**Tarefa 1: preparar o nó mestre e o balde**

Conecte-se ao nó mestre

Um cluster Dataproc foi preparado para você. Você estará executando a maioria das etapas de laboratório do nó mestre do cluster em uma janela do terminal SSH.

1. No Console, no **menu Navegação** ( 7a91d354499ac9f1.png), clique em **Dataproc** > **Clusters** .
2. Localize o cluster denominado cluster de **dataproc** .
3. Clique no nome **dataproc-cluster** para ir para a página Detalhes do cluster.
4. A página de detalhes do cluster é aberta na guia "Visão geral". Clique na guia "Instâncias de VM".
5. Na linha da VM denominada **dataproc-cluster-m** , você verá que ela tem a função de mestre e que há um link SSH ao lado dela. Clique em **SSH** para abrir uma janela de terminal para o nó mestre.

Copie os arquivos de amostra para o diretório inicial do Nó Principal

Os arquivos de amostra necessários já foram arquivados no nó mestre. Você precisará copiá-los em seu diretório de usuários com o seguinte comando.

1. Na janela do terminal SSH do Nó Mestre.

cd

cp -r /training .

ls

1. Agora você deve ter o diretório **/ training** em seu diretório inicial. E deve ter arquivos dentro dele.

Definir a variável de ambiente $ BUCKET

Um intervalo do Cloud Storage já foi criado para você. Tem o mesmo nome que o ID do projeto. Você criará uma variável de ambiente para facilitar a referência do bucket a partir da linha de comando no nó mestre.

1. No console, no **menu Navegação** ( 7a91d354499ac9f1.png), clique em **Armazenamento** > **Navegador** . Localize o intervalo com o mesmo nome que o ID do projeto. Destaque o nome do intervalo e use-o para definir a variável de ambiente.
2. Na janela do terminal SSH do Nó Mestre, digite:

BUCKET=<bucket-name>

echo $BUCKET

Você pode usar $ BUCKET em comandos CLI e recuperar o nome com **echo $ BUCKET** .

**Tarefa 2. O laboratório de duas letras**

Isso é chamado de "laboratório de duas letras", pois envolve apenas adicionar as duas letras "gs" à frente de uma referência de arquivo para instruir o Spark a usar o Cloud Storage em vez do HDFS.

O HDFS é baseado em um white paper publicado em 2003 no Google File Storage (GFS), um dos primeiros ancestrais do Cloud Storage. O Cloud Storage pode ser usado no lugar do HDFS para grande benefício. Os clusters do Dataproc têm o conector para o Cloud Storage já instalado por padrão.

**Por que você deseja usar o Cloud Storage em vez do HDFS?**

* Você pode desligar o cluster quando não estiver executando trabalhos. O armazenamento persiste mesmo quando o cluster é desligado, portanto, você não precisa pagar pelo cluster apenas para manter os dados no HDFS.
* Em alguns casos, o Cloud Storage oferece melhor desempenho que o HDFS.
* O Cloud Storage não requer a sobrecarga de administração de um sistema de arquivos local.

Substitua o HDFS por GS

1. Coloque uma cópia do seu arquivo de dados de amostra em um intervalo do Cloud Storage em vez de no HDFS.
2. Na janela do terminal Nó Principal, digite o seguinte comando gsutil para copiar os arquivos de texto de amostra para o intervalo do Cloud Storage.

gsutil cp /training/road-not-taken.txt gs://$BUCKET

1. No terminal SSH do nó mestre, use **nano** ou **vi** para criar o arquivo **wordcount.py** .
2. Copie e cole o seguinte código no arquivo.

from pyspark.sql import SparkSession

from operator import add

import re

print("Okay Google.")

spark = SparkSession\

.builder\

.appName("CountUniqueWords")\

.getOrCreate()

lines = spark.read.text("/sampledata/road-not-taken.txt").rdd.map(lambda x: x[0])

counts = lines.flatMap(lambda x: x.split(' ')) \

.filter(lambda x: re.sub('[^a-zA-Z]+', '', x)) \

.filter(lambda x: len(x)>1 ) \

.map(lambda x: x.upper()) \

.map(lambda x: (x, 1)) \

.reduceByKey(add) \

.sortByKey()

output = counts.collect()

for (word, count) in output:

print("%s = %i" % (word, count))

spark.stop()

Este aplicativo PySpark realiza contagem de palavras nos dados contidos no HDFS. Observe que o aplicativo atualmente espera encontrar os dados no HDFS em um diretório chamado **/ sampledata** . Você modificará o programa para trabalhar nos dados que você acabou de enviar para o Cloud Storage.

1. Primeiro, verifique se o arquivo de dados não existe no HDFS.

hadoop fs -ls

Não há nenhum diretório **/ sampledata** no HDFS neste cluster.

1. Em seguida, use o comando do sistema de arquivos Hadoop para visualizar os arquivos por meio do conector hadoop no Cloud Storage. Isso verifica se o conector está funcionando e se o arquivo está disponível no bloco.

hadoop fs -ls gs://$BUCKET

1. Edite **wordcount.py** em **nano** ou **vi** .

Substitua esta linha que se refere ao arquivo no HDFS.

lines = spark.read.text("/sampledata/road-not-taken.txt").rdd.map(lambda x: x[0])

Com uma linha, refere-se ao arquivo no Cloud Storage. Lembre-se de remover "/ sampledata" porque esse diretório não existe. Lembre-se de usar o nome do depósito real e não a variável de ambiente. Os Nós Trabalhadores no cluster em que o programa será executado não conhecem o valor da variável de ambiente local no Nó Principal.

lines = spark.read.text("gs://<YOUR-BUCKET>/road-not-taken.txt").rdd.map(lambda x: x[0])

1. Execute o trabalho.

spark-submit wordcount.py

Observe que o aplicativo funciona, embora os dados não estejam localizados no HDFS. O HDFS ainda está sendo usado para arquivos temporários durante o processamento. Os dados de entrada e saída podem estar localizados no Cloud Storage. A razão pela qual isso é possível é que o Google Cloud Network oferece uma transferência de dados super rápida, de forma que o armazenamento fora do cluster seja quase tão rápido e, em alguns casos, mais rápido do que o armazenamento no cluster.

Alavancar esse recurso do GCP torna o cluster sem estado, porque todos os dados persistentes são mantidos fora do cluster. E isso significa que (a) o cluster pode ser desligado quando não estiver em uso, resolvendo o problema de utilização do Hadoop e (b) um cluster pode ser criado e dedicado a um único trabalho, resolvendo o problema de configuração e ajuste do Hadoop.

**Tarefa 3. Executar um aplicativo PySpark no Cloud Storage**

Organize um programa no Cloud Storage e execute-o

Na tarefa anterior, você criou um aplicativo PySpark em um ambiente de desenvolvimento (no nó mestre). Você testou o aplicativo usando o **envio de faísca** .

Nesta tarefa, você migrará o aplicativo do ambiente de desenvolvimento para um ambiente de produção. Você irá preparar o arquivo do aplicativo de trabalho no Cloud Storage. E você executará o trabalho de produção no Console.

1. No terminal Nó Mestre, use o seguinte comando para copiar o aplicativo wordcount.py **PySpark** testado para o bloco.

gsutil cp wordcount.py gs://$BUCKET

1. No Console, no **menu Navegação** ( 7a91d354499ac9f1.png), clique em **Dataproc** > **Clusters** . Anote a região em que o cluster está localizado. Você precisará disso nos próximos passos.
2. Você também precisará do nome do intervalo. Você também pode recuperar o nome do intervalo do terminal do nó mestre, digitando o seguinte. Realce o nome do intervalo e copie-o.

echo $BUCKET

Você também pode encontrar o nome do intervalo no console. No **menu Navegação** ( 7a91d354499ac9f1.png), clique em **Armazenamento** > **Navegador** .

1. No Console, no **menu Navegação** ( 7a91d354499ac9f1.png), clique em **Dataproc** > **Trabalhos** .
2. Clique em **Submit Job** .
3. Especifique o seguinte e deixe as configurações restantes como seus padrões:

|  |  |
| --- | --- |
| **Propriedade** | **Valor**  (digite o valor ou selecione a opção conforme especificado) |
| **Região** | <your-region> |
| **Grupo** | dataproc-cluster |
| **Tipo de emprego** | PySpark |
| **Arquivo principal do Python** | gs://<your bucket>/wordcount.py |

1. Clique em **Submit** .
2. Você pode ver o progresso na página **Dataproc** > **Jobs** no Console.

**Termine seu laboratório**

Quando você tiver concluído seu laboratório, clique em **Finalizar Laboratório** . O Qwiklabs remove os recursos que você usou e limpa a conta para você.

Você terá a oportunidade de avaliar a experiência do laboratório. Selecione o número de estrelas aplicável, digite um comentário e clique em **Enviar** .

O número de estrelas indica o seguinte:

* 1 estrela = muito insatisfeito
* 2 estrelas = insatisfeito
* 3 estrelas = neutra
* 4 estrelas = Satisfeito
* 5 estrelas = muito satisfeito

Você pode fechar a caixa de diálogo se não quiser fornecer feedback.

Para comentários, sugestões ou correções, use a guia **Suporte** .

Última atualização: 2018-09-12

© 2018 Google LLC Todos os direitos reservados. Google e o logotipo do Google são marcas registradas do Google LLC. Todos os outros nomes de empresas e produtos podem ser marcas registradas das respectivas empresas com as quais estão associados.