Google Cloud Dataproc

Google Cloud 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.

Features

- Fast, easy, managed way to run Hadoop and Spark/Hive/Pig.
- Benefit from cloud integration (Storage, Stackdriver, etc).
- Customize and configure clusters with initialization actions.
- Create clusters in 90 sec or less.
- Pay-per-minute billing.
- Scale clusters up and down even when jobs are running.
- Tools including RESTful API and GCP SDK integration.

When to Use?

- Migrate on-prem Hadoop jobs to the cloud.
- Analyze data stored in Cloud Storage.
- Use Spark/Spark SQL to quickly perform data mining and analysis.
- Use Spark Machine Learning Libraries for classification models.