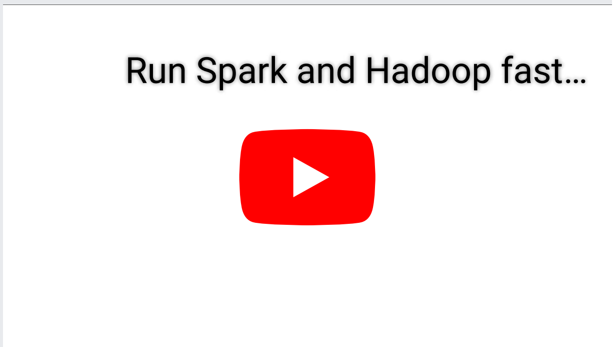
What is Dataproc?

Dataproc is a managed Spark and Hadoop service that lets you take advantage of open source data tools for batch processing, querying, streaming, and machine learning. Dataproc automation helps you create clusters quickly, manage them easily, and save money by turning clusters off when you don't need them. With less time and money spent on administration, you can focus on your jobs and your data.

Why use Dataproc?

When compared to traditional, on-premises products and competing cloud services, Dataproc has a number of unique advantages for clusters of three to hundreds of nodes:



- **Low cost** — Dataproc is priced (/dataproc/docs/resources/pricing) at only 1 cent per virtual CPU in your cluster per hour, on top of the other Cloud Platform resources you use. In addition to this low price, Dataproc clusters can include preemptible instances (https://cloud.google.com/preemptible-vms) that have lower compute prices, reducing your costs even further. Instead of rounding your usage up to the nearest hour, Dataproc charges you only for what you really use with second-by-second billing and a low, one-minute-minimum billing period.
- **Super fast** — Without using Dataproc, it can take from five to 30 minutes to create Spark and Hadoop clusters on-premises or through IaaS providers. By comparison, Dataproc clusters are quick to start, scale, and shutdown, with each of these operations taking 90 seconds or less, on average. This means you can spend less time waiting for clusters and more hands-on time working with your data.
- **Integrated** — Dataproc has built-in integration with other Google Cloud Platform services, such as BigQuery (/bigquery), Cloud Storage (/storage), Cloud Bigtable (/bigtable), Cloud Logging (/logging), and Cloud Monitoring (/monitoring), so you have more than just a Spark or Hadoop cluster—you have a complete data platform. For example, you can use Dataproc to effortlessly ETL terabytes of raw log data directly into BigQuery for business reporting.
- **Managed** — Use Spark and Hadoop clusters without the assistance of an administrator or special software. You can easily interact with clusters and Spark or Hadoop jobs through the Google Cloud console, the Cloud SDK, or the Dataproc REST API. When you're done with a cluster, you can simply turn it off, so you don't spend

money on an idle cluster. You won't need to worry about losing data, because Dataproc is integrated with Cloud Storage (/storage), BigQuery (/bigquery), and Cloud Bigtable (/bigtable).

- **Simple and familiar** — You don't need to learn new tools or APIs to use Dataproc, making it easy to move existing projects into Dataproc without redevelopment. Spark, Hadoop, Pig, and Hive are frequently updated, so you can be productive faster.

What is included in Dataproc?

For a list of the open source (Hadoop, Spark, Hive, and Pig) and Google Cloud Platform connector versions supported by Dataproc, see the Dataproc version list (/dataproc/docs/concepts/dataproc-versions).

Getting Started with Dataproc

To quickly get started with Dataproc, see the Dataproc Quickstarts (/dataproc/docs/quickstarts). You can access Dataproc in the following ways:

- Through the REST API (/dataproc/docs/reference/rest)
- Using the Cloud SDK (/sdk/gcloud/reference/dataproc)
- Using the Dataproc UI (/dataproc/docs/guides/create-cluster#using_the_console_name)
- Through the Cloud Client Libraries (/dataproc/docs/quickstarts/create-cluster-client-libraries)

Last updated 2022-09-30 UTC.