Spark na Google Cloud

Friends don't let friends build data centers

Data Scientist @Avenue Code

Evandro Caldeira

Cientista de dados na Avenue Code. Formado em Engenharia da Computação e louco por café

E-mail: ecaldeira@avenuecode.com

AvenueCode

Brasil

- Belo Horizonte
- São Paulo
- Porto Alegre

EUA

Canadá

TOC

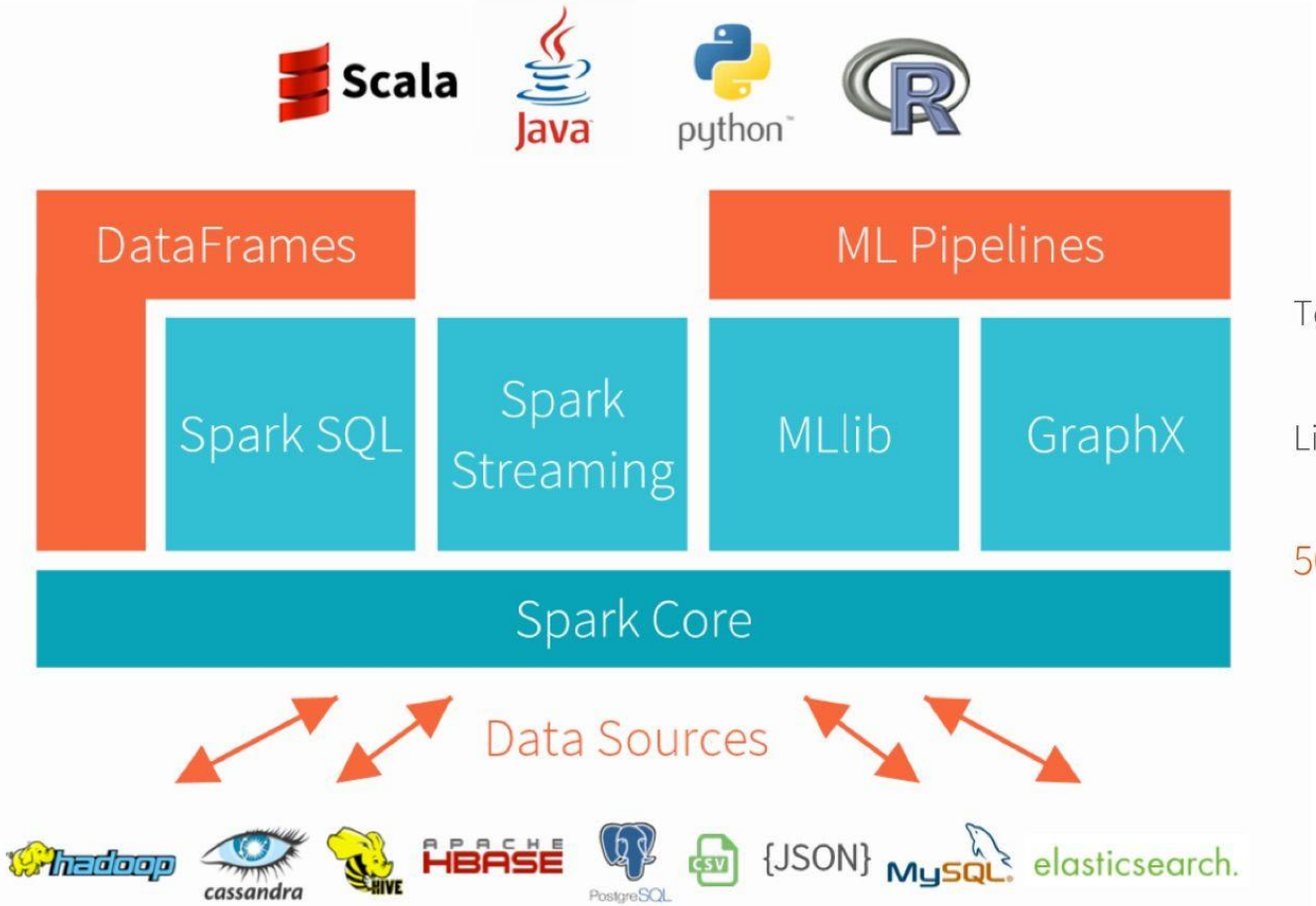
Overview

On premise vs cloud

Como migrar

Demo

Por que Spark?



Total contributors: 150 → 500

Lines of code: 190K → 370K

500+ active production deployments

VS.

On Premise

Cloud

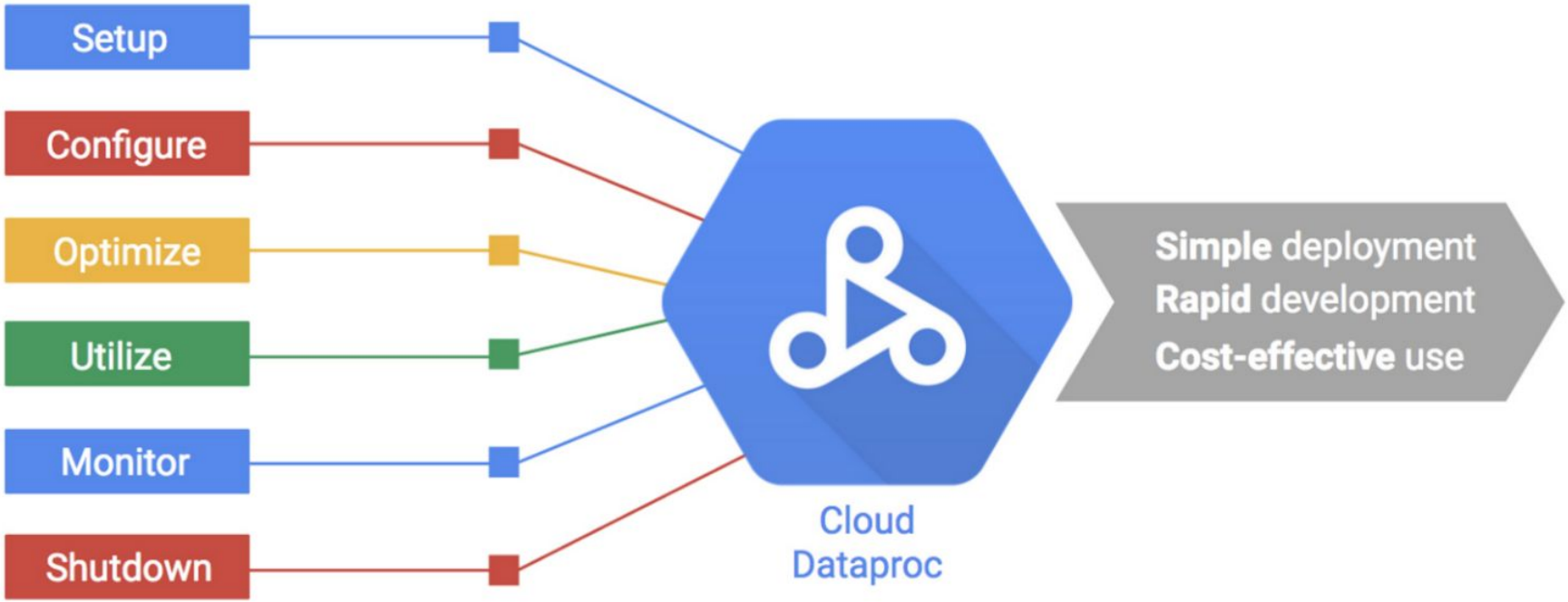
On premise

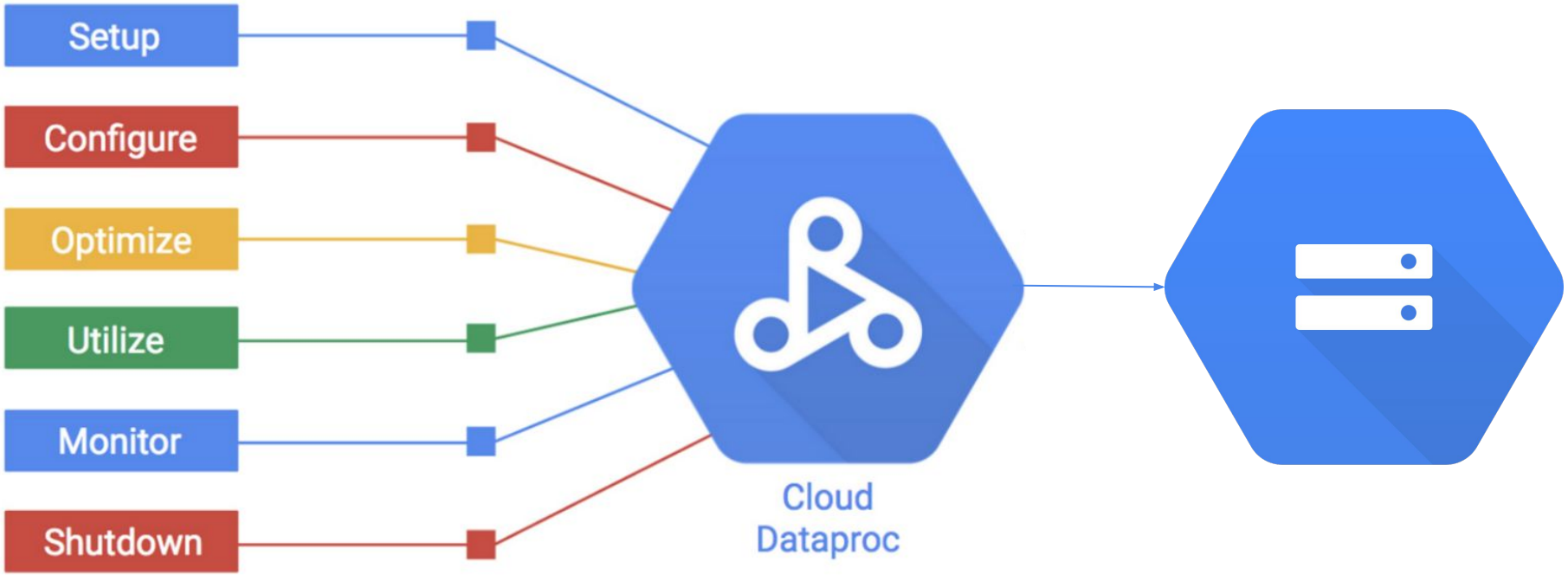
1. Equipamentos
2. Gerenciamento
3. Picos de uso
4. $$$

Spotify

Migragação para GCP:
Descomissionamento de um datacenter em 2018


Primeiros passos
para GCP


Setup
Configure
Optimize
Utilize
Monitor
Shutdown
Cloud
Dataproc
Simple deployment
Rapid development
Cost-effective use


Setup
Configure
Optimize
Utilize
Monitor
Shutdown
Cloud
Dataproc

O que fazer

1. Mova os dados

2. Experimente

O que fazer

3. Use clusters efêmeros
4. Workers preemptivos
5. Delete o cluster ao finalizar


Hands on

Créditos grátis!

12 meses

Crédito de US$ 300 para começar a usar qualquer produto do GCP.

+

Sempre gratuito

Há limites de uso gratuito dos produtos participantes para clientes qualificados, durante e após a avaliação gratuita. Oferta sujeita a alterações.

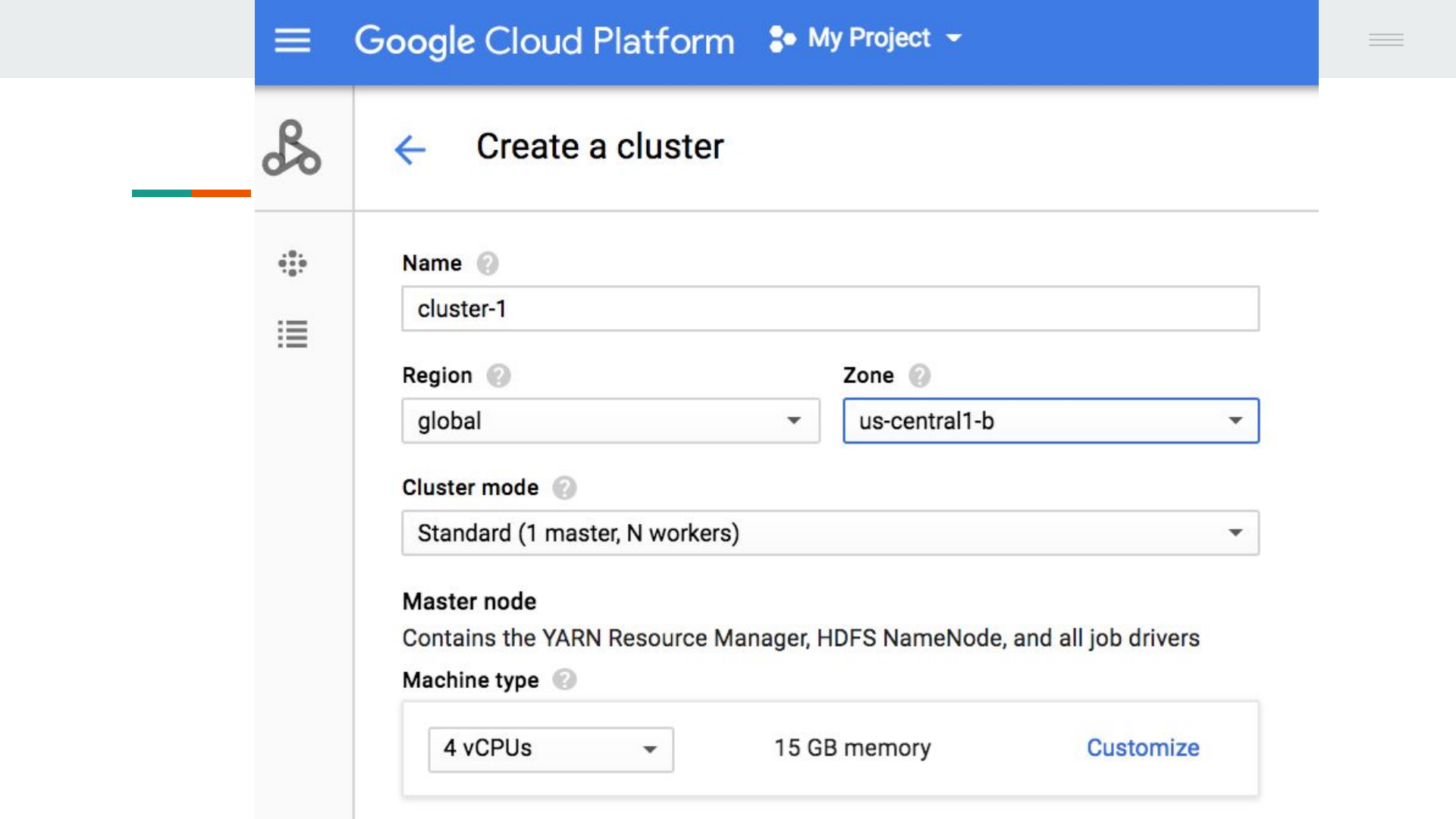
LEIA AS PERGUNTAS FREQUENTES

Instalação

1. Google SDK

2. Spark standalone

Criação do cluster


Google Cloud Platform
My Project
Create a cluster
Name
cluster-1
Region
global
Zone
us-central1-b
Cluster mode
Standard (1 master, N workers)
Master node
Contains the YARN Resource Manager, HDFS NameNode, and all job drivers
Machine type
4 vCPUs
15 GB memory
Customize

Primary disk size (minimum 10 GB)

500 GB

Worker nodes

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

Machine type

4 vCPUs | 15 GB memory | Customize

Primary disk size (minimum 10 GB)

500 GB

Nodes (minimum 2)

2

Local SSDs (0-8)

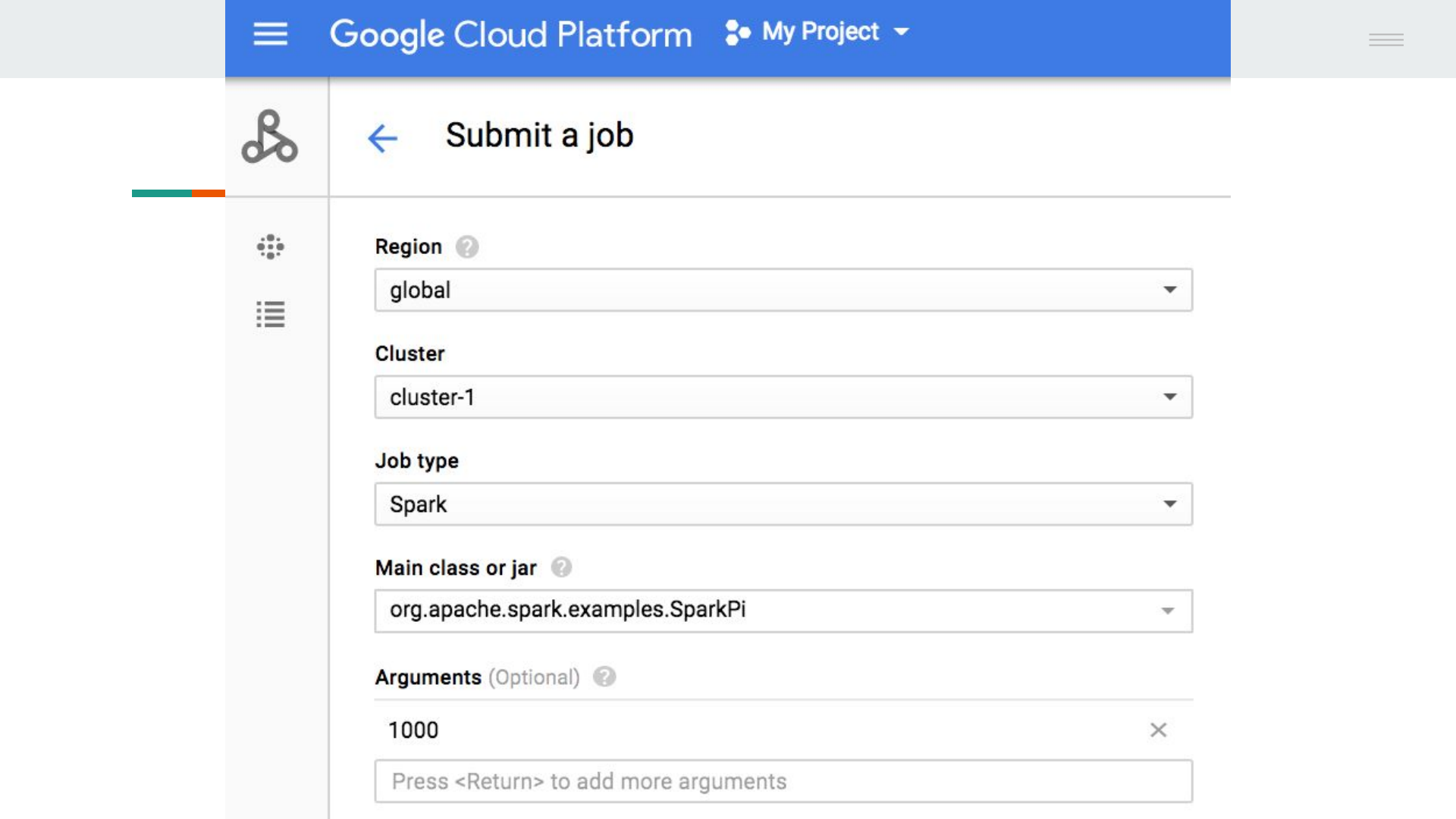
0 x 375 GB

YARN cores
8

YARN memory
24.0 GB

Preemptible workers, bucket, network, version, initialization, & access options

Execução de *job*


Google Cloud Platform
My Project
Submit a job
Region
global
Cluster
cluster-1
Job type
Spark
Main class or jar
org.apache.spark.examples.SparkPi
Arguments (Optional)
1000
Press <Return> to add more arguments

Jar files (Optional)

file:///usr/lib/spark/examples/jars/spark-examples.jar

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

Properties (Optional)

+ Add item

Labels (Optional)

+ Add label

Max restarts per hour (Optional)

Leave blank if you don't want to allow automatic restarts on job failure. Learn more

1-10

Submit Cancel

Equivalent REST

Source

- https://github.com/evandroc/tdc-spark

Obrigado.