of the list of services](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=39.879) [that you see here. Click on Clusters](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=42.96400000000002) [within Dataproc, and let's create our](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=45.49672727272728) [first cluster. Our Clusters page is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=48.09418181818183) [completely empty here. We've enabled the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=50.169142857142866) [Dataproc API, so we can click on the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=51.95585714285715) [button here to create our cluster. If you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=54.302142857142876) [haven't enabled the Dataproc API, you'll](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=57.2495) [see an option here to enable the API first](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=59.64286666666667) [before you create your cluster. Here is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=62.5756666666667) [the configuration page on the web console](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=64.76357142857144) [where you can set up what your cluster](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=66.71857142857148) [looks like. You can give your cluster a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=68.82124999999999) [name. I'm going to call it](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=70.70458333333332) [spikey-cluster-one. Remember, you can](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=72.86349999999997) [create clusters on the fly for specific](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=74.40229999999995) [workloads. You ought to name your cluster](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=77.12614285714284) [based on the workload that you plan to run](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=79.34262500000004) [here. The next option that you see here](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=82.0825625000001) [asks you to specify a region for the nodes](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=84.1001333333333) [in your cluster. A region on the GCP](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=86.4675) [refers to independent geographic areas,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=89.74349999999995) [and these geographic areas are made up of](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=91.83636363636363) [zones. In addition to a region, you also](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=95.29381818181817) [have the option to specify specific zones](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=97.94230769230772)[where your cluster nodes will be located.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=100.227625) [A zone can be thought of as a single](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=102.927) [deployment area within a geographic](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=105.92279999999995) [region. A zone is a single failure domain](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=107.58713333333326) [on the GCP. Notice that the default option](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=110.72349999999996) [for region has been set to global. Global](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=113.74583333333331) [is a special multi-region endpoint. And](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=116.67972727272728) [when you use this endpoint for your](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=118.98809090909093) [cluster, your cluster resources can live](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=121.6929090909091) [in any user-specified zone, and your](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=124.02472727272729) [cluster can interact with these resources.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=126.75960000000002)[This is a special multi-region namespace.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=129.696) [You also have the option here to indicate](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=132.1467272727272) [that you want all your cluster resources](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=135.00730769230768) [to be within a specific region, and here](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=136.91669230769227) [are the various regions that are available](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=139.13500000000002) [in the US, in Europe, in Asia, and so on.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=141.30818181818182) [When you choose a specific region from](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=146.24) [this list, you isolate your cluster](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=148.15741666666656) [resources to that region. This includes](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=149.90854545454545) [any VM instances that your cluster](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=152.1978181818182) [instantiates and any cloud storage buckets](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=154.56781818181815) [that you use. The advantage of isolating](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=157.09927272727265) [your cluster resources to a specific](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=159.72972727272725) [region is improved performance. Because](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=161.7768181818181) [all your nodes are located in the same](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=163.39327272727274) [region, they can communicate with each](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=166.0758181818182) [other with low latency. Once you've chosen](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=168.78072727272735) [a specific region, you can use this](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=171.49749999999997) [drop-down to choose a specific zone within](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=173.6289999999999) [the region. We'll have all our cluster](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=176.38728571428572) [nodes be in asia-east1-b, but our cloud](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=179.25908333333328) [storage buckets can live in any other](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=182.076) [region because of the global namespace.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=184.8523076923077) [Below that, you have a drop-down for the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=187.232) [different kinds of clusters that you can](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=189.15866666666673) [instantiate. The standard configuration is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=190.9600909090909) [when we have 1 master and N workers.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=192.35690909090903) [That's what we're going to choose here.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=194.71124999999995) [You can have a single node cluster or](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=196.834) [configure your cluster for high](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=198.91769230769222) [availability with three masters as well.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=200.22) [If you scroll down a bit, you'll see that](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=202.35249999999994) [you can choose the kind of VM instance you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=205.2162) [want to set up your master node. Here I've](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=207.4097999999999) [chosen 2 vCPUs. These are predefined](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=211.17966666666663) [machine types that you see here on this](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=213.59475000000003) [list. You can use the customize link to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=215.98341666666676) [customize your machine type as well. You](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=218.90323076923073) [can then move on to configuring the size](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=221.381) [of the persistent disk that you want](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=223.22899999999996) [attached to your master node. I'm going to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=224.90749999999997) [change the size to be 10 GB. I'm not](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=226.80175000000003) [planning to use it to store very much](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=228.71200000000013) [data. For your master node, you can choose](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=230.7784285714286) [to go with a standard persistent disk, or](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=233.21328571428586) [you can choose an SSD persistent disk.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=236.1966153846153) [Solid-state drives are for when you want](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=238.967) [higher performance and lower latency for](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=241.26109090909083) [your reads. SSDs are, of course, more](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=243.3307272727273) [expensive than standard persistent disks.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=245.98627272727285) [Once you're done with the master node, you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=247.96600000000004) [can then go on and configure the kind of](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=251.03546666666668) [machine that you want for your worker](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=253.0196666666668) [nodes. I'm going to go with the less](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=254.73000000000002) [powerful, predefined machine type here](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=257.1619999999999) [with just 1 vCPU. I'll reduce the size of](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=259.63) [my persistent disk to 10 GB. Remember,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=263.0912727272728) [persistent disks are pay as you allocate.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=264.98487500000005) [So if you allocate a larger sized disk,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=267.343) [you'll end up paying for it as well. It's](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=269.6295000000002) [preferable to go with the lower sized](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=272.15515384615384) [persistent disk in order to reduce your](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=273.8292307692307) [costs, especially if you plan to store](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=275.8231333333333) [your data on cloud storage buckets. We'll](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=278.6166) [accept the default options for all of the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=281.2645454545455) [other settings. We are creating a cluster](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=283.2929090909093) [with one master and two workers. Two](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=285.87545454545455) [worker nodes are the minimum requirement.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=288.8642857142857) [Click on the Advanced options link that](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=291.854) [you see here onscreen, and let's explore](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=293.8973636363636) [what other options are available to us.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=296.01089999999994) [You can see at the very top that you can](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=298.218) [add preemptible worker nodes using the UI](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=301.140857142857) [when you create your cluster. We'll leave](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=303.31499999999994) [it 0 now. We'll see how to create a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=305.4078461538462) [cluster with preemptible nodes later on](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=307.2881538461541)[using the command line. When you create a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=309.4865000000001) [project on the GCP, a VPC is created for](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=312.39713333333333) [you by default, and it's preconfigured](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=314.93993333333333) [with some standard firewall rules. This](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=317.3703333333334) [VPC is referred to as the default network,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=319.9930714285714) [and all of the cluster nodes that you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=322.3216428571426) [instantiate will by default be on this](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=325.1354545454546) [network. The default network on the Google](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=328.305181818182) [Cloud comes with a bunch of preconfigured](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=330.141) [firewall rules. These firewall rules](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=331.9674) [enable VMs on the network to communicate](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=334.1844) [with each other and allow only limited](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=337.04481818181824) [external traffic such as SSH connections](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=339.5807272727274) [to your VM. There are other options you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=342.184875) [can configure here as well. You can](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=345.1135000000001) [specify a cloud storage bucket that](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=347.4518333333333) [processes on Dataproc can use for staging.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=349.7163333333332) [You can also use the web console to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=352.39475000000004) [configure initialization actions that](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=355.51141666666683) [allow you to customize the software that](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=357.0368) [you have installed on your cluster VMs.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=359.687) [Let's hide these advanced options here by](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=362.7829090909091) [clicking on Less. And you can see here at](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=365.5160714285715) [the bottom that you have a command line](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=368.40828571428585) [option. If you click on the command line,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=370.54776923076935) [this will give you the gcloud command that](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=372.5915) [you can run from any terminal window in](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=374.9715000000001) [order to create this cluster that you've](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=377.72774999999996) [just configured. You can use this gcloud](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=380.6881666666665) [command in order to write a script that](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=382.7432857142858) [will create your cluster before you run](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=384.67814285714303) [your workload. If you don't want to use](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=387.43249999999995) [the gcloud command line utility, you have](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=389.79509090909073) [other options as well such as the REST](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=391.6963636363636) [API. You can use the REST API. You can see](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=394.21781818181825) [the POST request that you need to use for](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=396.76920000000007) [cluster creation right here on the screen.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=399.4159000000001) [Close this window, and let's click on the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=402.113) [Create button and create our very first](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=404.4163846153844) [Dataproc cluster. You might have to wait](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=406.48784615384614) [for a couple of minutes for all of the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=408.6993076923076) [cluster nodes to be up and running. And](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=411.30966666666677) [here it is, our very first Dataproc cluster.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=1&mode=live&start=414.195)

[Using SSH to Connect to the Master Node](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live)

[Now that we have our first Dataproc](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=2.664) [cluster, let's click through on our](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=4.642666666666667) [Cluster dashboard page and see what this](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=6.369454545454546) [cluster looks like. You can see a number](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=8.563636363636364) [of tabs on top here which gives you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=10.559714285714284) [information on this cluster. The first tab](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=12.248857142857139) [is the Overview tab. This will give you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=14.602599999999995) [graphical information on various cluster](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=17.545800000000007) [metrics. They haven't been turned on yet](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=19.439300000000017) [because you haven't done anything with](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=21.472714285714286) [your cluster. If you click on the Jobs](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=23.12742857142857) [tab, that will show you what jobs have](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=24.965400000000006) [been submitted to your cluster, what jobs](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=26.680600000000016) [are currently running. The Configuration](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=29.74899999999999) [tab will give you configuration details](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=31.7732) [for the cluster that you just created, the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=33.5648) [number of master nodes, the worker nodes,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=36.03340000000001) [the kind of machines that you set up for](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=38.07685714285715) [the master, as well as worker nodes, and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=40.464428571428606) [other such information. You can see an](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=43.27633333333333) [interesting thing at the bottom here. A](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=45.60220000000001) [default cloud storage staging bucket has](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=47.68508333333334)[been created for you along with the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=49.51558333333335) [cluster. This has been done by GCP under](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=51.90083333333334) [the hood. If you click on the VM Instances](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=55.340166666666704) [tab that you see here onscreen, this will](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=57.93546153846156) [give you information of all of the nodes](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=59.85514285714285) [that make up your cluster, the master](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=61.58371428571425) [node, as well as the worker nodes. You can](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=63.439800000000005) [see here that SSH has been enabled by](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=66.47875) [default only for the master node. Your](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=69.37775) [firewall rules have been configured such](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=71.85033333333334) [that you can only SSH into the master](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=73.64033333333333) [node, not into worker nodes. You can use](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=76.41525000000001) [the links on top to perform other](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=79.449) [administrative actions as well, such as](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=81.53733333333332) [deleting nodes from your cluster, viewing](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=83.23433333333331) [logs from your cluster, or refreshing this](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=84.88483333333328) [page to get the latest updates. There are](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=87.7368181818182) [different ways that you can use to SSH](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=90.92614285714285) [into the master node of your cluster. You](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=92.63471428571427) [can use the gcloud command line utility,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=94.72245454545455) [or you can use another SSH client that you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=96.881) [have installed on your machine. The](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=99.64850000000004) [simplest way is to open the SSH terminal](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=101.504) [window in a new browser window, and you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=104.04799999999999) [can simply click on the SSH button here,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=106.37128571428572) [and that will open up another terminal](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=108.32671428571429) [window on the browser. This is the master](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=110.719) [node of your Hadoop cluster, and a bunch](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=113.32569230769231) [of software that Hadoop requires to run](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=115.22107692307692) [has already been enabled. Java has already](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=118.04879999999997) [been installed, and you can see the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=120.54599999999998) [current version of Java by running java](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=122.3426666666666) [--version. You'll see that jdk 1. 8 is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=125.1694375) [powering this cluster. The master node](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=128.8709375) [also has Python installed. The current](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=131.22371428571424) [version of Python is Python 2. 7. In](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=133.32755555555553) [addition to Hadoop, your cluster also has](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=136.5291) [Spark running. PySpark, Sparks Python](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=138.90980000000005) [shelf, is also available here. You can](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=140.97974999999997) [check the version of PySpark. You can see](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=144.00866666666667) [it's version 2. 2. 1. Your cluster also](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=146.85933333333338) [comes with Pig installed. You can write](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=149.81442857142855) [Pig Latin scripts to perform extract,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=152.4146) [transform, and load operations. If you are](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=154.48580000000004) [currently using Hive as your data](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=157.04950000000005) [warehouse on your on-premises setup, Hive](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=159.268) [is available here as well. You can see](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=161.4541818181818) [that Hive version 2. 1. 1 has already been](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=164.54800000000003) [preinstalled. The version of Hadoop on the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=168.3640000000001) [GCP at the time of this recording is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=171.1086666666667) [version 2. 8. 4. GCP is responsible for](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=174.20288888888894) [managing and updating the versions of all](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=177.16899999999998) [of the software that's installed on this](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=179.65224999999995) [cluster. Hadoop, Hive, Pig, Spark, all of](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=182.56642857142862) [these will be independently updated. The](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=186.53300000000002) [exit command will allow you to exit this](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=189.74238461538462) [SSH connection and go back to your cluster](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=191.81746153846157) [dashboard. Now let's go back to our main](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=194.11333333333332) [dashboard on the Google Cloud Platform.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=196.63) [Under Resources, you can see that we have](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=198.38800000000003) [three compute engine instances running.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=201.338375) [These correspond to the three nodes of our](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=203.814) [cluster, and we have one cloud storage](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=205.7056923076922) [bucket. If you click on the three compute](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=207.5961818181818) [engine instances, you'll be taken to the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=210.428909090909) [VM instances dashboard. This will show you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=213.36433333333335) [all of the VMs you have in this project.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=215.55992307692307) [Right now we only have the cluster VMs](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=217.596) [corresponding to the nodes in our Dataproc](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=219.8649230769231) [cluster. Let's head back to our main](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=222.1838333333333) [project dashboard page. And under](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=225.33674999999988) [Resources, let's click on the cloud](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=227.36379999999997) [storage bucket. Remember, we didn't start](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=229.3911999999999) [off with a bucket, but once you go in](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=231.1175714285714) [here, you can see that Dataproc clusters](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=233.48135714285706) [automatically create a staging bucket that](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=236.33900000000003) [they can use. In addition, clusters also](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=238.79533333333333) [have installed a cloud storage connector,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=241.33750000000003) [which is software which allows them to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=243.534) [work with cloud storage buckets, connect](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=246.117) [to cloud storage buckets, get data from them, add data to them.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=2&mode=live&start=248.08599999999996)

[Creating a Firewall Rule to Enable Access to Dataproc](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live)

[When you're working with Hadoop, you need](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=3.027) [access to YARN's resource manager UI, as](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=5.05190909090909) [well as the HDFS web UI in order to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=7.333800000000001) [monitor your cluster. You need to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=10.708199999999998) [configure firewall rules so that you can](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=12.670545454545456) [access information from your cluster, and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=14.624818181818187) [that's what we'll do in this demo. Click](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=16.164571428571428) [on the hamburger icon on the top left to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=18.377142857142864) [access all of GCP services. And under](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=20.80977777777778) [Networking, go to the VPC network option.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=23.01177777777778) [Under here, let's go to Firewall rules.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=26.126) [This is where we can create new firewall](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=28.334230769230764) [rules. Even though we haven't explicitly](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=30.953500000000002) [set up a single firewall rule for our](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=33.133) [project so far, you can see a number of](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=35.643714285714296) [rules are already available here. These](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=38.023571428571465) [are preconfigured firewall rules for the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=40.630222222222216) [default VPC that has been automatically](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=43.38044444444444) [created. Notice that these rules allow SSH](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=46.00977777777778) [and Remote Desktop access to our VMs. By](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=48.94353846153847) [default, these rules also allow all VMs on](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=52.148999999999994) [the same network to communicate with each](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=54.50099999999997) [other. Click on the Create Firewall Rule](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=56.772999999999996) [on top in order to create a new firewall](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=59.33033333333332) [rule that will apply to our cluster. If](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=62.14700000000001) [you're interested in firewall rules on the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=64.50572727272727) [GCP in general, here is a link that you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=66.41481818181813) [can study. It's good to give your firewall](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=69.14142857142858) [rule a meaningful name so that you know](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=71.686) [what it represents. I'm going to call it](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=73.71145454545453) [default-allow-dataproc-access, allow us](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=76.438) [access to the Dataproc cluster. Give it a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=77.6182) [meaningful description if you want to.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=81.1325) [This rule that we're going to create](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=83.86) [applies to the default network because](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=85.75991666666667) [that's where our cluster nodes live.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=87.52928571428572) [That's the only network available in our](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=90.003) [project at this point in time. Firewall](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=92.2629230769231) [rules can be used to allow or deny access](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=94.53546153846155) [to network resources. And every firewall](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=97.55461538461545) [rule is associated with a priority. Lower](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=99.799) [numbers indicate higher priority. I'm](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=102.631) [going to assign a priority of 1000 for](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=104.41099999999997) [this firewall rule. This gives me some](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=107.44985714285714) [room to play around. Higher priority rules](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=109.64933333333335) [will have lower values than 1000. Lower](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=111.3789166666667) [priority rules will have higher values](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=113.71509090909089) [than 1000. Firewall rules are explicitly](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=116.05266666666667) [configured applying to ingress or egress](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=119.63266666666665) [traffic. Ingress traffic applies to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=122.34360000000001) [incoming traffic, traffic incoming to your](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=124.24060000000003)[cluster nodes. And egress traffic talks](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=126.75785714285713) [about traffic outgoing from your cluster](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=129.10136363636363) [nodes. We'll apply an ingress rule here to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=131.33390909090903) [allow certain kinds of traffic to reach](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=133.57230769230773) [our cluster nodes. You then specify the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=135.50246153846155) [kind of action that you want to apply to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=137.8156923076924) [this ingress traffic. You can choose to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=140.6156666666666) [allow specific kinds of traffic or deny](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=142.6240833333332) [specific kinds of traffic. I'm going to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=145.1868) [choose an allow rule here. I want to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=147.63976923076925) [explicitly allow certain kinds of traffic.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=150.13515384615388) [You can fine-tune your firewall rules](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=152.51020000000003) [further by specifying the kind of targets](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=155.137) [to which you want your firewall rules to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=156.93716666666674) [apply. You can have your firewall rule](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=159.1836923076923)[apply to all instances that live on this](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=161.42530769230763) [network, to instances with specific tags,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=164.44418181818187) [or to instances which use a specific](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=166.914) [service account. In order to keep things](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=169.8485555555556) [simple, we'll apply this firewall rule to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=171.9027692307693) [all instances that are on this network. If](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=173.60484615384632) [you want this firewall rule to apply to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=176.31415384615394) [only traffic from a specific source,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=179.10683333333333) [that's possible as well. You can specify a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=181.14333333333332) [source filter to make your rule more](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=184.04099999999997) [specific. In the case of egress rules, you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=186.7359999999999) [can specify a corresponding destination](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=188.8671818181818) [filter as opposed to a source filter. For](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=190.35025000000002) [our ingress firewall rule, we can filter](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=194.08725) [based on IP ranges, subnets, source tags,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=196.469) [or service accounts from where the traffic](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=199.46200000000005) [originated. Here I want to be a little](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=203.27000000000012) [specific. I want to allow traffic only](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=205.45499999999996) [from my local machine to this cluster. So](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=207.096) [I'm going to filter based on IP ranges. In](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=209.51323076923086) [order to figure out what my IP is, I'm](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=212.4351875) [going to search for my IP on the browser.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=214.7948750000001) [Here is my public IP address. Your public](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=217.59999999999997) [IP address will be a little different.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=220.62288888888887) [Copy this IP address over, and we'll](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=222.579) [specify this in the IP range. I'll be able](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=225.45492307692317) [to connect to instances on this network](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=228.64184615384613) [only from my local machine, not from any](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=230.3261538461538) [other machine. You can choose an](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=233.7944285714285) [additional source filter if you want to.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=235.95445454545458) [For the purposes of this demo, we'll just](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=237.812) [stick with one. I only want to enable](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=240.6083636363636) [traffic which follows certain protocols to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=243.28745454545455) [certain ports on our cluster. I want to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=245.244) [use TCP to connect to the resource manager](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=247.51714285714283) [UI on my Hadoop cluster, which is present](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=249.86983333333333) [on port 8088 and to the HDFS UI present on](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=252.46116666666663) [port 9870. We are now finally ready to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=258.52180000000004) [create the firewall rule allowing us](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=261.46799999999985) [access to our cluster. Click on the Create](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=263.061) [button and wait for a couple of minutes for the firewall rule to be up.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=3&mode=live&start=265.0942307692306)

[Accessing the Resource Manager and Name Node UI](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=4&mode=live)

[We now have the firewall rule configured](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=4&mode=live&start=2.573) [that allows us access to the cluster nodes](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=4&mode=live&start=4.306846153846155) [from our local machine. We are now on the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=4&mode=live&start=6.5001428571428574) [Dataproc dashboard page. I'm going to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=4&mode=live&start=9.682285714285712) [click through to spikey-cluster-one, and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=4&mode=live&start=11.475818181818184) [I'm going to click on the VM Instances](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=4&mode=live&start=12.697181818181823) [tab. You can see our one master and two](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=4&mode=live&start=15.104285714285712) [worker nodes here. Click through to the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=4&mode=live&start=18.118) [master node, and you can see from the VM](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=4&mode=live&start=20.32569230769231) [instance details page that we have access](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=4&mode=live&start=23.164153846153848) [to the external IP of this master node VM.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=4&mode=live&start=25.37184615384616) [This is the IP address that we can use to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=4&mode=live&start=28.19262500000001) [connect to this VM instance from our local](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=4&mode=live&start=31.317) [machine. So copy over this IP address, and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=4&mode=live&start=34.33566666666666) [let's open up a new browser window and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=4&mode=live&start=36.666538461538465) [specify the 8088 port on this VM instance.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=4&mode=live&start=39.37730769230771) [This is the port on the master node where](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=4&mode=live&start=42.4892857142857) [the Hadoop resource manager is available.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=4&mode=live&start=46.146857142857144) [We are now going to access this web](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=4&mode=live&start=48.618) [console from our local machine. And you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=4&mode=live&start=50.01888888888892) [can see this familiar web console](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=4&mode=live&start=52.52938461538463) [onscreen. This is the same web console](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=4&mode=live&start=54.864769230769255) [that you'd use with on-premises Hadoop.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=4&mode=live&start=57.43250000000001) [It's available for you on Dataproc as](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=4&mode=live&start=59.611) [well. This is the same UI that you have on](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=4&mode=live&start=62.64637499999999) [your on-premises cluster. If you click on](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=4&mode=live&start=65.57899999999998) [the About link on the left, this will give](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=4&mode=live&start=67.20025000000001) [you all of the information you need on the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=4&mode=live&start=68.97381250000002) [cluster. If you click on the Nodes link to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=4&mode=live&start=71.48823076923077) [the left here, you'll see that this](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=4&mode=live&start=74.29923076923077) [cluster has two data nodes, exactly what](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=4&mode=live&start=76.28453846153849) [we configured. Both of these nodes are](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=4&mode=live&start=78.56118181818181) [currently in the running state, and you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=4&mode=live&start=81.2262727272727) [can see that the node addressed for each](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=4&mode=live&start=83.05471428571425) [of these nodes is a GCP-specific one, one](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=4&mode=live&start=84.985) [that GCP has assigned. In addition to the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=4&mode=live&start=88.30433333333336) [resource manager web console, our firewall](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=4&mode=live&start=91.094) [rule was also configured to allow our](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=4&mode=live&start=92.78899999999999) [local machine to access the NameNode web](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=4&mode=live&start=95.424) [console available at port 9870. Accessing](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=4&mode=live&start=98.322) [this URL brings up the familiar NameNode](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=4&mode=live&start=102.53975) [web console for Hadoop. Once again, this](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=4&mode=live&start=105.77199999999996) [UI should be very familiar to you. You can](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=4&mode=live&start=108.2868) [scroll down here and get a bunch of](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=4&mode=live&start=110.35585714285713) [different pieces of information. You can](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=4&mode=live&start=112.39928571428567) [see that this cluster has two live nodes.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=4&mode=live&start=114.11560000000001) [You can click on the Datanodes tab on top](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=4&mode=live&start=116.894) [in order to get information on your worker](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=4&mode=live&start=119.4365) [nodes. Having these familiar web consoles](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=4&mode=live&start=122.12700000000001)[available to your Hadoop developers and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=4&mode=live&start=124.68300000000002) [administrators makes migration to cloud Dataproc easier.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=4&mode=live&start=129.535)

[Upload Data and MapReduce Code to Cloud Storage](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live)

[With our cluster set up, let's see how we](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=1.815) [can run our first MapReduce on Dataproc.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=4.6736249999999995) [When working with on-premises Hadoop, one](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=6.897) [standard practice that organizations might](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=8.859666666666667) [follow would be to log in to the master](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=10.423461538461536) [node to kickstart their MR jobs or](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=13.044538461538453) [MapReduce jobs, and that's exactly what](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=15.43563636363636) [we'll do here. Later we'll see how we can](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=17.606) [use the gcloud command line utility to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=20.104785714285725) [kickstart MapReduce jobs without logging](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=22.423) [in to the master node. When we migrate to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=24.747999999999998) [the cloud, though, we want to store our](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=28.001307692307698) [data, as well as our scripts on cloud](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=29.857923076923093) [storage buckets. Go to the hamburger icon,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=32.614666666666665) [and click on Storage and the Browser](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=34.943307692307705) [option within storage, and this will show](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=37.193538461538495) [you all of the buckets that we have](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=39.15339999999999) [created. This is the staging bucket that](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=41.33793333333331) [our cluster created. We're going to create](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=43.60033333333332) [a brand-new one where we'll store the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=45.66300000000001) [scripts and the data for our MapReduce.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=47.53200000000003) [I'm going to call this bucket spikey-data.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=50.12792307692309) [It's a multi-regional bucket, and this is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=52.845) [fine because our cluster is in the global](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=54.94791666666664) [namespace. It should be able to access](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=57.41023076923078)[data in a multi-regional bucket that's](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=59.65076923076926) [based in the US. Go ahead and click on the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=61.77114285714285) [Create button, and create your spikey-data](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=65.70415384615386) [bucket. Within the spikey-data bucket, I'm](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=68.09030769230772) [going to create a folder where I'll store](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=69.67633333333335) [my MapReduce scripts. I'm going to call it](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=71.8045) [mapr\_scripts. This is what will hold our](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=75.52050000000001)[mapper and reduce our code, which we'll](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=77.49481818181819) [write in Python. Create another folder](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=79.77081818181817) [here called mapr\_input. This is the folder](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=82.82427272727269) [where we'll store the input data that our](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=85.48546153846155) [MapReduce will act on. And here is the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=87.75438461538465) [data that we're going to upload to this](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=91.32074999999999) [cloud storage e-commerce bucket. This is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=93.61024999999997) [some sample data that we have from our](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=95.16966666666667) [Spiky Sales organization containing](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=96.82833333333338) [e-commerce transactions. Every record here](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=99.3545) [has the following bits of information, the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=102.351) [time at which a particular purchase was](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=104.41600000000001) [made, the location of the customer, what](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=106.69054545454544) [the customer bought, the product category,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=109.05383333333333) [the price, and how the customer paid,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=111.10483333333329) [Amex, Visa, or MasterCard. This is typical](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=113.88766666666666) [information for a sale day, and our](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=117.02309090909091) [MapReduce operation will calculate how](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=119.32863636363638) [much customers spent on each product. Here](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=121.57800000000002) [is the code for our mapper, which reads in](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=124.71593333333333) [one line at a time. It splits the data in](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=127.03733333333327) [that line, extracts that information, and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=130.16950000000006) [outputs just the item and the price. Here](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=132.30842857142858) [is the code for the reducer, which is a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=134.6798571428572) [script written once again in Python. We](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=137.59400000000002) [are tracking the total\_price paid per](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=140.20175) [item. The keys are the products or the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=142.96024999999992) [items that the customer purchased. The](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=146.41133333333335) [output of the mapper is fed into the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=148.8696153846154) [reducer, extracts the product and the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=151.4425384615386) [price at which it was purchased, and then](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=153.42115384615389) [performs an aggregation. This is the sum](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=156.18976923076937) [aggregation to calculate how much money](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=158.66400000000004) [customers spent on a per-item basis. The](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=160.8180000000001) [exact structure and functionality of this](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=163.33972727272734) [code is beyond the scope of this](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=165.4706)[particular course. You have the mapper and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=167.77079999999995) [reducer Python files available for](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=169.9716363636364) [download, though, if you want to try them](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=171.51527272727282) [out. Let's switch over to our GCP web](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=174.222375) [console and go to the mapr\_input folder](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=176.69407692307686) [within our spikey-data bucket. We'll](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=178.9399) [upload the input data set, the e-commerce](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=181.04439999999994) [data set that we saw earlier. Our sample](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=184.05566666666664) [sale data for a particular day is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=187.233) [available in itemData. txt, and I'm going](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=189.05999999999997) [to go ahead and upload this. Once the data](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=191.62836363636367)[has been uploaded, it should be available](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=194.8688461538462) [in your cloud storage bucket and should be](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=196.70015384615397) [accessible by a MapReduce operation](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=199.46549999999996) [running on your cluster. Go back one level](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=201.446) [to the spikey-data bucket, and within](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=204.95266666666666) [here, let's go to mapr\_scripts. Here is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=207.6248571428571) [where we'll upload the Python code that](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=210.57199999999997)[runs our mapper, as well as our reducer.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=212.8469999999999) [Click on Upload files, choose mapper. py](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=215.517) [and reducer. py, and upload them. We now](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=217.89) [have our MapReduce code written in Python,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=220.80791666666664) [as well as our input data available in](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=223.42446153846154) [cloud storage buckets. And that's where](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=226.5223076923077) [we're going to access them when we run our](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=228.43123076923078)[MapReduce on our Dataproc cluster. Observe](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=230.67292307692313) [that we follow Dataproc best practices.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=233.40427272727277) [We've separated storage from compute.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=236.383) [We're storing our data on cloud storage](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=238.29492307692306) [buckets. We're going to run our processes on our Dataproc cluster.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=5&mode=live&start=240.97161538461532)

[Running MapReduce on Dataproc](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live)

[We're now ready to run MapReduce on our](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=2.359) [spikey-cluster-one. Start off in the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=5.501222222222223) [Dataproc dashboard, click through to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=7.470000000000001) [spikey-cluster-1, and click on VM](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=9.440000000000001) [Instances. We'll SSH into the master node](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=11.529) [in order to run MapReduce. When you're](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=14.49525) [using Dataproc on the cloud, HDFS is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=17.300461538461537) [available to you on all of the cluster](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=19.703076923076914) [nodes. If you run the hadoop fs command,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=22.312000000000005) [you'll see that HDFS is available, and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=24.784) [you'll see the directories that are](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=26.750416666666663) [available in HDFS. You can choose to use](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=28.52407142857143) [HDFS to store your data and to store your](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=31.476642857142846) [scripts as well, but that is tightly](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=34.39461538461538) [coupling your storage with compute. We'll](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=36.271692307692284) [store our data on the cloud. We'll use our](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=38.365199999999994) [cluster only for compute. And here is the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=41.7715) [Hadoop command that you would run to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=44.62616666666668) [kickstart your MapReduce, the Hadoop JAR.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=47.51399999999999) [We'll reference Hadoop's streaming JAR](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=50.045) [that comes preinstalled on your master](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=51.96249999999999) [node. This is the Java archive that will](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=54.26349999999998) [kickstart the MapReduce operation on your](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=56.76036363636364) [cluster. Now you need to specify where](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=58.73815384615385) [your files are located. Use the --files](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=61.43369230769231) [command line argument to specify where](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=64.112) [your mapper and reducer scripts live. You](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=66.388) [can point to a cloud storage bucket. There](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=69.04890909090908) [is no additional configuration required](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=72.11953846153845) [for Hadoop to work with cloud storage. The](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=74.14723076923073) [cloud storage connector comes preinstalled](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=77.4939) [with your Dataproc cluster. We've written](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=80.03340000000003) [our mapper and reducer code in Python. Use](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=82.81766666666667) [the --mapper and --reducer command line](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=85.87736363636364) [options to specify the command to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=88.42954545454549) [kickstart your mapper and reducer](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=91.126) [processes. These are both Python](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=93.61350000000002) [processes. We use the Python shell command](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=95.64299999999999) [for this. The --input command line](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=98.32399999999996) [argument points to where our input data is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=100.43333333333337)[located on which we want the MapReduce](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=103.12) [operation to run. You can point to a cloud](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=105.50769999999999) [storage bucket using gs://. Specifying](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=109.92579999999998) [data on cloud storage buckets as opposed](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=112.55954545454544) [to HDFS should simply involve changing](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=115.1533636363636) [your hdfs:// to gs://. We also want the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=119.77071428571428) [output of our MapReduce operation to be](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=125.57269230769231)[stored on cloud storage buckets rather](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=127.54615384615386) [than on HDFS. You can simply point to the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=129.88542857142863) [folder where you want the output to be](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=132.77219999999997) [placed. Mapr\_output doesn't exist yet.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=134.8457999999999) [Dataproc will automatically create this](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=137.27549999999997) [folder on cloud storage. Go ahead and hit](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=139.84507692307693) [Enter, and wait for your MapReduce process](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=143.0376923076924) [to complete. You'll see that the mapper](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=145.5619) [process runs first. You can see the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=147.727) [progress of your operation. Based on the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=149.8721818181819) [logs that you see on your screen, map has](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=152.0019375) [gone up to 87%. Now mapper has reached](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=154.71374999999995) [100%, the reduce operation has started and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=157.6553333333333) [completed. And you can see the final](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=160.69391666666652) [result in your cloud storage bucket. Let's](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=163.2115) [switch back to our GCP web console. Click](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=165.52908333333332) [on the hamburger icon, choose Storage and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=167.40174999999996) [Browser. Here is where all of our buckets](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=169.6517142857143) [are listed. Let's click through to the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=172.74714285714285) [spikey-data bucket. Here is the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=175.79100000000003) [mapr\_output folder. This did not exist](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=177.95510000000002) [before. This was created by Hadoop when it](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=180.46730000000002) [ran our MapReduce job. And within this](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=183.3678181818181) [folder here are the results of our](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=185.69074999999998) [MapReduce operation that completed](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=187.80183333333326) [successfully as you can see from the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=189.09799999999998) [\_SUCCESS file that was placed here in this](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=191.52349999999993) [cloud storage folder. Let's switch back to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=194.49237499999998) [where we SSHed into the master node of our](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=196.79840000000004) [cluster. You'll see here that we can use](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=199.61492857142855) [the hdfs dfs command with cloud storage](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=202.53207142857138) [buckets exactly as we would with HDFS. We](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=205.56418181818177) [use the -cat command here to list the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=209.32114285714286) [contents of our MapReduce output. This is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=212.24228571428583) [the first file that MapReduce outputted,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=214.72153846153847) [and here are the contents. You can see the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=216.68215384615385) [products and how much customers spent on](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=219.23) [those products for that sale day. You can](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=220.82949999999997) [see from this example that the change to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=222.96899999999997) [managed Hadoop on the cloud is not](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=225.30800000000002) [intrusive at all. If you've been using](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=227.7615000000001) [on-premises Hadoop, you can move to cloud](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=229.96866666666665) [storage buckets and work with them in exactly the same way as you did with HDFS.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=6&mode=live&start=231.9910833333333)

[Running MapReduce Using the gcloud Command Line Utility](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live)

[If you're using Dataproc on the GCP, you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=2.263) [won't typically log in to the master node](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=4.6378571428571425) [in order to kickstart your MapReduce jobs.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=7.132555555555556) [Instead you'll use the gcloud command line](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=9.63) [utility that allows you to do this from](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=12.168374999999997) [any terminal window that can connect to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=14.433923076923076) [your Dataproc cluster. Click on the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=16.388166666666667)[hamburger icon on the top left, and go to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=18.32716666666666) [the navigation menu that shows you all of](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=21.068142857142863) [the services. Let's go to Dataproc, and go](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=23.318428571428587) [to the Jobs page that will allow us to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=25.3194) [monitor the jobs that we submit to our](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=27.477) [cluster. Once we're in this page, activate](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=30.639666666666667) [the cloud shell. We'll use this as our](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=32.861) [terminal window in order to use the gcloud](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=35.13746153846155) [command line utility. In order to use](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=37.319) [gcloud, you need to make sure that you're](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=39.58215384615386) [authenticated. And you can do this by a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=42.26753846153846) [gcloud auth login. Now if you're working](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=45.24969230769229) [on the cloud shell, you don't really need](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=47.48285714285715) [this step. But if you're working with](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=49.712571428571444) [gcloud on your local machine, you need to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=51.5096153846154) [authenticate yourself. So I'm going to go](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=53.493) [ahead and click on Yes to continue. This](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=55.386000000000024) [will pop up a link. Click on this link to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=57.522333333333336) [get the authentication token. You need to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=59.95566666666666) [sign in with your GCP account. Yes,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=62.93600000000001) [indeed, I do want gcloud to perform all of](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=65.93) [these actions. I'm going to click on](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=68.24300000000005) [Allow, and this will generate a token that](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=70.03311111111108) [I need to copy/paste into my cloud shell](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=72.62169230769231) [window. If you now click on Enter, gcloud](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=75.75707692307694) [has now authenticated you as](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=78.35341666666663) [spikeysales@loonycorn. com. We're ready to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=79.928) [get started. Using the gcloud command line](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=82.58871428571429) [utility, I'm going to set my current](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=84.93133333333336)[project as spikey-dataproc. This is the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=86.721) [project that I want my current cloud shell](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=89.05990909090912) [session to use. All Dataproc jobs can be](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=92.23027272727273) [submitted using the gcloud command line.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=95.487) [Note that we are working from any terminal](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=97.65300000000002) [window. We haven't logged into our master](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=100.54454545454547) [node. Gcloud dataproc jobs submit hadoop](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=103.07981818181824) [will allow us to submit Hadoop MapReduce](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=105.8139230769231) [jobs. Specify the cluster where you want](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=109.13461538461546) [to submit this job, that is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=111.18579999999999) [spikey-cluster-one, our only cluster at](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=112.96245454545453) [this point in time. For this demo, we'll](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=114.77609090909088) [run one of the standard Java examples that](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=117.36300000000003) [are present when your cluster is set up.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=119.63500000000008) [This JAR file is located on the master](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=121.91149999999999) [node of the Dataproc cluster that we set](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=124.192) [up and can be accessed locally by our](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=126.49599999999997) [MapReduce job. This JAR contains the word](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=129.2205454545455)[count MapReduce that comes preinstalled](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=132.21781818181827) [with your Dataproc cluster. You can point](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=134.58700000000002) [to any other JAR which has your MapReduce](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=137.48624999999996) [code. You can point to JAR files that you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=140.0415833333332) [uploaded to cloud storage as well using](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=144.0970000000001) [gs://. JAR files can also live on HDFS. We](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=147.36600000000018) [want to run the word count MapReduce from](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=150.77071428571426) [within this JAR FILE, and we want to count](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=153.61969230769228) [the number of words in itemDated. txt that](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=156.52530769230756) [is located in our cloud storage bucket. We](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=159.84146153846154) [want the output of this MapReduce](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=163.24699999999999) [operation to be stored to our spikey-data](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=164.98699999999994) [bucket in mapr\_java\_output. Our MapReduce](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=167.54344444444445) [job has been successfully submitted to the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=171.62566666666666) [Dataproc cluster. It's running here as you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=176.1323333333333) [can see from the log output. If you click](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=178.19645454545454) [on Refresh within your Jobs page, you'll](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=180.25499999999997) [be able to see the current status of your](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=182.1589999999999) [job here. This job might take a minute or](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=184.8359285714286) [two to run. Instead of viewing logs on the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=187.696) [console window, you can click through to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=190.69872727272735) [the job here on this Jobs page, and you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=192.6896666666667) [can see logs right here on your web](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=195.07226666666676) [browser. Just wait a couple of minutes](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=197.722923076923) [more, and you'll see that your job will](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=200.1338) [complete, job output is now complete.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=203.06020000000007) [Let's go back to our main Jobs page. You](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=204.84599999999995) [can see from the green checkmark that the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=207.399) [job has completed. Let's go to our storage](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=209.64700000000005) [bucket to view the output. Within the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=212.44499999999996) [spikey-data bucket, we have the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=215.68488888888888)[mapr\_java\_output folder that has been](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=218.89711111111106) [created to store the final results of our](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=221.30614285714282) [word count operation. You can see that our](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=224.30385714285703) [MapReduce completed successfully. On your](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=227.22612499999997) [cloud shell terminal window, you can use](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=228.9016153846154) [the gsutil command to view the contents of](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=230.98976923076933) [this file. Gsutil is a command line](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=233.4103846153846) [utility that you can use to work with](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=235.55830769230766) [cloud storage buckets. Here we are going](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=238.56266666666662) [to view the output of our MapReduce here,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=240.95366666666663) [and you can see the word counts associated](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=243.345) [with the words in that itemData. txt. If](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=246.43193333333326) [you're scripting your MapReduce](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=249.4815) [operations, prefer to use the gcloud command line. It's simple and easy to use.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=7&mode=live&start=250.87550000000002)

[Creating a Cluster with Preemptible Instances Using gcloud](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live)

[If you want to add additional compute](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=2.682) [capacity to your Dataproc cluster, and you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=4.396363636363636) [want to do so cheaply, the easiest way to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=6.171600000000001) [do this is using preemptible VMs. In this](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=8.5584) [demo, we'll see how we can use the gcloud](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=10.790153846153844) [command line utility in order to create a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=13.556) [Dataproc cluster, and we'll have this](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=17.3)[Dataproc cluster have preemptible](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=20.13155555555556) [instances. We are on our main project](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=22.02711111111112) [page, and we have cloud shell activated](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=23.99114285714285) [and have set a current project to be](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=25.729642857142842) [spikey-dataproc. You can use gcloud](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=29.385250000000006) [dataproc clusters create](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=32.102666666666664) [spikey-cluster-two to create a brand-new](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=33.77566666666667) [cluster named spikey-cluster-two. The](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=37.0865)[num-preemptible-workers flag allows you to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=39.634) [indicate how many preemptible instances](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=41.39400000000002) [you want for this cluster. We'll choose](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=43.28485714285715) [just one here. We want this cluster to be](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=46.406571428571425) [created in asia-southeast1-a. We'll use](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=51.06214285714284) [the n1-standard-1 predefined machine type](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=53.309400000000004) [for our master node, and our](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=55.04790000000002) [master-boot-disk is 10 GB in size. We'll](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=57.81433333333334) [have two permanent worker nodes in](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=61.54172727272728) [addition to our preemptible worker node,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=63.226090909090935) [and the machine type here is once again](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=65.07866666666668) [n1-standard-1 for these worker nodes. And,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=67.99733333333337) [finally, the worker-boot-disk size is also](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=70.47033333333334) [10 GB. Hit Enter and wait for a minute or](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=73.08533333333335) [so for our cluster nodes to come up. We'll](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=76.17700000000002) [go to the navigation menu and go to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=78.75392857142857) [Dataproc to view our clusters. We already](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=81.60935714285712) [have spikey-cluster-one, and here is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=84.54480000000001) [spikey-cluster-two being created. Click](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=87.42180000000002) [through to the spikey-cluster-two Dataproc](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=89.51989999999999) [cluster, and here on the Configuration](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=91.37939999999998) [page, you can see the current](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=93.65833333333335) [configuration of your cluster. You can see](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=96.03233333333337) [that the preemptible worker nodes are](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=98.48549999999999) [equal to 1. And if you click on the VM](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=100.22649999999997) [Instances tab, you can see that along with](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=102.82273333333332) [the one master and the two permanent](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=104.795) [workers, we also have a preemptible](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=107.32949999999998) [worker. And our preemptible worker has](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=109.50192857142854)[already been preempted. You can see that](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=111.76587500000001) [it's currently in the stopped state.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=114.15979999999998) [Preemptible workers are lower cost VM](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=116.039) [instances, but they can be reclaimed at](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=118.13409090909092) [any point in time by GCP, which is why you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=120.434) [never store data in preemptible workers,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=123.20400000000001) [you only use it for compute. Let's restore](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=125.63079999999998) [our cloud shell window here, and we'll see](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=128.8623076923077) [how we can use the gcloud command line to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=130.55153846153848) [edit the configuration settings of our](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=133.40823076923084) [cluster. Here I'm going to set the number](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=135.63146153846168) [of preemptible workers to 0. I'm going to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=138.09383333333335) [get rid of the one preemptible worker that](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=140.52635714285717) [our cluster has. On our Cluster details](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=143.02864285714296) [page, our UI hasn't yet updated. If you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=145.25619999999998) [click on the Refresh button, you'll find](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=147.68766666666664) [that our preemptible instance has](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=149.4294999999999) [disappeared. We only have the one master](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=150.95892307692304) [and two workers now. I can also use the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=153.69915384615373) [gcloud dataproc command in order to delete](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=156.87600000000003) [this entire cluster. I'm going to delete](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=159.1314000000001) [spikey-cluster-two. It shows me a warning.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=161.53800000000007) [Yes, I do want to go ahead and delete it.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=163.698) [Click on Yes here, and then your cluster](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=166.53799999999993) [will be deleted. If you hit Refresh on](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=168.8456923076923)[your Dataproc Clusters page, we'll find](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=171.1890769230769) [that we are left with just the one cluster](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=173.04546153846158) [now, spikey-cluster-one. Cluster two has been deleted.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=8&mode=live&start=176.547)

[Monitoring Clusters Using Stackdriver](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live)

[We have our Dataproc cluster, and we know](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=1.47) [how to run MapReduce jobs on this cluster.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=4.3165) [Let's see how we can monitor the cluster](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=7.163) [and its metrics using Stackdriver.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=10.350076923076928) [Stackdriver monitoring collects metrics,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=12.342) [events, and other metadata from your cloud](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=14.481499999999999) [platform resources and allows you to view](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=17.544454545454542)[this information graphically and configure](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=19.69918181818181) [alerts based on predefined thresholds. In](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=22.264777777777777) [your navigation menu, go to Monitoring](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=25.646) [under Stackdriver. Wait for Stackdriver to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=29.000000000000007) [instantiate for your current project.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=31.77844444444445) [Choose the project, here I'm choosing](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=33.854) [spikey-dataproc, and create a new](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=36.18079999999999) [Stackdriver account. This is needed if](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=38.0406) [you're using Stackdriver for the very](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=39.89219999999999) [first time on this project. My current](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=42.12425) [project here, spikey-dataproc, has been](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=45.17216666666668)[selected. I'm going to click on Continue](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=47.1179) [and move on to the next step. Stackdriver](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=49.626) [monitoring also works with AWS resources,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=52.64733333333336) [so you can add an AWS account and have](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=54.86016666666667) [Stackdriver monitor that as well. I'm](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=58.01991666666669) [going to go ahead and skip AWS and move on](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=59.78733333333335) [to the next step. If you want to collect](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=62.62)[additional information in the form of](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=65.60350000000001) [events and metrics from your VM instances](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=67.30985714285715) [that you are to monitor, then you can](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=68.97085714285718) [install Stackdriver agents on your VMs.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=71.77719999999998) [We're going to skip this step as of now.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=74.563) [I'm going to hit Continue and move on. I'm](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=76.31542857142855) [not really interested in Stackdriver](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=79.07799999999999) [reports at this point in time. I'm going](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=81.31424999999997) [to hit Continue and move on. Wait for a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=84.45645454545458) [bit while Stackdriver configures](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=87.61763636363638) [monitoring for your project. You can now](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=88.95181818181821) [click on the Launch monitoring button.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=91.465) [This takes us to the main monitoring](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=93.823) [dashboard for our spikey-dataproc project.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=96.58041666666668) [You can explore this dashboard and see](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=98.55) [what all is possible to monitor using](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=101.0513333333333) [Stackdriver. For our case, though, we're](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=103.55266666666661) [going to create a new dashboard to monitor](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=105.3553846153846) [our Dataproc cluster. Our dashboard is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=107.668) [empty at this point in time. Click on the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=110.30350000000001) [Add Chart button to the top right to add a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=113.72949999999997) [new chart that you'll used to monitor your](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=116.36449999999991) [cluster. We'll call this the Log\_Chart.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=119.1322727272727) [We'll use it to track logging information.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=121.505) [You can choose the resource type and the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=123.785) [metric you want to track for that](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=126.10799999999998) [resource. I'm going to choose Cloud](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=128.14062499999997) [Dataproc Cluster here. There are a variety](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=130.04912499999998) [of other options available. You can use](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=132.42200000000005) [Stackdriver to track metrics on your](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=134.7946363636364) [storage buckets, VM instances, any GCP](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=136.73590909090922) [resource. Once you've chosen the Dataproc](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=139.35580000000002) [cluster as an option, here are all of the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=142.315)[metrics that you can track within your](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=144.58749999999998) [cluster, the HDFS capacity, the DataNodes,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=146.701) [storage utilization, log bytes, log](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=149.254) [entries. There are a whole host of metrics](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=152.14233333333328) [available here. For our demo here, we'll](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=155.50444444444443) [choose to track the number of user-defined](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=158.04250000000002) [log entries. So that's our option. Once](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=160.48200000000003) [you've made your choice, you can see that](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=162.5270769230769) [the chart for this is immediately](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=164.684) [available for preview on the right. In](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=167.06050000000002) [addition, I'm going to now filter this by](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=169.7649090909091) [cluster name. I only want to track my](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=172.3881818181818) [spikey-cluster-one, not other clusters](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=175.7341428571428) [that I might have. Click on Apply, and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=177.52757142857135) [notice how the graph on the right changes](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=180.36700000000005) [now to track just spikey-cluster-one.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=182.45900000000015) [There are other options available here for](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=184.60600000000002) [grouping, as well as aggregation. I'm just](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=188.1111666666667) [going to go with aggregation none for now.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=189.91133333333343) [I want to count the number of log entries,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=192.23353846153844) [I don't want other aggregations to be](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=195.144) [performed. Go ahead and hit Save. And here](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=197.08825) [is our first Log\_Chart. The spike in log](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=200.3810909090909) [counts that you see at the center here](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=203.09800000000004) [corresponds to when we ran our MapReduce](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=205.0860000000001) [jobs. You can get additional details on](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=207.22415384615385) [the log count if you click on the icon](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=209.52176923076925) [here at the top right. And here are the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=212.10538461538462) [individual log counts for info logs, name](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=214.83930769230767) [node logs, and so on. You can add multiple](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=217.15684615384612) [charts to your dashboard to track all of](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=220.03150000000002) [the metrics that you find interesting. I'm](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=222.58814285714286) [going to now track some HDFS-specific](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=224.95862499999998) [metrics. The HDFS\_Chart is going to track](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=226.79237499999994) [the Cloud Dataproc Cluster as before. And](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=229.90699999999998) [this time I'm going to track HDFS storage](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=233.2228)[utilization. We'll track HDFS storage](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=237.2772) [utilization for spikey-cluster-one. We](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=240.01828571428575) [want to ensure that our developers don't](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=242.0187692307692) [use HDFS beyond a point. They should be](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=244.29215384615375) [using cloud storage. This time around,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=247.00846153846152) [we'll perform an aggregation. We want to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=249.19338461538456) [track average storage utilization. We'll](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=252.01285714285711) [use the mean aggregation. Go ahead and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=254.12872727272728) [click on Save, and we have our second](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=256.21981818181814) [chart. Clicking on the details icon at the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=258.82278571428566) [top right of this chart will show us that](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=261.55364285714273)[our storage\_utilization is very low. So](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=264.7134615384616) [that's good. Once you have your custom](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=266.81483333333335) [dashboard, you can rename this dashboard](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=269.1936666666667) [to be something meaningful,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=271.601) [Spikey-Dataproc Dashboard. And you can](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=273.697) [view information for different periods of](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=275.63929999999993) [time, for example, for 6 hours, 1 day,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=277.4278999999998) [1-hour periods, and so on. Stackdriver](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=281.1454999999999) [monitoring makes a host of cluster metrics available to you out of the box.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=9&mode=live&start=283.94300000000004)

[Stackdriver Monitoring Groups and Alerting Policies](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live)

[Stackdriver monitoring has a feature](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=1.901) [called groups that allows you to monitor a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=3.4528181818181816) [set of resources as a single entity. In](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=6.151666666666666) [this demo, we'll see how we can set up a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=9.327857142857143) [monitoring group to monitor our cluster](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=11.80642857142857) [resources. We want to be able to collect](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=13.500727272727273) [additional metrics and events from our](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=16.036) [cluster nodes. And for this, we'll set up](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=18.213249999999995) [a Stackdriver monitoring agent using](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=20.78972727272727) [initialization actions on our nodes. We](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=22.332) [start off in the main Monitoring Overview](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=26.343444444444444) [page. This shows us our Spikey-Dataproc](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=28.53055555555556) [Dashboard that we set up earlier. On the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=30.64709090909091) [left navigation pane, click on the Groups](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=33.42892307692307) [link, which will take us to monitoring](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=35.58815384615382) [groups. Here is where we'll set up a new](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=37.71415384615385) [group. Click on the Create Group option.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=40.39061538461539) [The cool thing about using monitoring](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=42.669090909090926)[groups to monitor a group of resources is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=44.62233333333333) [the fact that this group is dynamic. You](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=47.041000000000025) [can set up membership criteria so that you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=49.50069230769232) [know which resources will belong to this](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=51.985) [group at any point in time. Based on](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=54.34292307692307) [whether they meet this membership](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=56.961999999999996) [criteria, resources can fall in and out of](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=58.456999999999994) [this particular group. Once I get to the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=61.98439999999999) [Create group page, go ahead and give this](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=64.45456250000001) [group a name. I'm going to call it the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=66.19306250000005) [spikey-clustergroup. And you can specify](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=69.75242857142857) [the criteria to determine membership](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=71.65233333333336) [within this group. The default filter](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=73.42150000000007) [condition allows us to monitor any](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=75.59409090909091) [resource that matches any of the rules](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=77.81627272727272) [that we're going to set up here. Here our](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=79.81491666666665) [very first rule is that the name contains](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=82.638) [spikey-cluster-two. You can modify this](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=86.88600000000005) [rule. Instead of Contains, you can have](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=88.40733333333336) [other options configured as well, such as](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=90.14041666666672) [Does Not Contain, Equals, Starts With,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=92.77720000000002) [Ends With, and so on. There's a special](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=95.02830769230769)[checkbox here to allow you to customize](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=97.66153846153844) [your monitoring when you're dealing with](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=99.84754545454547) [nodes that belong to the same cluster. If](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=101.64536363636367) [there is one or more nodes in this cluster](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=104.07549999999999) [which differ significantly in performance](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=106.94355555555556) [from others, that will be flagged. Go](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=108.78744444444446) [ahead and save this group. We've created a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=111.297) [new monitoring group. A very useful](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=114.06063636363638) [feature of monitoring is the setting up of](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=116.04390909090912) [alerting policies. Alerting policies can](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=119.13350000000003) [highlight significant changes in your](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=121.40669999999999) [resource performance. We are going to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=123.84619999999997) [create a new alerting policy, which will](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=125.74323076923076) [be triggered under specific conditions.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=127.3553846153846) [Click on the Add Condition button here,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=129.259) [and this will bring up a page that will](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=132.626) [allow you to configure the condition under](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=134.3970714285714) [which your alert will be triggered. There](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=136.27775)[are two broad categories based on which](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=139.30984615384614) [your alert can be triggered. Either a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=141.70276923076918) [metric has crossed a particular threshold,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=144.1797142857143) [or a metric is entirely absent. We're](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=146.48285714285723) [going to choose the Metric Threshold](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=149.12627272727272) [alerting policy for now. Go ahead and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=151.16790909090903) [click on the Select button here. The first](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=153.762) [configuration setting is to specify the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=156.7669090909091) [target of this alerting policy, which is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=158.3836363636364) [the resource that you want to monitor.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=160.47090909090912) [There are a number of resources you can](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=162.709)[choose from. We want to specifically](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=164.52580000000006) [monitor the consumed API, how many API](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=166.681) [calls are made from our Dataproc cluster.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=169.65775000000005) [If the number of requests made by the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=172.36253846153838) [processes running on our Dataproc cluster](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=175.38299999999998) [exceed a certain threshold within a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=177.54899999999992) [certain time period, we want our](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=179.9560909090909) [engineering team to be alerted. So go](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=182.6042727272727) [ahead and choose Consumed API. Now let's](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=185.2465) [set up our metric. If a metric goes beyond](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=187.648) [a certain condition here, if request count](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=190.53699999999998) [goes beyond a certain condition here,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=193.3003636363636) [we'll choose the condition to be above.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=195.846) [And here, the threshold we set to 1000](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=197.9506666666667) [requests. So if the number of requests](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=200.4001538461538) [made by the processes on our cluster go](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=202.60784615384603) [beyond 1000 requests per minute, then we](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=205.9425) [want to be alerted. Accept the default](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=208.9385714285714) [values for all of the other configuration](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=211.06507142857137) [settings. If we go down to the PROJECT\_ID](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=213.3307272727273) [field, we want to explicitly monitor](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=216.3176363636365) [resources in the spikey-dataproc project,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=218.8904444444445) [and that's what we're going to choose](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=220.999) [here. We can also specify the service](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=223.44199999999992) [where we want to track the consumed APIs.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=225.72000000000008)[We want to track APIs at dataproc.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=228.292) [googleapis. com. That is our choice here.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=232.06966666666673) [Once you've set up your configuration,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=234.82490909090905) [you'll notice that a visualization](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=236.836) [tracking this metric pops up here at the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=238.261) [bottom. Go ahead and click on Save](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=240.66814285714287) [Condition, and let's go on with setting up](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=243.1081428571429) [our alerting policy. The next step is to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=245.8151538461538) [specify how we want to be alerted when](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=248.38623076923062) [this threshold is exceeded. Under](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=251.14160000000004) [Notifications, go ahead and click on the +](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=252.81484615384616)[button here that will allow us to add a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=254.93361538461536) [method of notification. We want to be](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=257.4530000000001) [notified by email to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=259.4970000000003) [spikeysales@loonycorn. com. And, finally,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=261.35255555555557) [we'll give this alerting policy a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=263.89566666666667) [meaningful name so that we can identify](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=266.90464285714285) [what exactly it tracks. This is the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=269.4751428571428) [APIConsumption\_AlertPolicy. Click on the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=272.4665)[Save Policy button. We've created an](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=274.49649999999997) [alerting policy for our cluster group. You](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=277.46099999999996) [can always click on the Groups link on](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=280.5936153846154) [your left navigation pane in order to get](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=282.58253846153855) [quick access to your monitoring groups.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=284.92820000000006) [Under Groups, click on Groups Overview,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=286.955) [and this will take you to a dashboard that](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=288.569) [will list all of the monitoring groups](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=291.0801666666666) [that you have. Here is our](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=293.27874999999983) [spikey-clustergroup. Click through to this](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=295.2833333333334) [group, and you'll notice something. There](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=297.0538888888891) [are no matching criteria, no events are](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=299.84077777777765) [being monitored. This is because when we](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=302.732923076923) [set up this monitoring group, our](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=304.9400769230767) [filtering criteria basically said we want](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=307.1661)[to match those resources which have the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=309.72629999999987) [name spikey-cluster-two. We just have](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=313.15150000000006) [spikey-cluster-one. There is no resource](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=315.37139999999994) [which matches the membership criteria for](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=317.4053999999998) [this group. There is nothing to monitor here. Let's fix that in the next clip.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=10&mode=live&start=319.71199999999993)

[Configuring Initialization Actions for Dataproc](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live)

[In this clip, we're going to create a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=1.414) [cluster called spikey-cluster-two, which](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=4.541200000000002) [we'll then monitor using Stackdriver. For](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=6.298714285714286) [every node in spikey-cluster-two, we want](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=9.0635) [to customize the software that is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=11.016499999999999) [installed in the VMs that make up this](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=12.890230769230769) [cluster so that the Stackdriver monitoring](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=15.282846153846151) [agent is installed. The monitoring agent](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=18.149) [can be installed using commands in a shell](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=20.697909090909093) [script, and we'll have this shell script](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=22.959) [run as an initialization action. So](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=25.660999999999994) [whenever a VM on this cluster comes up,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=27.921666666666667) [this initialization action will be](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=30.566999999999993) [executed. Here are the shell commands that](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=32.182) [we'll use to install the Stackdriver](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=34.36250000000002)[monitoring agent. This can be for any VM](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=37.0142) [instance. We'll apply it to the VM](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=39.819285714285705) [instances that make up our cluster. This](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=41.756285714285696) [command uses the curl utility that all our](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=44.195769230769216) [VMs come preinstalled with to download the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=47.07018181818182) [monitoring agent and then install the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=49.52845454545457) [monitoring agent on the VM. These shell](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=51.932375) [commands are present in the stackdriver.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=55.0524) [sh file. We'll now switch to our GCP web](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=57.660599999999995) [console. We start off in the Dataproc](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=60.77481249999999) [Clusters page. We have the one cluster](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=63.01962499999998) [here, spikey-cluster-one. Click on the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=65.79528571428573) [Activate Cloud Shell icon to the top right](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=67.27379999999998) [so that we can work with the terminal](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=68.93993333333326) [window. We set the current project on this](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=71.15266666666665) [cloud shell session to be spikey-dataproc.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=73.912) [We'll use the gcloud command line utility](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=77.808) [to create a new Dataproc cluster. We'll](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=80.11692307692304)[call this cluster spikey-cluster-two.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=82.653) [Remember, this cluster will match the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=84.881) [membership criteria of our monitoring](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=87.47550000000001) [group. This cluster will be created within](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=89.63758333333335) [the spikey-dataproc project. That's where](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=92.58333333333331) [we're doing all of our work. And we have](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=94.22650000000002) [specified initialization-actions for all](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=96.55900000000001) [of the VMs on this cluster. We want to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=98.12300000000003) [customize those VMs to install certain](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=101.7533) [software. The GCP has set up a public](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=104.3219) [cloud storage bucket with commonly used](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=107.05353846153848)[initialization actions, and installing the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=109.028) [monitoring agent is one such script file.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=111.14577777777777) [We'll access this script file directly](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=113.81909090909092) [from the dataproc-initialization-actions](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=115.77727272727275) [bucket. We need to enable certain OAuth](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=117.22785714285715) [scopes on the cluster machines so that](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=122.4533) [Stackdriver monitoring allows publishing](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=125.52139999999999) [metric data to the GCP projects. The scope](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=127.41925) [that we want here is the monitoring. write](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=131.01115384615383) [scope that will enable our monitoring](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=133.12376923076914) [agent to write data to our GCP project.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=135.44245454545455) [The other configuration parameters that](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=139.513) [you see here relate to the cluster](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=140.87981818181814) [creation itself. And you are already](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=142.86360000000002) [familiar with these. Hit Enter, and go](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=144.92520000000007) [ahead and create this cluster in the zone](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=147.3482666666666) [of your choice. I'm going to create this](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=150.13279999999983) [cluster in the asia-southeast1-a zone.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=153.19500000000005) [Creating this cluster might take a minute](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=155.213) [or so. Observe the progress of your](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=157.49422222222228) [cluster creation on the UI, and wait for](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=159.50715384615378) [your cluster to be up and running. We now](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=161.685) [have spikey-cluster-two. Now let's go back](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=165.03299999999987) [to our Stackdriver monitoring page. And if](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=167.23655555555553) [you hit Refresh on this](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=169.74316666666667) [spikey-clustergroup, you'll see that we](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=171.44858333333332)[now have monitoring enabled for all of our](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=173.154) [nodes in spikey-cluster-two. The](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=177.07763636363643) [monitoring agent that we set up on the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=178.8) [cluster thanks to our initialization](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=180.80800000000005) [action made these metrics available to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=182.53287500000002) [Stackdriver. You can click through to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=186.85812500000006) [individual cluster nodes from the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=188.94945454545459) [Stackdriver monitoring group UI, and you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=190.31990909090916) [can view the API requests made by this](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=192.48707692307693) [particular node to your Google Dataproc](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=195.516) [APIs. These metrics are only available](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=199.4922857142857) [because we have the monitoring agent](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=201.7272727272727) [installed via our initialization action.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=203.614) [And with this, we come to the very end of](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=207.021) [this hands-on module. We saw how we could](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=209.97869230769223) [work with managed Hadoop on the Google](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=212.57850000000002) [cloud using Dataproc clusters. We started](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=214.97600000000003) [off by creating a cluster using the web](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=216.93072727272727) [console. We also saw how we could create](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=219.26963636363632) [and manage clusters using the gcloud](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=221.97538461538463) [command line utility. We saw how we could](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=224.044) [submit MapReduce jobs to the cluster by](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=226.88866666666664) [SSHing into the master node, as well as](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=229.1846923076923) [using the gcloud command line. We](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=231.51453846153842) [connected to cloud storage buckets to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=233.59336363636365) [access data, and we worked with](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=235.8379090909091) [preemptible workers. In addition, we](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=238.78057142857142) [learned how we can use Stackdriver](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=240.70250000000001) [monitoring in order to monitor our cluster](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=242.49150000000006) [resources. We also saw how we could](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=244.85866666666666) [customize our Dataproc cluster nodes with](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=247.43583333333333) [specific software using initialization](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=250.76049999999995) [actions. In the next module, we'll see how](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=253.1256999999999)[we can create and run Spark applications on our Dataproc cluster.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97183a94-07c8-4a11-9c37-8c9d3a4a7581&clip=11&mode=live&start=255.23546153846158)

[Working with Apache Spark on Google Dataproc](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=0&mode=live)

[Module Overview](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=0&mode=live)

[Hi, and welcome to this module where we'll](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=0&mode=live&start=0) [see how we can work with Apache Spark on](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=0&mode=live&start=4.008) [Google's Dataproc cluster. Apache Spark is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=0&mode=live&start=6.631) [a great general purpose computing engine](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=0&mode=live&start=10.358714285714285) [that is built on top of basic Hadoop](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=0&mode=live&start=12.2075) [framework. Apache Spark uses HDFS, YARN](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=0&mode=live&start=14.899666666666668) [under the hood. It does not use MapReduce,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=0&mode=live&start=18.012624999999996) [though it has its own analytics](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=0&mode=live&start=20.68475) [computation engine that allows for big](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=0&mode=live&start=22.759500000000003) [data processing. Apache Spark is simple](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=0&mode=live&start=25.598800000000004) [and easy to use. It's very intuitive](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=0&mode=live&start=27.952499999999993) [because it completely abstracts away the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=0&mode=live&start=30.293999999999997) [concept of multiple machines on the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=0&mode=live&start=32.283) [cluster from the developer, unlike](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=0&mode=live&start=34.45985714285715) [MapReduce operations. Dataproc on the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=0&mode=live&start=36.801) [Google Cloud supports Spark using the YARN](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=0&mode=live&start=39.83614285714285) [cluster manager. Spark applications that](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=0&mode=live&start=43.193) [you might have been running on your](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=0&mode=live&start=44.96314285714285) [on-premises datacenter can be lifted and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=0&mode=live&start=47.287499999999994) [shifted to the cloud very easily. There is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=0&mode=live&start=49.591333333333324) [built-in support for PySpark, that is a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=0&mode=live&start=52.42533333333333) [Python shell for Spark, as well as Spark](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=0&mode=live&start=54.89485714285714) [Scala. In addition, GCP provides](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=0&mode=live&start=57.791999999999994) [connectors so you can connect to BigQuery,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=0&mode=live&start=59.678) [Cloud Storage, and other GCP technologies](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=0&mode=live&start=62.70933333333334) [in order to access your data from within Spark applications.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=0&mode=live&start=64.94257142857144)

[Spark for Distributed Processing](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live)

[Before we move on to using Spark on](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=2.83) [Dataproc, let's briefly talk about Spark](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=4.466) [as a data analytics engine and understand](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=6.7211428571428575) [why it's so popular. Spark is one of a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=9.508142857142857) [suite of technologies that is built on top](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=12.006888888888884) [of the Hadoop ecosystem, but it's by far](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=13.689222222222226) [the most popular and very widely used. One](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=16.605249999999998) [important reason for this is the fact that](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=20.216124999999998) [Spark makes it very easy to develop and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=22.16225) [prototype big data applications on Hadoop.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=25.253) [This is a distributed computing engine](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=28.541) [that is built on top of Hadoop, but it has](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=30.15499999999999) [an interactive shell that allows you to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=32.635299999999994) [quickly process datasets. You can develop](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=34.691) [your big data processing applications in a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=37.71357142857143) [programming language of your choice using](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=39.982499999999995) [something familiar such as the Jupyter](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=42.343500000000006) [iPython notebook for Python code, as well](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=44.690999999999995) [as Scala. In addition to a distributed](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=47.33440000000001) [processing framework, Spark offers a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=49.38649999999999) [number of built-in libraries in different](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=51.29166666666668) [languages for very common use cases, such](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=54.17340000000001) [as machine learning, processing streaming](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=56.33274999999999) [data, graph processing, and so on. These](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=59.053749999999994) [built-in libraries make it very easy to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=62.195)[get up and running with your application.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=64.4285) [We've already spoken of the three main](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=66.875) [components that make up the Hadoop](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=68.485) [distributed processing framework. We have](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=70.029) [HDFS, which is a distributed file system](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=72.4195) [to manage data storage. We have MapReduce,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=74.504) [which is the framework to define a data](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=77.07042857142858) [processing task. And we have YARN, which](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=80.08033333333331) [is a cluster manager or a framework, which](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=82.32287499999998) [executes the data processing task that](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=85.18971428571432) [we've defined. Now Apache Spark is built](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=87.82933333333335) [on top of YARN and HDFS. The Spark Core](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=90.34200000000001) [engine can be thought of as a replacement](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=93.82625) [for MapReduce. Spark Core is what takes](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=96.89399999999998) [care of general purpose computing on a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=99.28775000000002) [distributed platform. Spark Code can use](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=102.43266666666668) [YARN as its cluster manager to allocate](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=104.71949999999998) [resources across the different processing](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=106.95999999999998) [tasks. It can use other cluster managers](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=109.30533333333334) [as well, such as Mesos or Spark](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=111.43975) [Standalone, which runs on a single node.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=114.30749999999999)[When you run Spark on top of Hadoop, it](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=116.409) [continues to use HDFS in order to store](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=119.16577777777778) [data in a distributed manner. Where Spark](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=121.748) [really differentiates itself and wins over](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=124.886) [basic Hadoop, though, is in its libraries.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=127.30775) [There are libraries to perform SQL](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=129.572) [transformations on your data. There are](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=131.68633333333332) [streaming libraries, machine learning](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=134.312) [libraries, and libraries for graph](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=135.9342) [manipulation, for social network graphs](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=138.14633333333333) [for example. If you talk to a developer or](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=140.085375) [data scientist today, chances are that](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=142.77439999999999) [they use Spark for all of their big data](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=144.84233333333333) [processing requirements. So what does](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=147.80639999999997) [Spark have to offer? Spark gives us](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=149.64128571428574)[real-time, as well as batch processing of](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=151.69362500000003) [data. Hadoop is mainly for batch](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=153.95457142857143) [processing. Spark also gives us an](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=155.8764285714286) [interactive read/evaluation/print/loop](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=158.159)[environment, a REPL environment as it's](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=161.061) [called. This environment gives us fast](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=163.6718571428572) [feedback allowing us to build prototypes](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=165.92866666666663) [very quickly. And, lastly, Spark supports](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=168.29800000000006) [a variety of common programming languages.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=170.53771428571426) [Python, Java, Scala, and R are among the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=173.303) [few. Hadoop MapReduce operations always](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=176.96788888888898) [act as data in the form of files. These](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=179.108) [files can be stored on cloud storage in](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=181.796125) [the case of Dataproc, or they'll be on](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=183.87539999999998) [HDFS. All operations in Spark are](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=186.55859999999993) [performed not on files but on in-memory](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=188.26828571428572) [objects called RDDs or resilient](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=191.20871428571422) [distributed datasets. Working on in-memory](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=194.31179999999998) [objects is what allows Spark to be so](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=196.67085714285716) [performant and efficient. An RDD's not a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=199.60742857142858) [single object. Instead, it can be thought](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=202.07666666666665) [of as a collection of entities. These](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=203.53024999999997) [entities can be thought of as rows and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=205.45533333333336) [records of the RDD, and these rows and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=207.65766666666661) [records are partitioned and spread across](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=210.56885714285718) [multiple machines on your cluster. All of](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=213.359) [these are in the memory of your cluster](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=216.55144444444446) [machines, not on disk. One of the most](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=218.545) [important characteristics that you have to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=221.6276666666667) [remember about RDDs is the fact that it's](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=223.28611111111115) [partitioned or split across data nodes in](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=225.60142857142858) [the cluster. This is what allows for](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=228.0144) [parallel processing on multiple nodes.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=230.33757142857147)[RDDs are immutable. Once created, they](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=232.80800000000002) [cannot be changed. You can only apply](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=235.55599999999998) [transformations to an RDD and create a new](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=237.60014285714283) [RDD with new information. And, finally,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=241.09566666666666) [RDDs are resilient. That's why they are](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=244.1278) [called resilient distributed dataset. If](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=246.44533333333334) [the node on which a portion of our RDD](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=247.8964285714286) [lives crashes, the RDD can be](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=250.71333333333334) [reconstructed from scratch. It has a lineage; it knows where it came from.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=1&mode=live&start=252.76160000000002)

[Running a Spark Scala Job Using the Web Console](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live)

[In this demo, we'll see how we can submit](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=2.809) [a Spark application to the Dataproc](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=4.812399999999999) [cluster using the web console. This Spark](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=7.186666666666667) [application will be built using Scala. We](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=9.729) [start off in the Dataproc cluster state,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=13.479875) [and here is the spikey-cluster-one that we](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=15.74) [created earlier. Click through, and let's](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=17.625) [take a look at the VM instances in the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=20.308) [cluster. What we want to do here is to SSH](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=22.586799999999997) [into our master node and create a new](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=25.142363636363633) [Scala Spark application. A master node](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=27.967428571428567) [comes preinstalled with Scala. If you type](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=30.850500000000007) [Scala on our terminal window here, we'll](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=33.54449999999999) [get into the Scala interactive shell.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=35.653000000000006) [Let's define a very simple Hello, world!](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=38.551) [Scala object here, which prints Hello,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=41.042) [world! to screen. This will be our Scala](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=43.134) [application. Let's save this object in the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=47.025555555555556) [Scala file. We'll call it HelloWorld.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=49.19725000000001) [scala, and then go ahead and quit this](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=51.16600000000001) [Scala interactive shell by hitting :q.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=54.28499999999999) [Compile the Scala file by calling scalac](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=57.053625000000004)[HelloWorld. scala, and if you run an ls](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=60.484857142857145) [command here, you'll see that there are.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=63.70049999999999) [class files available here in the current](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=65.92928571428571) [working directory. We know that Java is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=69.2945) [already installed on our master node here.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=71.77649999999997) [Let's use the jar command in order to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=73.65249999999997) [create HelloWorld. jar. Once our Java](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=77.40942857142858) [archive has been created, we can now copy](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=80.35757142857145) [this JAR file over to our spikey-data](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=82.6358) [bucket. We copy the JAR file over using](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=84.59914285714285) [the gsutil cp command. We can now head](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=88.71866666666666) [back over to our GCP web console, use the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=91.768) [hamburger icon, go to the navigation menu,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=94.77485714285716) [and let's take a look at our storage](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=97.15900000000002) [buckets. If you look within the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=100.40528571428575) [spikey-data bucket that we have here, you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=102.30699999999999) [can see that HelloWorld. jar is present.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=103.90375000000002) [Let's submit this Spark Scala application](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=107.856) [to our Dataproc cluster. Go to Dataproc,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=110.05) [Jobs, and here you'll find a link to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=112.51562499999999) [submit a job from the web console. Click](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=115.4336) [on Submit job, and it'll take you to a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=118.13722222222223)[place that will allow you to configure](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=120.05137500000001) [your job settings. Specify a unique ID for](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=121.678) [your job. We are going to use the global](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=124.503) [endpoint, and we are going to submit this](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=126.88) [job to spikey-cluster-one. There is a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=129.171) [drop-down here that allows you to specify](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=132.57562500000003) [the kind of job you want to run on your](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=134.26500000000001) [cluster. You can choose from Hadoop jobs,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=136.69214285714284) [Spark jobs, PySpark jobs, and so on. For a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=138.92044444444446) [Scala Spark application, the job type](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=143.32925) [should be Spark. For Spark job, you need](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=145.421) [to specify the JAR file or the main class](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=149.239) [where your Spark application can be](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=152.419) [invoked. Point to the HelloWorld. jar](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=154.92357142857136) [that's present in your spikey-data bucket.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=156.87599999999995) [If you have other command line arguments](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=160.035) [that you need to pass into your JAR file,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=161.834) [you can specify those here as well. We](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=163.833)[have nothing, so we can leave all of these](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=166.49833333333333) [fields empty. Scroll down and click on the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=168.6065714285714) [Submit button, and you will have](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=171.7325) [successfully submitted your Scala Spark](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=173.162) [application to your Dataproc cluster. If](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=175.774) [you go over to the page that lists all of](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=179.96710000000002) [your jobs, you can see that your](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=181.664375) [sparkScala job is just starting up. You](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=182.892) [can monitor the status of your job here.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=186.43655555555554) [If you're interested in viewing logs, you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=188.593) [can just click through to the job, and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=190.844) [you'll see your log files right there.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=192.903125) [There's just one log file entry here.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=195.067) [Hello, world! has been printed to screen. A job output is complete.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=2&mode=live&start=197.647)

[Executing a Spark Application Using gcloud](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live)

[In this demo, let's see how we can run a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=2.052) [simple Spark application on the Dataproc](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=3.806) [cluster. We'll submit this job to the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=6.765714285714286) [cluster using the gcloud command line](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=8.753000000000002) [utility. We start off in the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=11.646200000000002) [spikey-dataproc project, and we are in the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=13.659166666666664) [Dataproc Clusters page, which lists all of](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=15.681142857142856) [the clusters that we have up and running.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=17.51042857142858) [We just have the single cluster,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=19.875) [spikey-cluster-one, at this point in time.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=21.867) [We've deleted any additional clusters that](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=23.806) [we had created earlier. We'll run a Spark](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=25.712) [Python application, which is available as](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=29.44) [a part of the examples that GCP provides](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=30.875000000000004) [to us in order to test out Dataproc. The](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=32.945499999999996) [Python code for this Spark application is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=36.123125) [available in the Dataproc staging bucket](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=38.30628571428571) [that was created when we set up our](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=40.557500000000005) [spikey-cluster-one. I'm going to use](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=44.132499999999986) [gsutil here and list the contents of this](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=45.59077777777779) [bucket. And we'll find the right Python](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=48.04749999999999) [script that we want to run. Within the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=50.11550000000001)[source folder here on our staging bucket,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=52.88925000000002) [there is a folder named PySpark, which](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=54.589500000000015) [contains a few Python examples. The](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=56.59442857142858) [specific example that we're looking for is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=58.96249999999999) [the hello-world. py. This is the Spark](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=62.20171428571428) [application that we'll execute on our](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=64.65342857142858) [Dataproc cluster. As its name implies,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=67.5546) [this is a very simple Hello, world! Spark](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=70.05455555555558) [application. I'm going to call gsutil cat](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=73.15700000000001) [in order to list out the contents of this](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=75.14600000000003) [Python file. You can see the Python code](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=77.96224999999998) [right here onscreen. We import the PySpark](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=80.368) [library, set up a SparkContext,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=82.3915) [parallelize a new RDD with the words hello](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=84.11566666666667) [and world, and then print it out to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=87.77499999999998) [screen. You can use the gcloud dataproc](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=90.27766666666668) [jobs command in order to submit a PySpark](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=92.629875) [job. We want to submit PySpark, and we](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=95.6982857142857) [want to submit this job to the one cluster](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=99.2094) [that we have up and running, that is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=101.49271428571429) [spikey-cluster-one. You'll also need to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=104.60033333333332) [specify the Spark Python program that we](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=106.30385714285714)[want to execute on our Dataproc cluster.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=108.15249999999997) [We simply point to hello-world. py on our](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=110.46571428571427) [staging bucket. When you're executing this](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=114.46) [code on your local machine, remember that](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=116.47085714285713) [your staging bucket will have a slightly](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=118.67866666666666) [different name. You need to copy over the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=121.30928571428569) [right name for this to work. Our job has](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=123.13099999999997)[been successfully submitted to the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=125.77214285714285) [Dataproc cluster. We can head over to the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=127.81433333333332) [web console and click on the Jobs link to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=129.9563333333333) [see the status of its progress. You can](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=132.07039999999995) [see onscreen here that our PySpark job is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=134.79600000000005) [currently running. It has been running for](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=138.2477142857143) [about 15 seconds. Wait for a little bit](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=140.52587500000004) [for this application to run through to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=142.70837499999996) [completion. You can see here onscreen](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=144.75400000000005) [Hello and world! have been printed out. Our Spark job has executed successfully.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=3&mode=live&start=146.60912500000003)

[Creating a BigQuery Table](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live)

[In this demo, we'll build our own Python](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=2.819) [Spark application, which shows how we can](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=4.74) [connect to other GCP services in order to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=7.161) [access data. We'll connect to BigQuery, as](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=10.01) [well as cloud storage from within Spark.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=12.702125) [We'll start off by creating a new BigQuery](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=15.643) [table where we'll store our input data](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=17.884)[that we want to process using a Spark](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=19.86325) [application. BigQuery is Google's data](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=22.849750000000004) [warehouse, and it's comparable to Redshift](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=25.192800000000002) [on Amazon. One big advantage of BigQuery,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=27.914142857142863) [though, is the fact that it's serverless.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=30.52033333333333) [You don't have to instantiate machines](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=32.366) [that will hold your data. Your data is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=34.00142857142859) [always available to you in a serverless](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=36.57000000000002) [manner. You only pay for data storage only](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=39.21914285714288) [for what you use and for processing.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=41.558999999999976) [Clicking on the BigQuery link on your](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=44.839) [navigation pane will open up the BigQuery](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=46.418) [UI. We are within the spikey-dataproc](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=49.55925000000002) [project. All tables in BigQuery live](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=52.88549999999999) [within a dataset. A dataset can contain](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=55.287250000000014) [one or more tables. Let's create a new](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=57.60150000000001) [dataset where we'll set up our table.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=59.95900000000001) [Click on the Create dataset button here,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=61.951874999999994) [and specify the spikeysales\_dataset. You](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=64.169) [have the option to choose where you want](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=66.7932) [your data to be located in case you have](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=70.08633333333334) [jurisdiction issues. The options are](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=73.14314285714285) [United States, European Union, and Tokyo.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=74.97319999999999) [We'll go with the default option here,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=77.503) [which should be the US. Go ahead and click](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=79.313) [on the Create dataset button. That'll](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=81.209) [create a new BigQuery dataset for you.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=83.8515) [Click on the Go to dataset link here, and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=86.312) [this will show you our spikeysales\_dataset](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=88.57283333333334) [that we just created. We'll now create a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=90.5066) [table within this dataset, and our data](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=94.70957142857145) [will be stored within this BigQuery table.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=96.78849999999997) [Click on the Create table button, and this](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=98.92442857142856) [will open up a UI allowing you to create](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=101.09744444444443) [your first BigQuery table. You can specify](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=103.79450000000003) [where the data in this table comes from.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=106.05475000000001) [You can choose to leave the table empty,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=108.24150000000002) [but we're going to upload a CSV file with](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=110.57887500000001) [our data. So I'm going to choose the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=113.59700000000001) [Upload option here on this menu. I have](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=115.85049999999997) [some sales data available here on my local](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=118.64166666666668) [machine. Click on the Browse button. This](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=121.26916666666668) [will allows you to browse your computer.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=123.15537499999999) [And I'm going to choose salesdata. csv.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=125.538) [This contains CSV records for the sale of](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=128.078) [various products on the Spiky Sales site.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=130.556) [You can indicate to BigQuery the format of](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=134.461) [your data. It's in the CSV format. I'm](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=136.353) [going to choose the CSV option from this](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=139.254) [menu. This table is in the right project](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=141.268375) [and the right dataset. I'm going to now](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=143.6802) [name this table. It'll be called the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=146.22842857142854) [salesdata table. You can click on this](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=148.9463333333333) [checkbox here that allows BigQuery to auto](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=150.78837499999995) [detect the schema for your table. It will](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=153.13157142857145) [set up this schema based on the records in](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=155.71888888888884) [your CSV file. Go ahead and click on the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=158.55999999999995)[Create table button. Wait for a little bit](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=161.30999999999992) [for your data to be uploaded to your newly](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=163.11350000000004) [created table. This is in the form of a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=166.05966666666666) [job, so if you click on](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=168.05072727272724)[spikeysales\_dataset, you'll find that you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=169.72400000000002) [now have this salesdata table within it.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=171.99757142857143) [Observe that the column names from our CSV](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=174.42257142857144) [file have been picked up by BigQuery. That](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=177.6028) [is the schema for our table. There are](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=180.33871428571427) [five columns here, time, location, item,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=182.9476666666667) [cost, and mode. This is the same](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=185.55740000000003) [sale-specific information that we've seen](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=189.11866666666666) [earlier but now in a CSV format. You can](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=190.7691) [preview the table as well, and here is a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=193.23655555555558) [little preview of the data that stored in](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=195.5811111111111) [this BigQuery table. The Spiky Sales](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=197.702) [organization is seriously considering](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=200.23250000000002) [moving its analytics data to BigQuery, and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=201.63633333333334) [having Spark applications process this](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=204.6334) [data easily is an important factor in their decision-making.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=4&mode=live&start=206.97385714285713)

[Pyspark Application Using BiqQuery and Cloud Storage Connectors](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live)

[And here is the Python code for our Spark](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=2.47) [application. This Spark application will](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=4.795) [connect to BigQuery, read in sales data,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=7.544166666666665) [perform some kind of aggregation operation](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=9.946499999999999) [on it, and write the results out to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=12.135333333333335) [BigQuery. In order to work with BigQuery](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=15.47871428571429) [tables, we'll use a BigQuery connector.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=17.473571428571425) [The Spark application will take the data](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=20.112) [from BigQuery, place it in a cloud storage](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=22.347) [bucket, perform its operations, and write](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=24.439) [the final results from cloud storage back](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=27.386) [to BigQuery. In this way, we'll see how](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=30.08) [Spark integrates both with BigQuery, as](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=31.586) [well as cloud storage. We start off by](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=33.98466666666667) [instantiating a SparkContext, which is our](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=36.238) [bridge to the Spark environment. We need](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=39.033) [to point Spark to the bucket created by](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=41.755) [our Hadoop cluster, the default staging](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=44.16028571428572) [bucket. The name of the staging bucket is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=46.294714285714285) [available as a property in the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=48.602500000000006) [hadoopConfiguration file, the fs. gs.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=50.149) [system. bucket, and that's the property](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=53.15575) [that we are accessing here. We then access](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=55.5955) [the current project ID to which the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=58.417375) [cluster belongs. That is our](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=61.0365) [spikey-dataproc project. This is also](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=63.33659999999999) [present in the hadoopConfiguration file of](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=64.98850000000002) [our Dataproc cluster. In order to use the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=67.25600000000001) [BigQuery connector from Spark, you need to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=70.12287500000002) [specify a bunch of information to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=71.88100000000001) [BigQuery, which you do via a config Python](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=73.63725000000001) [dictionary. These are BigQuery-specific](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=77.25733333333332) [configuration settings, all of the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=79.73975)[information that BigQuery needs to connect](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=81.51600000000002) [to the right table within the right](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=83.3172857142857) [dataset in the right project. We connect](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=86.02455555555557) [to BigQuery within our spikey-dataproc](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=88.822) [project, specify the staging bucket and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=90.57600000000001) [the input\_directory within the staging](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=94.0242) [bucket where data will be temporarily](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=95.99657142857141) [stored. When we read in data from](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=97.87942857142858) [BigQuery, we'll temporarily store data in](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=100.13) [the staging bucket. And here is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=102.79679999999999) [information for the BigQuery connector on](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=104.92885714285715) [where the input data is present in the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=106.52028571428572) [salesdata table within](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=109.3945) [spikeysales\_dataset. The output of our](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=110.483) [Spark application will be written out to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=113.736) [the spikeysales\_dataset in the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=116.386) [totalcost\_output table. You can use](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=118.272) [Spark's newAPIHadoopRDD function in order](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=123.85219999999998) [to create an RDD from data that's present](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=126.14725) [in BigQuery. Just specify the right input](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=129.29340000000002) [arguments as you see here onscreen. We](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=131.7378571428571) [specify that the BigQuery input is in a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=133.99137500000006)[JsonTextBigQueryInputFormat. This is a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=136.896) [format that our BigQuery connector on](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=140.99557142857142) [Spark Dataproc understands. BigQuery is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=143.44740000000002) [the input format. Here are classes which](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=146.03799999999998) [represent the output key-value pairs. Keys](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=148.33600000000004) [are LongWritables, values are JsonObjects.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=150.83740000000003) [Make sure you specify a config object that](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=154.319)[you set up earlier for your BigQuery](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=156.244) [connection. And here are some simple Spark](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=159.59612499999992) [transformations on our input data. This](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=161.437) [can be anything based on your use case. We](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=164.23399999999998) [are calculating here the sales per](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=166.9265) [location. We'll write the output](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=170.88950000000008) [temporarily to cloud storage before we](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=172.66199999999998) [move it from cloud storage to BigQuery.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=174.31414285714283) [The output generated by Spark will be](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=176.836) [written to our temporary staging bucket](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=179.714) [within the pyspark\_output folder within](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=182.32580000000002) [this temporary staging bucket. We'll use](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=185.24266666666668) [the sql\_context in Spark to generate](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=188.17714285714283) [output in a JSON format, which we'll then](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=190.134) [write out to our cloud storage bucket. We](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=193.23049999999998) [can then execute a shell script on our](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=195.7417142857143) [cluster to spawn off a process on the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=198.54133333333337) [command line. We'll load the JSON result](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=201.26479999999998) [from cloud storage buckets into a BigQuery](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=203.82457142857143) [table. Bq load is the BigQuery command](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=207.11683333333337) [that will run on the command line. Bq is a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=209.23325000000003) [built-in command line utility available as](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=212.68488888888885) [a part of the gcloud SDK. The bq load](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=214.6375) [command takes the source data, which is in](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=218.48424999999995) [a JSON format in our cloud storage bucket,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=221.29766666666671) [and places it into a BigQuery table. This](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=223.82950000000005) [is the output table that we have specified](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=226.27524999999994) [earlier. The output\_files will be placed](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=229.85233333333332) [within our Dataproc staging bucket in the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=231.75499999999997) [pyspark\_output folder. And, finally, once](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=234.475) [our data has been safely stored within a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=238.52300000000002) [BigQuery table, we can cleanup all of the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=240.81714285714284) [cloud storage folders that we used, all of](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=243.55657142857143) [the extra folders that were generated in our staging bucket.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=5&mode=live&start=245.77644444444442)

[Executing a Spark Application to Get Results in BigQuery](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live)

[We have our input data in BigQuery. We](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=2.742) [have our PySpark code. We are now ready to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=5.0676) [submit this Spark job to our cluster.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=7.933333333333332) [Activate the Cloud Shell, and launch the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=10.399857142857144) [code editor beta. This is what we'll use](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=13.5685) [to edit configuration files and other](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=16.126714285714282) [little bits of code from within the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=18.017666666666663) [browser. Here is code editor beta. It's a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=20.246166666666667) [nifty new feature that you should](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=22.819) [definitely try out. I'm going to upload a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=24.13) [new file by clicking on the three-dot menu](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=26.853) [on the top right. This time, we'll store](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=29.92) [the Python program which contains our](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=32.35) [Spark application on the local machine on](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=34.28028571428572) [our Cloud Shell VM. So I'm going to upload](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=36.88371428571427) [totalcost. py, which is on my local](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=39.541) [machine here, to the Cloud Shell VM. This](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=41.99887500000002) [code editor points to the home directory](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=45.779714285714284) [on my Cloud Shell. If you click on File](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=47.6096) [and then Refresh, you'll see that](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=50.05512500000001) [totalcost. py is present here. You can](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=51.96650000000001) [click through to this file and see that](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=54.23966666666667) [your Spark application has been](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=55.7525) [successfully uploaded to the home](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=57.18) [directory in your Cloud Shell. You can now](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=59.05166666666666) [run this application on your Dataproc](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=61.258444444444464) [cluster using a gcloud command. We'll](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=63.83666666666666) [first set the current project on our Cloud](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=66.4194285714286) [Shell session to be spikey-dataproc, and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=69.04171428571429) [we'll call gcloud dataproc jobs submit to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=70.86600000000001) [submit our PySpark job. Point to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=73.75842857142858) [totalcost. py that is located in this](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=76.32849999999999) [current folder, that is our home](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=79.01637499999998) [directory, and the cluster to which we](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=80.992875) [want to submit this job is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=82.854) [spikey-cluster-one. Once the job has been](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=85.944) [submitted, we can use the navigation menu](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=87.49233333333332) [to go to Dataproc, Jobs in order to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=89.50571428571429) [monitor the status of its progress. You](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=92.5144) [can see from this web console here that](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=94.85644444444443) [our job is currently running and has been](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=96.476) [running for about 36 seconds. If you click](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=98.691) [on the running job, you'll be able to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=102.07300000000001) [monitor logs that it's generated from your](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=103.86585714285715) [Spark application. Here are the logs from](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=106.6185) [our BigQuery PySpark application. If you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=110.19733333333336) [switch back to Cloud Shell, you'll be able](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=112.8862857142857) [to monitor the job from the command line](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=114.6996) [as well. You can see some of the output](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=116.79549999999998) [that we've printed out from within our](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=119.10077777777778) [Spark application right here on screen.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=120.88550000000004) [Wait for a little bit for the job to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=123.876) [complete execution. You can switch over to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=125.769) [the Jobs dashboard and see that the job](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=128.498) [has run through successfully as you can](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=130.75222222222231) [see from the green checkmark here. The](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=132.57600000000002) [results of our Spark processing](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=135.7215) [application should now be available in](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=137.309) [BigQuery. Click on BigQuery from your](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=139.92500000000004) [navigation menu, and let's see if we can](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=141.73016666666663) [query the results. Under the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=144.80049999999991) [spikey-dataproc project, we have the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=146.66733333333335) [spikeysales\_dataset and the salesdata](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=148.62199999999999) [table. The output table doesn't seem to be](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=151.09399999999997) [available yet. I'm going to click on](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=153.475) [Refresh on my browser here to see if the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=155.439) [table has been updated. And if you click](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=157.705) [on the spikeysales\_dataset now, you'll see](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=159.889) [that totalcost\_output is now available. It](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=162.429) [has been populated with our final results.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=166.356625) [You can click on this table and observe](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=168.461) [the schema. We have the location and the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=170.76099999999997) [sales per location. And you can click on](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=173.72333333333341) [the Preview link in order to see a few](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=175.8425555555555) [rows from this table. We integrated our](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=178.19744444444441) [Python Spark application with BigQuery in](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=180.67600000000002) [order to calculate the total sales per location for our sale days.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=6&mode=live&start=182.75228571428573)

[Monitoring Spark Jobs on Dataproc](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live)

[We are already familiar with how](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=2.282) [Stackdriver monitoring works. Let's use](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=3.463) [Stackdriver to monitor our Spark](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=6.561249999999999) [applications. Start off in your main](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=9.304374999999997)[project dashboard page. We'll use the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=11.571124999999997) [navigation menu to go to the Monitoring](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=13.341333333333333) [option. And here within Monitoring, we'll](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=15.3006) [choose the Spikey-Dataproc Dashboard that](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=17.982) [we had set up earlier. Here are the two](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=19.922999999999995) [charts that we had set up earlier. We're](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=23.415749999999992) [going to add a new chart this time to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=25.315333333333335)[monitor Spark applications. Click on the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=28.27) [Add Chart button in the top right and set](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=30.13733333333333) [up a PySparkJob\_Chart. We are going to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=33.63099999999999) [continue monitoring our Dataproc cluster,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=36.729714285714294) [that is our resource type. We'll choose](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=38.309428571428576) [Cloud Dataproc Cluster here. With the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=41.403) [cluster chosen, the next step is to choose](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=43.688) [a metric that we want to monitor within](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=45.44166666666667) [the cluster. There are several job-related](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=47.4325) [metrics that you could choose from. One](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=49.88157142857143) [metric is job duration to monitor how long](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=52.312625) [your jobs take to complete. You can also](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=55.1475) [monitor submitted jobs, running jobs. You](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=58.336000000000006) [can monitor jobs based on their status.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=60.792714285714275) [Here is how you'd monitor your submitted](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=63.098625) [jobs, and you can choose the running jobs](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=65.0195) [as an option as well. This is the metric](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=67.5822857142857) [that we'll choose to monitor, our running](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=70.278) [jobs. We can specify additional filtering](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=72.97400000000002) [conditions for the jobs that we want to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=75.332) [monitor. Our filter condition here is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=77.71244444444449) [going to be based on cluster\_name. We're](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=79.06062500000003) [only interested in those jobs that are](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=81.26185714285714) [running on spikey-cluster-one. We've](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=84.07724999999999) [configured our chart. Go ahead and click](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=87.04975) [on the Save button, and you'll have a new](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=89.09450000000002) [chart added to your Spikey-Dataproc](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=91.34944444444443) [Dashboard. You can click on the details](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=93.59700000000001) [link here, and this will show you all of](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=96.30924999999998) [the jobs that are currently running in](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=97.93660000000003) [your cluster. Now the values that you see](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=100.08066666666664) [might be a little different based on the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=101.96600000000004) [jobs that you have running at this point](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=103.63825000000001) [in time. And with this we come to the very](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=106.46228571428574) [end of this module where we saw how we](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=108.37799999999996) [could use Dataproc to run our Spark](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=110.24911111111108) [applications on the cloud. Now Spark is a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=112.60514285714288) [great general-purpose computing engine,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=115.09885714285718) [and it's one of the most popular big data](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=116.72179999999999) [processing frameworks. Dataproc supports](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=119.533375) [Spark on the cloud. It uses YARN as its](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=121.2342857142857) [cluster manager. Dataproc has built-in](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=125.19314285714286) [support for Spark jobs which are written](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=127.31200000000001) [in Python, as well as Scala. You can use](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=129.79199999999997) [the web console or the gcloud command line](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=132.0313333333333) [utility to submit your jobs to the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=135.0485714285714) [cluster. The Spark applications that you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=137.43328571428566) [run on Dataproc have access to other GCP](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=138.8078571428571) [services as well. You can connect to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=141.7367142857142) [BigQuery, as well as cloud storage, in](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=144.12557142857136) [order to access the data that you want to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=145.87749999999994) [process, and to write out your final](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=148.79499999999993) [results. In the next module, we'll see how](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=151.33828571428575) [we can work with Pig, as well as Hive on cloud Dataproc.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=b36aff0c-cd4f-4049-a29d-a55933eef6a8&clip=7&mode=live&start=153.24477777777787)

[Working with Pig and Hive on Google Dataproc](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=0&mode=live)

[Module Overview](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=0&mode=live)

[Hi, and welcome to this module where we'll](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=0&mode=live&start=0) [see how we can use Google's Cloud Dataproc](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=0&mode=live&start=4.822) [to work with Pig and Hive. Now Pig and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=0&mode=live&start=7.904) [Hive are important technologies in the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=0&mode=live&start=10.876750000000001) [Hadoop ecosystem. In fact they're built on](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=0&mode=live&start=13.211285714285715) [top of Hadoop, and whatever script you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=0&mode=live&start=15.380874999999996) [write in Pig or query you right in Hive](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=0&mode=live&start=17.434000000000005) [actually runs MapReduce jobs under the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=0&mode=live&start=20.431142857142863) [hood in order to process your data. Pig is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=0&mode=live&start=22.480500000000003) [a scripting technology that runs in the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=0&mode=live&start=25.72785714285715) [Hadoop ecosystem. You write code in Pig](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=0&mode=live&start=28.57071428571428) [Latin, and you use Pig for extract,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=0&mode=live&start=30.875555555555554) [transform, and load operations where you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=0&mode=live&start=33.23199999999999) [clean up unstructured data in order to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=0&mode=live&start=35.5345) [load it into a data warehouse. Hive is a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=0&mode=live&start=38.07299999999999) [data warehouse that you use with Hadoop.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=0&mode=live&start=41.01660000000001) [Hive offers a SQL interface to any data](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=0&mode=live&start=42.929) [that you have stored within HDFS. This](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=0&mode=live&start=45.531) [allows data analysts with no knowledge of](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=0&mode=live&start=48.20728571428572) [MapReduce to extract information from](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=0&mode=live&start=50.79885714285714) [Hadoop. When you create a Dataproc cluster](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=0&mode=live&start=53.30814285714284) [on the GCP, it comes built in with support](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=0&mode=live&start=55.136000000000024) [for Pig, as well as Hive. You can use these technologies out-of-the-box.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=0&mode=live&start=58.420875000000024)

[Pig for Extract Transform Load](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live)

[Every technology that runs on top of the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=2.104) [Hadoop ecosystem has been built for a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=3.679) [special use case, and Pig is no exception.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=5.878) [Pig is what you'd use when you want to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=9.071) [script your extract, transform, and load](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=10.778) [operations in order to clean up the data](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=13.224) [before you store it in a data warehouse.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=15.919666666666664) [Pig can be thought of as a data](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=18.09455555555556) [manipulation language, which transforms](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=20.85166666666667) [unstructured data, which might be](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=23.137999999999998) [generated by sensors or log files into a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=25.381999999999998)[structured format. The Pig Latin scripting](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=28.528399999999998) [language has been tailored to work with](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=30.77114285714286) [raw, unstructured data. It allows you to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=33.548285714285704) [extract, parse, clean up this data, and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=36.22850000000001) [store the structured data into a data](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=38.729142857142854) [warehouse such as Hive. Once data is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=41.56728571428572) [present in Hive, it's accessible to your](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=43.87250000000001) [data analysts who can then extract this](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=46.00571428571429) [information using languages such as](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=48.45866666666666) [HiveQL. Data which might be collected from](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=51.147000000000006) [log files, sensors, or even your website](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=53.82742857142856)[might be in an unstructured format. The](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=56.61057142857145) [schema might be unknown. The data might be](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=59.9358) [incomplete. The data might be](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=62.87100000000001) [inconsistent. When you're working with](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=64.97299999999998) [data that has these characteristics, Pig](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=66.13757142857143) [is what comes to the rescue. Apache Pig is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=67.86) [a high-level scripting language that has](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=71.57585714285713) [been explicitly tailored to work with data](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=73.98233333333333) [that has an unknown or inconsistent](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=76.53199999999997) [schema. The exact term for this is that](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=79.886) [Apache Pig is used for extract, transform,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=82.41822222222223) [and load operations. This is where you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=85.15524999999997) [pull unstructured, inconsistent data from](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=87.566) [another source, you clean up this data,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=89.548) [extract the relevant information, and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=91.86985714285717) [place it in another database where the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=93.64949999999999) [data can be analyzed further. This is what](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=95.90942857142855) [a typical log file looks like. You have a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=99.34166666666668)[bunch of text messages that are stored](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=101.77600000000001) [somewhere on file. Now this log file, if](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=103.65357142857142) [you see, has some format even though it's](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=106.46124999999999) [unstructured. Here you have the server IP](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=108.95200000000001) [address. You have the date and time at](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=111.17114285714284) [which this log was generated. You have the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=113.03166666666671) [request type that was made to your website](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=115.5087142857143)[or application. You have the URL that was](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=118.1406666666667) [called. Pig can work with data of this](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=121.0425714285714) [type, extract all of the relevant parts of](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=123.06025000000004) [the information from these log messages,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=125.23400000000001) [and place it into a tabular format where](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=128.527) [it can then be queried. The scripting](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=130.31188888888886) [language that you use to work with Pig is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=133.28300000000002) [called Pig Latin. This is a procedural](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=135.55299999999997) [dataflow language to extract, transform,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=138.00885714285718) [and load data. It has a series of](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=139.99699999999999) [well-defined steps to perform data](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=142.87800000000004)[transformations. It does not have if](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=145.24166666666665) [statements or for loops. Pig Latin runs](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=147.7038888888889) [MapReduce jobs under the hood in order to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=150.11128571428569) [process your data. Data from one or more](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=152.85471428571432) [sources can be read, processed, and stored](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=155.288) [in parallel. You'll use Pig to clean your](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=158.13000000000002) [data, perform common aggregations, and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=161.08142857142857)[perform other pre-computations that will](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=163.21239999999995) [help you before you store this data in a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=165.2273333333333) [data warehouse. Here we have a visual](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=168.9988888888889) [showing the basic components of Hadoop.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=170.9365) [Pig is a technology that runs on top of](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=172.76371428571426) [Hadoop. It uses the Hadoop distributed](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=176.51149999999996) [computing framework in order to perform](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=178.94700000000003)[its operations. Pig processes data that is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=181.4612857142857) [stored in files on HDFS. Any intermediate](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=183.73887499999995) [results produced by Pig are also stored in](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=186.39000000000004) [HDFS. The final output will also be](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=189.58571428571423) [written to HDFS. Pig runs MapReduce tasks](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=192.38833333333326) [in order to perform its data](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=194.72714285714284) [transformation tasks. Pig scripts can run](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=197.46520000000004) [very fast because they leverage the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=199.65499999999997) [parallel computing framework that](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=201.6268) [MapReduce offers. Pig has built-in](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=203.87) [MapReduce implementations for common](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=206.86599999999999)[operations, which are run very, very efficiently.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=1&mode=live&start=208.65885714285716)

[Running Pig Scripts on Dataproc](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live)

[In addition to Hadoop and Spark, Dataproc](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=1.714) [comes preinstalled with Pig. Let's see how](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=4.902200000000001) [we can run Pig scripts on our Dataproc](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=7.649900000000001) [cluster. Here is the data that we're going](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=9.803) [to process using a Pig script. This](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=11.989874999999996) [contains some product data from our Spiky](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=14.723714285714285) [Sales e-commerce site. We have the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=17.288) [original product category, such as](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=20.47) [clothing, the actual product. Then we have](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=22.293999999999997) [the MSRP and the discounted rate and the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=24.4995) [brand for that particular product. The](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=26.894875000000003) [brands that are available here are Puma](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=29.448222222222224) [Nike. Switch over to our GCP web console](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=32.704777777777785) [dashboard, click on the hamburger icon,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=35.242) [and use the navigation menu to go to our](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=37.63) [cloud storage buckets. We're going to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=40.366) [upload this file into the spikey-data](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=42.512499999999996) [bucket that we created earlier. I'm going](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=44.546)[to create a new folder here and call it](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=48.121111111111105) [pig\_input. This is where the input files](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=50.21125) [that our Pig scripts will process will be](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=52.670857142857145) [placed. Click through to pig\_input and use](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=55.56914285714286) [the Upload files button in order to upload](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=58.32033333333334) [itemdetails. txt to the pig\_input folder.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=61.030142857142856) [Once the file has been uploaded](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=65.249) [successfully, we can activate the Cloud](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=66.68) [Shell using this button on the top right,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=69.19) [and then run our first Pig job on the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=71.736) [Dataproc cluster. We'll use the gcloud](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=74.429)[command line utility as we've done before,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=77.588) [gcloud dataproc jobs submit pig and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=80.02) [specify that we want to submit this Pig](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=83.02) [job to spiky-cluster-one. The --e or the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=85.291) [--execute option for gcloud dataproc jobs](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=89.43116666666666) [submit allows you to specify the Pig](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=92.57633333333332) [script right here on the command line. And](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=94.98825000000002) [here are the different lines of our Pig](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=96.98566666666667) [script. We first load the data from](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=99.98466666666664) [itemdetails. txt in our spikey-data bucket](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=102.81300000000003) [and specify a field name for each column.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=105.14614285714288)[Along with the field name, we specify the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=109.503) [field data type as well. This can be](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=111.253) [chararray, float, int, and so on. We then](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=114.62200000000001) [perform a GROUP BY action on the data that](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=118.11533333333333) [we read in. We want to group by the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=120.1912) [category of product, which is specified in](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=122.36971428571428) [the type field, and we want to dump the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=124.45599999999999) [groups out to our terminal window. This we](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=126.724) [do using the DUMP GROUPS command in Pig.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=130.82500000000002) [Hit Enter. The Pig job will be submitted](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=134.109) [to our Dataproc cluster. You can go to the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=136.768) [navigation menu, go to Jobs under](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=138.991) [Dataproc, and you should see the job](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=141.67214285714286) [running right there. You can see its](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=143.94566666666665) [status is running. It has been running for](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=145.73842857142856) [about 18 seconds. And you can see that the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=147.88299999999998) [type of jog is a Pig job. Click on the job](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=150.0754545454546) [ID, and you can view the log outputs from](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=153.3332) [your Pig script. Notice that under the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=156.65262499999997) [hood, your Pig scripts run a MapReduce](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=158.98985714285712) [job. You can see the mapper and reducer](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=161.42580000000004) [running here, and you can see the same](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=163.67950000000008) [thing on your Cloud Shell terminal window.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=165.40144444444445) [You'll find that the job will run through](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=168.448) [successfully in a minute or so. You've](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=170.078)[successfully submitted your first Pig script to your Dataproc cluster.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=2&mode=live&start=172.5032)

[Storing Pig Output to Cloud Storage](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=3&mode=live)

[We'll continue with our previous example](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=3&mode=live&start=2.43) [and run a Pig script on our Dataproc](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=3&mode=live&start=4.048) [cluster once again, but this time we'll](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=3&mode=live&start=6.366) [specify the script in a file. We'll write](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=3&mode=live&start=8.471) [our Pig script using code editor beta.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=3&mode=live&start=12.007142857142858) [Launch code editor beta and wait for it to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=3&mode=live&start=14.301285714285715) [be up and running. Click on the File](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=3&mode=live&start=17.240499999999997) [button, and click on the New option in the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=3&mode=live&start=19.256777777777778) [menu to create a new file. Pig scripts](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=3&mode=live&start=21.8278) [have the. pig extension. We'll call this](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=3&mode=live&start=25.07166666666667) [file itemdiscounts. pig. We'll write the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=3&mode=live&start=29.033) [code in our Pig script here. We'll load](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=3&mode=live&start=31.951999999999995) [data from the itemdetails. txt file that](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=3&mode=live&start=34.30457142857142) [is stored in our cloud storage bucket. We](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=3&mode=live&start=36.97544444444445) [then filter out the header row from this](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=3&mode=live&start=40.0068)[file. We only want the data; we don't want](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=3&mode=live&start=42.29320000000001) [the header. Extract each of the fields](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=3&mode=live&start=45.39659999999998) [from the file that we've read in, and add](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=3&mode=live&start=47.33850000000001) [a new column which calculates how much of](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=3&mode=live&start=49.381666666666675) [a discount we got off the retail price.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=3&mode=live&start=51.606999999999985) [That's the last column here. Filter our](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=3&mode=live&start=54.277) [data to only keep the Nike and the Puma](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=3&mode=live&start=57.26411111111111)[brands, and then group this resulting](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=3&mode=live&start=59.79628571428571) [information by category or type. This time](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=3&mode=live&start=62.0655) [we want to store our scripting results](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=3&mode=live&start=65.7255) [back to our spikey-data bucket in the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=3&mode=live&start=67.594) [pig\_output folder. Itemdiscounts. pig is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=3&mode=live&start=71.228) [present in the home directory of our cloud](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=3&mode=live&start=73.9132857142857) [shell. We can now submit this Pig script](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=3&mode=live&start=76.73400000000002) [to Dataproc using gcloud dataproc jobs](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=3&mode=live&start=79.58671428571427) [submit pig, specify that you want to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=3&mode=live&start=82.3114) [submit this to spiky-cluster-one, and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=3&mode=live&start=84.463375) [point to our itemdiscounts file. This file](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=3&mode=live&start=85.9776) [is present in the current working](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=3&mode=live&start=89.47500000000002) [directory. Hit Enter, and this job will be](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=3&mode=live&start=93.02099999999999) [submitted to your cluster. You can go to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=3&mode=live&start=95.368) [your cluster jobs web console, and you can](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=3&mode=live&start=97.21799999999999) [see that this job is currently running.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=3&mode=live&start=99.68699999999995) [Wait for the script to run through](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=3&mode=live&start=102.943) [successfully. If you scan the cloud shell](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=3&mode=live&start=105.43937500000004) [terminal window, you'll find that we've](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=3&mode=live&start=107.67571428571428) [successfully stored two records in our](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=3&mode=live&start=109.69816666666671) [spikey-data/pig\_output folder. Go to cloud](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=3&mode=live&start=111.864) [storage using the navigation pane on the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=3&mode=live&start=116.8587142857143) [left. This will show you all of the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=3&mode=live&start=118.95966666666668) [buckets that we have on the cloud. Within](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=3&mode=live&start=121.26944444444445) [spikey-data, you should now see the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=3&mode=live&start=123.29042857142858) [pig\_output folder. And if you click](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=3&mode=live&start=124.991) [through to pig\_output here, you'll find](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=3&mode=live&start=127.823) [that our MapReduce job, which was](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=3&mode=live&start=129.799) [triggered by our Pig script, has been](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=3&mode=live&start=132.4458571428572) [successfully completed, and the output](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=3&mode=live&start=134.64928571428572) [generated as well. You can see here the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=3&mode=live&start=137.1927142857143) [technologies in the Hadoop ecosystem have](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=3&mode=live&start=139.90437500000007) [been seamlessly integrated with your cloud](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=3&mode=live&start=141.9596)[storage buckets, thus, effectively separating your compute from storage.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=3&mode=live&start=144.63666666666668)

[Hive to Query Big Data](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live)

[In all of the demos we've seen so far,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=2.938) [we've seen how easy it is to work with](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=4.728) [open-source Hadoop technologies on](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=7.161) [Google's Cloud Dataproc. And the same](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=9.246599999999999) [thing is true for Hive as well. Hive is a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=11.9168) [data warehouse that has been built on top](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=14.335125000000001) [of Hadoop, and Hive provides a SQL](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=16.53375) [interface to data that has been stored in](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=18.642) [files in HDFS. Hive can be thought of as a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=21.409999999999997) [way to perform great data processing in](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=24.54019999999999) [Hadoop for folks who don't have exposure](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=26.603857142857144) [to object-oriented programming in Java.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=29.687666666666665) [You can use MapReduce on Hadoop by just](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=31.886142857142858) [writing a few queries in Hive. Hive in a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=35.11540000000001) [way serves to democratize MapReduce by](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=38.33714285714285) [eliminating the need to write Java code.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=40.158571428571435) [Just like other technologies in the Hadoop](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=42.89642857142858) [open source ecosystem, Hive is built on](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=45.84214285714286) [top of Hadoop. Hive uses HDFS, MapReduce,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=48.58819999999999) [and YARN in order to run queries. Data](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=51.764285714285705) [that you store in Hive is stored in the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=55.11525000000001) [form of files. These can be text files or](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=57.83066666666667) [binary files. All of these files are](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=60.727750000000015) [partitioned across machines in the Hadoop](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=62.75357142857142) [cluster. The files are replicated for](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=65.11699999999999) [fault tolerance, and any processing tasks](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=66.96033333333335) [that you run using Hive is parallelized](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=69.2286666666667) [across multiple machines in your cluster.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=71.71328571428573) [All of the processing that Hive performs](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=74.997) [on your data is run in the form of](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=76.721625) [MapReduce jobs under the hood. As you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=79.7857777777778) [already know, MapReduce jobs are run in](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=82.26050000000004) [parallel across all cluster machines, and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=84.5322857142857) [they work on subsets of your data. The](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=86.71857142857145) [best thing about using Hive, though, is](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=89.87650000000002) [that you can write MapReduce jobs using](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=92.324)[queries without actually writing any](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=95.4234) [MapReduce Java code. Instead of working](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=98.69300000000001) [with MapReduce code, you'll write](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=101.69250000000001) [structured queries using HiveQL. HiveQL](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=103.81866666666666) [else stands for Hive Query Language, and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=106.512) [it's a SQL-like interface to the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=108.96071428571429) [underlying data that's stored on HDFS.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=110.693) [Data that is stored in the form of files](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=113.741) [in HDFS is exposed to the user in Hive](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=115.44649999999996) [\_\_\_\_\_ tables. So Hive actually creates](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=119.94480000000001) [tables and structures the underlying file](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=122.59333333333332) [data so that it can be queried by the user](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=125.21385714285712) [in a tabular format. As a user of Hive,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=129.2124545454546) [you'll write a SQL-like query in order to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=132.32433333333333) [operate on the underlying data. You'll do](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=135.21455555555553)[this in HiveQL, and you'll then submit](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=137.752) [this to Hive. Hive will then translate](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=139.72257142857143) [this query to MapReduce jobs under the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=142.30733333333333) [hood, which it'll run in parallel across](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=145.08425) [the Hadoop cluster. MapReduce will process](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=147.16725000000002) [the files that are stored in HDFS in](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=149.42571428571432) [Hadoop and then return the results to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=152.3525) [Hive, which will then display these](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=154.68014285714287) [results to the user. Hive completely](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=156.52066666666667) [abstracts away the details of the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=159.483) [underlying MapReduce jobs that are run to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=161.131) [process data from the user. You'll work](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=163.451) [with Hive almost exactly like you would with a traditional relational database.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=4&mode=live&start=166.4252857142857)

[Executing Hive Queries on Dataproc](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live)

[In this demo, we'll see how we can run](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=2.216) [Hive queries on our Dataproc cluster. We](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=4.124000000000001) [start this demo off in the Dataproc](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=7.73775) [Cluster page on the GCP web console. We](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=9.654) [are currently viewing spikey-cluster one.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=13.3135) [SSH into the master node. Hive has a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=16.861) [simple command line shell that can be](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=21.215375) [invoked by calling the hive command.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=23.419749999999997) [Calling Hive will take you to the Hive](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=25.981) [interactive shell where you can run](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=28.032) [queries. Let's create a new Hive table](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=31.181142857142863) [here to store our e-commerce items](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=33.18866666666666) [information. We'll create this table if it](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=35.97) [doesn't already exist, and we specify the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=37.542285714285704) [columns for this table. Remember that this](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=40.187857142857155) [Hive table is simply an interface or a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=42.71428571428572) [bridge on top of the data that is stored](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=45.190111111111115) [typically in the form of a text file or](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=47.49255555555557) [maybe even binary files. Specify the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=49.59477777777778) [format of the file that holds the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=51.67985714285713) [underlying data. Our Hive table has been](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=54.87699999999999) [successfully created here. If you're using](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=57.31699999999999) [Hadoop on-premises, and your running Hive](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=59.746999999999986) [on top of Hadoop, your text file and other](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=61.322857142857124) [files will be stored on HDFS. In our case,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=64.34487499999997) [though, our file lives in cloud storage](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=66.988625) [buckets, specifically the spikey-data](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=69.24139999999998) [bucket. Inside spikey-data in the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=71.2512) [pig\_input folder, we have the itemdetails.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=74.182) [txt file, and this is the data that we'll](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=76.614375) [use to load into our Hive table and query](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=80.28322222222221) [using the HiveQL language. Let's switch](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=84.187) [back to our terminal window here where we](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=87.49128571428572) [are logged into the Hive shell. I'm going](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=89.70199999999998) [to load data into the table that we just](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=92.24711111111112) [created. The table name was called items.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=94.99222222222224) [Our Hive table will simply point to the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=98.045) [data that is stored in this cloud storage](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=99.796) [bucket in the itemdetails. txt file. Once](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=102.032) [the data has been loaded into our table](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=106.134875) [successfully, we can query this using](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=108.187) [HiveQL. SELECT \* FROM items will display](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=111.30699999999997) [all of the items in our itemdetails. txt](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=113.905) [file. Notice that HiveQL queries are very](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=117.83066666666664)[similar to SQL queries. You can specify](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=120.87657142857144) [more complex queries as well using the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=123.72228571428569) [WHERE clause, the ORDER BY, and so on.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=126.54857142857146) [Observe that when you execute these](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=129.888) [queries, Hadoop runs MapReduce jobs under](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=131.51) [the hood in order to give you the results.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=134.392) [You can also run Hive commands without](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=137.532)[having to log in to the master node of](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=140.176) [your cluster. You can run from the command](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=143.06909090909087) [line using gcloud exactly like you did](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=145.23599999999993) [with Pig scripts, Spark jobs, and Hadoop](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=147.33257142857144) [MapReduce jobs. Here is what a gcloud](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=151.01371428571434) [command to execute a Hive query on your](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=153.4275555555555) [Dataproc cluster would look like. Hive](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=155.71657142857143) [queries can be specified as a command line](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=157.85042857142855) [argument, or if there are more complex](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=160.39011111111114) [queries, they can be specified in a file.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=162.34914285714288) [Let's clean up after ourselves. I'm going](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=165.04087499999997) [to go ahead and drop the items Hive table](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=167.2613333333333) [that we just created. Working with Hive on](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=170.077) [Dataproc is very similar to working with](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=173.13357142857143) [Hive on an on-premises Hadoop cluster. In](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=175.09428571428575) [fact, it's easier because you can have your data on cloud storage.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=5&mode=live&start=177.885875)

[Summary and Further Study](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=6&mode=live)

[And this brings us to the very end of this](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=6&mode=live&start=1.879) [module where we saw how we could use Pig,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=6&mode=live&start=4.715) [as well as Hive on our Dataproc cluster.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=6&mode=live&start=7.412) [Pig and Hive are extremely important](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=6&mode=live&start=10.377) [technologies in the open source world, and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=6&mode=live&start=12.002) [if you're migrating to cloud Dataproc,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=6&mode=live&start=14.289428571428571) [it's important that you have equivalents](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=6&mode=live&start=16.116)[available, and that's what Google offers.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=6&mode=live&start=18.563) [Hive offers a SQL interface to data that](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=6&mode=live&start=21.148) [is stored in Hadoop. It democratizes the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=6&mode=live&start=23.551555555555556) [process of MapReduce so that big data](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=6&mode=live&start=26.026) [processing is available to those handlers](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=6&mode=live&start=28.844) [who do not write Java code. Pig, on the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=6&mode=live&start=31.058) [other hand, is an important scripting](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=6&mode=live&start=33.937000000000005) [mechanism that allows you to perform](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=6&mode=live&start=35.31583333333333) [extract/transform/load operations on your](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=6&mode=live&start=38.15833333333333) [data. This allows you to clean up your](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=6&mode=live&start=40.48722222222223) [data so that it can be stored in a data](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=6&mode=live&start=42.699)[warehouse such as Hive. Cloud Dataproc](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=6&mode=live&start=44.8352) [offers built-in support for both Pig and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=6&mode=live&start=47.989333333333335) [Hive, which makes migrating to the cloud a](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=6&mode=live&start=50.925124999999994) [simple decision. As you work through the](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=6&mode=live&start=54.504200000000004) [demos of this course, we created Dataproc](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=6&mode=live&start=56.43649999999999) [clusters running on GCP VMs. Now if you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=6&mode=live&start=59.328999999999986) [leave these clusters running, you will](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=6&mode=live&start=62.463777777777764)[incur charges for your clusters, so it's](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=6&mode=live&start=65.04133333333334) [better to turn them down when you're not](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=6&mode=live&start=67.09575000000002) [using them. And on this note, we come to](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=6&mode=live&start=70.19228571428573) [the very end of this course on Google's](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=6&mode=live&start=72.44149999999999) [Cloud Dataproc. It's time for us to say](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=6&mode=live&start=74.375) [goodbye. But before I head out, here are](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=6&mode=live&start=77.14985714285714) [some other courses on Pluralsight that you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=6&mode=live&start=79.05111111111111)[might find interesting. If you're](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=6&mode=live&start=81.46874999999999) [interested in big data processing on other](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=6&mode=live&start=83.15085714285715) [cloud platforms, Big Data on Amazon Web](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=6&mode=live&start=85.20914285714285) [Services will show you how you can use](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=6&mode=live&start=87.76599999999999) [elastic MapReduce on AWS. If, instead, you](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=6&mode=live&start=90.77866666666667) [want to learn about Azure's managed Hadoop](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=6&mode=live&start=93.58928571428572) [offering, HDInsight Deep Dive is a good](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=6&mode=live&start=96.44000000000001)[course that you might want to see. If](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=6&mode=live&start=99.11500000000002) [you're interested in the GCP, and you want](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=6&mode=live&start=101.753) [to learn about other Google technologies,](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=6&mode=live&start=103.65899999999998) [here are some courses on Pluralsight that](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=6&mode=live&start=105.53449999999998) [you might want to watch. Creating and](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=6&mode=live&start=107.71657142857144) [Administering Google Cloud SQL Instances](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=6&mode=live&start=110.49619999999997) [will show you how you can move your](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=6&mode=live&start=112.77142857142856)[relational databases to the Google Cloud.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=6&mode=live&start=115.8197142857143) [Or if you're interested in cloud storage](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=6&mode=live&start=117.64657142857142) [solutions, Architecting Google Cloud](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=6&mode=live&start=120.34619999999998) [Storage Configurations is the course for](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=6&mode=live&start=122.27274999999999) [you. And that's it from me here today. Goodbye, and thank you for listening.](https://app.pluralsight.com/player?course=google-dataproc-architecting-big-data-solutions&author=janani-ravi&name=97ff9549-8a9b-44ba-a6ef-98d884e813d6&clip=6&mode=live&start=124.8618333333333)