Dataproc

Dataproc is a fully managed and highly scalable service for running Apache Spark, Apache Flink, Presto, and 30+ open source tools and frameworks. Use Dataproc for data lake modernization, ETL, and secure data science, at planet scale, fully integrated with Google Cloud, at a fraction of the cost.

- Open: Run open source data analytics at scale, with enterprise grade security
- Flexible: Use serverless, or manage clusters on Google Compute and Kubernetes
- Intelligent: Enable data users through integrations with Vertex AI, BigQuery, and Dataplex
- Secure: Configure advanced security such as Kerberos, Apache Ranger and Personal Authentication
- Cost-effective: Realize 54% lower TCO compared to on-prem data lakes with per-second pricing

BENEFITS

Modernize your open source data processing

Whether you need VMs or Kubernetes, extra memory for Presto, or even GPUs, Dataproc can help accelerate your data and analytics processing through on-demand purpose-built or serverless environments.

Intelligent and seamless OSS for data science

Enable data scientists and data analysts to seamlessly perform data science jobs through native integrations with Vertex AI.

Advanced security, compliance, and governance

Manage and enforce user authorization and authentication using existing Kerberos and Apache Ranger policies or Personal Cluster Authentication. Define permissions without having to set up a network node.

Key features

Fully managed and automated big data open source software

Serverless deployment, logging, and monitoring let you focus on your data and analytics, not on your infrastructure. Reduce TCO of Apache Spark management by up to 54%. Enable data scientists and engineers to build and train models 5X faster, compared to traditional notebooks, through integration with Vertex AI Workbench. The Dataproc Jobs API makes it easy to incorporate big data processing into custom applications, while Dataproc Metastore eliminates the need to run your own Hive metastore or catalog service.

Containerize Apache Spark jobs with Kubernetes

Build your Apache Spark jobs using Dataproc on Kubernetes so you can use Dataproc with Google Kubernetes Engine (GKE) to provide job portability and isolation.

Enterprise security integrated with Google Cloud

When you create a Dataproc cluster, you can enable Hadoop Secure Mode via Kerberos by adding a Security Configuration. Additionally, some of the most commonly used Google Cloud-specific security features used with Dataproc include default at-rest encryption, OS Login, VPC Service Controls, and customer-managed encryption keys (CMEK).

The best of open source with the best of Google Cloud

Dataproc lets you take the open source tools, algorithms, and programming languages that you use today, but makes it easy to apply them on cloud-scale datasets. At the same time, Dataproc has out-of-the-box integration with the rest of the Google Cloud analytics, database, and AI ecosystem. Data scientists and engineers can quickly access data and build data applications connecting Dataproc to BigQuery, Vertex AI, Cloud Spanner, Pub/Sub, or Data Fusion.

All features

Serverless Spark	Deploy Spark applications and pipelines that autoscale without any manual infrastructure provisioning or tuning.
Resizable clusters	Create and scale clusters quickly with various virtual machine types, disk sizes, number of nodes, and networking options.
Autoscaling clusters	Dataproc autoscaling provides a mechanism for automating cluster resource management and enables

	automatic addition and subtraction of cluster workers (nodes).
Cloud integrated	Built-in integration with Cloud Storage, BigQuery, Cloud Bigtable, Cloud Logging, Cloud Monitoring, and AI Hub, giving you a more complete and robust data platform.
Versioning	Image versioning allows you to switch between different versions of Apache Spark, Apache Hadoop, and other tools.
Highly available	Run clusters in high availability mode with multiple main nodes and set jobs to restart on failure to help ensure your clusters and jobs are highly available.
Cluster scheduled deletion	To help avoid incurring charges for an inactive cluster, you can use Dataproc's scheduled deletion, which provides options to delete a cluster after a specified cluster idle period, at a specified future time, or after a specified time period.
Automatic or manual configuration	Dataproc automatically configures hardware and software but also gives you manual control.
Developer tools	Multiple ways to manage a cluster, including an easy-to-use web UI, the Cloud SDK, RESTful APIs, and SSH access.
Initialization actions	Run initialization actions to install or customize the settings and libraries you need when your cluster is created.
Optional components	Use optional components to install and configure additional components on the cluster. Optional components are integrated with Dataproc components and offer fully configured environments for Zeppelin, Druid, Presto, and other open source software

	components related to the Apache Hadoop and Apache Spark ecosystem.
Custom images	Dataproc clusters can be provisioned with a custom image that includes your pre-installed Linux operating system packages.
Flexible virtual machines	Clusters can use custom machine types and preemptible virtual machines to make them the perfect size for your needs.
Component Gateway and notebook access	Dataproc Component Gateway enables secure, one-click access to Dataproc default and optional component web interfaces running on the cluster.
Workflow templates	Dataproc workflow templates provide a flexible and easy-to-use mechanism for managing and executing workflows. A workflow template is a reusable workflow configuration that defines a graph of jobs with information on where to run those jobs.

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.