

Как деплоить в Google Cloud

AWS, GCP, Azure,
Oracle



ВСЁ БЕСПЛАТНО?

Укажите кредитную или дебетовую карту. Мы выполним пробное списание средств с этой карты, но **оплата не будет взиматься, если вы не перешли на модель с оплатой по мере использования**. Мы не принимаем предоплатные карты.

Мы принимаем следующие карты:



Имя владельца карты

Номер карты

Срок действия

ММ

ГГ

CVV-код

[Что такое CVV?](#)

Куда деплоить Spark Job



Cloud Dataproc API

Google Enterprise API

Manages Hadoop-based clusters and jobs on Google Cloud Platform.

MANAGE

TRY THIS API [↗](#)

✓ API Enabled

OVERVIEW

DOCUMENTATION

Overview

Manages Hadoop-based clusters and jobs on Google Cloud Platform.

Additional details


Type: [SaaS & APIs](#)

Last updated: 7/23/21

Category: [Google Enterprise APIs](#)

Service name: dataproc.googleapis.com

Создание кластера: название и тип

 Dataproc

Jobs on Clusters

Clusters

Jobs

Workflows

Autoscaling policies

Serverless

Batches

Utilities

Component exchange

Metastore

Workbench

Release Notes

Create a cluster

Set up cluster
Begin by providing basic information.

Configure nodes (optional)
Change node compute and storage capabilities.

Customize cluster (optional)
Add cluster properties, features, and actions.

Manage security (optional)
Change access, encryption, and security settings.

CREATE

CANCEL

EQUIVALENT COMMAND LINE

Name

Cluster Name *
cluster-example

Location

Region *
europe-west1

Zone *
europe-west1-c

Cluster type

☒ Standard (1 master, N workers)

☐ Single Node (1 master, 0 workers)
Provides one node that acts as both master and worker. Good for proof-of-concept or small-scale processing

☐ High Availability (3 masters, N workers)
Hadoop High Availability mode provides uninterrupted YARN and HDFS operations despite single-node failures or reboots

Autoscaling


Automates cluster resource management based on an autoscaling policy.

Policy
None

Enhanced Flexibility Mode

Dataproc Enhanced Flexibility Mode (EFM) manages shuffle data to minimize job

Создание кластера: система и компоненты

 Dataproc

Jobs on Clusters

Clusters

Jobs

Workflows

Autoscaling policies

Serverless

Batches

Utilities

Component exchange

Metastore

Workbench

Release Notes

Create a cluster

Set up cluster

Begin by providing basic information.

Configure nodes (optional)

Change node compute and storage capabilities.

Customize cluster (optional)

Add cluster properties, features, and actions.

Manage security (optional)

Change access, encryption, and security settings.

CREATE

CANCEL

EQUIVALENT COMMAND LINE

Versioning

Use a custom image to load pre-installed packages. [Learn more](#)

Image Type and Version

2.0-debian10

Release Date

First released on 1/22/2021.

CHANGE

Components

Component Gateway

☐ Enable component gateway

Provides access to the web interfaces of default and selected optional components on the cluster. [Learn more](#)

Optional components

Select one or multiple components. [Learn more](#)

☐ Anaconda

☐ Hive WebHCat

☐ Jupyter Notebook

☐ Zeppelin Notebook

☐ Druid

☐ Presto


☐ ZooKeeper

☐ Ranger

☐ HBase

☐ Flink

Создание кластера: тип master node

 Dataproc

Jobs on Clusters

Clusters

Jobs

Workflows

Autoscaling policies

Serverless

Batches

Utilities

Component exchange

Metastore

Workbench

Release Notes

Create a cluster

- Set up cluster
Begin by providing basic information.
- Configure nodes (optional)
Change node compute and storage capabilities.
- Customize cluster (optional)
Add cluster properties, features, and actions.
- Manage security (optional)
Change access, encryption, and security settings.

CREATE CANCEL

EQUIVALENT COMMAND LINE

Master node

Contains the YARN Resource Manager, HDFS NameNode, and all job drivers.

Machine family

GENERAL-PURPOSE COMPUTE-OPTIMIZED MEMORY-OPTIMIZED

Machine types for common workloads, optimized for cost and flexibility


Series

N1

Powered by Intel Skylake CPU platform or one of its predecessors

Machine type

n1-standard-2 (2 vCPU, 7.5 GB memory)



vCPU

2

Memory

7.5 GB

CPU PLATFORM AND GPU

Primary disk size * 500 GB

Primary disk type Standard Persistent Disk

Number of local SS... x 375GB

Local SSD Interface

Worker nodes

Each contains a YARN NodeManager and a HDFS DataNode. HDFS replication factor is 2.

Machine family

Создание кластера: тип worker node

Dataproc

Jobs on Clusters

Clusters

Jobs

Workflows

Autoscaling policies

Serverless

Batches

Utilities

Component exchange

Metastore

Workbench

Release Notes

← Create a cluster

• Set up cluster
Begin by providing basic information.

• **Configure nodes (optional)**
Change node compute and storage capabilities.

• Customize cluster (optional)
Add cluster properties, features, and actions.

• Manage security (optional)
Change access, encryption, and security settings.

CREATE

CANCEL

EQUIVALENT COMMAND LINE

▼

Worker nodes

Each contains a YARN NodeManager and a HDFS DataNode. HDFS replication factor is 2.

Machine family

GENERAL-PURPOSE

COMPUTE-OPTIMIZED

MEMORY-OPTIMIZED

Machine types for common workloads, optimized for cost and flexibility

Series

N1

Powered by Intel Skylake CPU platform or one of its predecessors

Machine type

n1-standard-2 (2 vCPU, 7.5 GB memory)

vCPU

2

Memory

7.5 GB

✓ CPU PLATFORM AND GPU

Number of worker nodes *

2

?

Primary disk size *

500

GB

?

Primary disk type

Standard Persistent Disk

?

Number of local SS...

0

x 375GB

?


Local SSD Interface

SCSI

?

Secondary worker nodes

Кластер создан

 Databricks

Jobs on Clusters ^

Clusters

Jobs

Workflows

Autoscaling policies

Serverless ^

Batches

Utilities ^

Component exchange

Metastore

Workbench

Clusters

CREATE CLUSTER

REFRESH

START

STOP

DELETE

REGIONS

+ 5 RECOMMENDED ALERTS

HIDE INFO PANEL

Filter Search clusters, press Enter

☒

Name ↑

☒

cluster-ea17

Stopped

us-central1

us-central1-f

2

Off

PERMISSIONS

LABELS

Edit or delete permissions below or "Add Principal" to grant new

ADD PRINCIPAL

Show inherited permissions

Filter Enter property name or value

Role / Principal ↑

Inheritance

Databricks Administrator (1)

Databricks Service Agent (1)

Editor (3)

Owner (2)

7

Джобы

Dataproc

Jobs

SUBMIT JOB

REFRESH

STOP

DELETE

REGIONS

+ 2 RECOMMENDED ALERTS

SHOW INFO PANEL

Jobs on Clusters

Clusters

Jobs

Workflows

Autoscaling policies

Serverless

Batches

Utilities

Component exchange

Metastore

Workbench

Release Notes

Filter

Filter jobs

Job ID

Status

Region

Type

Cluster

Start time

Elapsed time

Labels

job-a0dcc666

Succeeded

us-central1

PySpark

cluster-ea17

Feb 4, 2022, 6:44:38 PM

31 sec

None

job-e1156058

Succeeded

us-central1

PySpark

cluster-ea17

Feb 4, 2022, 6:41:35 PM

30 sec

None

job-747ff136

Succeeded

us-central1

PySpark

cluster-ea17

Feb 4, 2022, 6:34:30 PM

32 sec

None

job-baacaa32

Succeeded

us-central1

PySpark

cluster-ea17

Feb 4, 2022, 6:30:57 PM

33 sec

None

job-13b2694f

Succeeded

us-central1

PySpark

cluster-ea17

Feb 4, 2022, 6:27:52 PM

32 sec

None

job-5250b645

Succeeded

us-central1

PySpark

cluster-ea17

Feb 4, 2022, 6:22:36 PM

31 sec

None

job-c4cbdf8f

Succeeded

us-central1

PySpark

cluster-ea17

Feb 4, 2022, 6:21:00 PM

31 sec

None

job-8db58bde

Failed

us-central1

PySpark

cluster-ea17

Feb 4, 2022, 6:19:02 PM

36 sec

None

job-9c6536c2

Failed

us-central1

PySpark

cluster-ea17

Feb 4, 2022, 6:15:34 PM

36 sec

None

job-f9f031ed

Failed

us-central1

PySpark

cluster-ea17

Feb 4, 2022, 6:12:23 PM

30 sec

None

job-6f079631

Failed

us-central1

PySpark

cluster-ea17

Feb 4, 2022, 6:09:43 PM

34 sec

None

job-1e328843

Failed

us-central1

PySpark

cluster-ea17

Feb 4, 2022, 6:06:50 PM

25 sec

None

job-2aa830cb

Failed

us-central1

PySpark

cluster-ea17

Feb 4, 2022, 6:03:58 PM

31 sec

None

b0fc047359704f0996839336eb7dba2f

Failed

us-central1

PySpark

cluster-ea17

Feb 4, 2022, 6:02:38 PM

26 sec

None

4a14016cb57446c4bf190c733fd7ce23

Failed

us-central1

PySpark

cluster-ea17

Feb 4, 2022, 6:01:32 PM

27 sec

None

job-e99199fd

Failed

us-central1

PySpark

cluster-ea17

Feb 4, 2022, 5:57:54 PM

32 sec

None

job-a1f56eb2

Failed

us-central1

PySpark

cluster-ea17

Feb 4, 2022, 5:56:30 PM

36 sec

None

job-9d35ccb1

Failed

us-central1

PySpark

cluster-ea17

Feb 4, 2022, 5:48:27 PM

34 sec

None

37fb46a1ec104a39bf917f2127539174

Failed

us-central1

PySpark

cluster-ea17

Feb 4, 2022, 5:41:54 PM

1 min 4 sec

None

job-dc7f9ad4

Failed

us-central1

PySpark

cluster-ea17

Feb 4, 2022, 5:20:08 PM

34 sec

None

a30fb69dfacb4abfa9d61c44cfd4d64c

Failed

us-central1

PySpark

cluster-ea17

Feb 4, 2022, 5:17:50 PM

33 sec

None

55c7d3220d7a40fa086fd901e40c611a

Failed

us-central1

PySpark


cluster-ea17

Feb 4, 2022, 5:14:05 PM


25 sec


None


Submit Job: название, регион и кластер


 Dataproc

Jobs on Clusters ^

 Clusters

 **Jobs**

 Workflows

 Autoscaling policies

Serverless ^

← Submit a job

Job ID *
job-1af35f97

Region *
us-central1 ▼
Specifies the Cloud Dataproc regional service, which determines what clusters are available.

Cluster *
cluster-ea17

Submit Job: тип и исполняемые файлы

Job type *

PySpark



Main python file *

gs://dmashkina/bq_analytics_avro.py

Can be a GCS file with the gs:// prefix, an HDFS file on the cluster with the hdfs:// prefix, or a local file on the cluster with the file:// prefix

Additional python files

gs://dmashkina/conftest.py



Enter file path, for example, hdfs://example/exam

NB: исполняемые файлы

Исполняемые файлы, будь то питоновский скрипт или джарник, должны быть загружены в Google Cloud Storage.

Browser [+ CREATE BUCKET](#)

Filter Filter buckets

<input type="checkbox"/>	Name ↑
<input type="checkbox"/>	dataproc-staging-us-central1-346533...
<input type="checkbox"/>	dataproc-temp-us-central1-34653344...
<input type="checkbox"/>	dmashkina



Buckets > dmashkina

[UPLOAD FILES](#) [UPLOAD FOLDER](#) [CREATE FOLDER](#)

Filter by name prefix only ▼

Filter Filter objects and fol

<input type="checkbox"/>	Name	Size
<input type="checkbox"/>	analytics.avro	16.1 MB
<input type="checkbox"/>	bq_analytics_avro.py	1,009 B
<input type="checkbox"/>	conftest.py	244 B
<input type="checkbox"/>	test_bq_analytics_avro.py	2.7 KB

Submit Job: аргументы программы

Подаются как угодно, логика их обработки определяется написанной программой, не Dataproc'ом.

Arguments

test



Press <Return> to add more arguments

Additional arguments to pass to the main class. Press Return after each argument.

Submit Job: дополнительные ресурсы

Jar files

gs://spark-lib/bigquery/spark-bigquery-latest_2.12.jar ✕

Enter file path, for example, hdfs://example/example.jar

Jar files are included in the CLASSPATH. Can be a GCS file with the gs:// prefix, an HDFS file on the cluster with the hdfs:// prefix, or a local file on the cluster with the file:// prefix.

Properties ?

Key 1 *

spark.jars.packages

Value 1


org.apache.spark:spark-avro_2.12:3

+ ADD PROPERTY

Версии компонентов для ОС кластера

- Проверить версию Spark и прочего для установленной ОС:
<https://cloud.google.com/dataproc/docs/concepts/versioning/dataproc-release-2.0>. Может пригодиться при подключении дополнительных библиотек при сабмите джобы.

Статус джобы

 Dataproc

Jobs on Clusters

Clusters

Jobs

Workflows

Autoscaling policies

Serverless

Batches

Utilities

Component exchange

Metastore

Workbench

Release Notes

Job details

CLONEDELETESTOPREFRESH

Job IDjob-c4cbdf8f

Job UUIDf045de59-80b4-4e48-ab4e-b7031473ceb9

TypeDataproc Job

StatusSucceeded

MONITORING

CONFIGURATION

The charts below represent the metrics from the cluster this job ran on, scoped to the time that this job was running. It is possible for more than one job to run on a cluster at a time, so these metrics may not reflect this job's resource usage accurately. Metrics for a job may lag behind the job run by several minutes.

RESET ZOOM

1 hour6 hours12 hours1 day2 days4 days7 days14 days30 days

2/4/22, 6:18 PM - 2/4/22, 6:23 PM

Output

LINE WRAP: OFF

===== test session starts =====

platform linux -- Python 3.8.12, pytest-6.2.5, py-1.11.0, pluggy-1.0.0 -- /opt/conda/default/bin/python

cachedir: .pytest_cache

rootdir: /tmp/job-c4cbdf8f

plugins: cov-3.0.0, anyio-3.5.0

collecting ... 22/02/04 14:21:06 INFO org.apache.spark.SparkEnv: Registering MapOutputTracker

22/02/04 14:21:06 INFO org.apache.spark.SparkEnv: Registering BlockManagerMaster

22/02/04 14:21:06 INFO org.apache.spark.SparkEnv: Registering BlockManagerMasterHeartbeat

22/02/04 14:21:06 INFO org.apache.spark.SparkEnv: Registering OutputCommitCoordinator


22/02/04 14:21:06 INFO org.sparkproject.jetty.util.log: Logging initialized @4162ms to org.sparkproject.jetty.util.log.Slf4jLog

Output is complete

EQUIVALENT COMMAND LINE

15

Конфигурация джобы

 Dataproc

Jobs on Clusters

Clusters

Jobs

Workflows

Autoscaling policies

Serverless

Batches

Utilities

Component exchange

Metastore

Workbench

Release Notes

← Job details

CLONE

DELETE

STOP

REFRESH

Job ID

job-c4cbdf8f

Job UUID

f045de59-80b4-4e48-ab4e-b7031473ceb9

Type

Dataproc Job

Status

✓ Succeeded

MONITORING

CONFIGURATION

EDIT

Start time:

Feb 4, 2022, 6:21:00 PM

Elapsed time:

31 sec

Status:

Succeeded

Region

us-central1

Cluster

[cluster-ea17](#)

Job type

PySpark

Main python file

gs://dmashkina/test_bq_analytics_avro.py

Jar files

gs://spark-lib/bigquery/spark-bigquery-latest_2.12.jar

Additional python files

gs://dmashkina/confest.py

Arguments

test

Labels

Output

LINE WRAP: OFF

↓

Output is complete

EQUIVALENT COMMAND LINE

https://console.cloud.google.com/dataproc/jobs/job-c4cbdf8f/configuration?region=us-central1&orgonly=true&project=dmashkina-ad297c4e&supportedpurview=organizationId

Хочу в консоли как труъ-прогер: создание кластера

```
gcloud dataproc clusters create CLUSTER_NAME --region=REGION
```

```
gcloud dataproc clusters start CLUSTER_NAME --region=REGION
```

Хочу в консоли как труъ-прогер: submit job

```
gcloud dataproc jobs submit pyspark \  
  gs://BUCKET/FILE.py \  
  --cluster=CLUSTER_NAME \  
  --region=REGION \  
  --jars gs://spark-lib/bigquery/spark-bigquery-latest_2.12.jar \  
  --properties spark.jars.packages='org.apache.spark:spark-avro_2.12:3.1.2'  
  -- ARG_1 ARG_2
```

NB: передача аргументов программы

После указания всех аргументов Dataproc и перед началом передачи аргументов непосредственно программы необходимо **поставить два дефиса**, и только потом передавать аргументы. Даже если первый аргумент сам начинается с двух дефисов.

```
--properties spark.jars.
```

```
-- ARG_1 ARG_2
```

Документация по консольным командам

<https://cloud.google.com/sdk/gcloud/reference/dataproc/clusters/create>

<https://cloud.google.com/sdk/gcloud/reference/dataproc/jobs/submit>

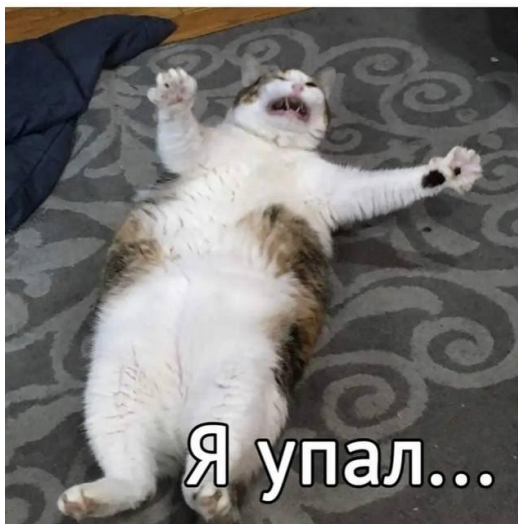
<https://cloud.google.com/sdk/docs/initializing> - первоначальная
настройка gcloud

Общий алгоритм

1. Подключить Dataproc API
2. Создать кластер
3. Запустить его
4. Засабмитить джобу
5. Дождаться окончания ее работы
6. Проверить результат выполнения в логах или на Google Storage

Вопросы?

Что умели
сервера раньше



Что умеют
сервера сейчас



Удачи!

