**Tarefa 1: preparação**

Um cluster Dataproc foi preparado para você. Se você fizer login no GCP antes de a barra de progresso informar que o "Laboratório está em execução", talvez seja necessário esperar alguns minutos para que o cluster faça a transição de "Fornecimento" para "Em execução" antes que o cluster conclua a configuração.

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** . Em qual região e zona ele está localizado? A região e a zona foram selecionadas automaticamente para você pela Qwiklabs.
3. Observe o intervalo de armazenamento temporário do Cloud Storage definido para este cluster. Esse intervalo tem o mesmo nome que o ID do projeto, que é uma maneira conveniente de tornar o nome globalmente exclusivo.
4. Clique no nome **dataproc-cluster** para ir para a página Detalhes do cluster.
5. A página de detalhes do cluster é aberta na guia **Visão geral** . Clique na guia denominada **instâncias de VM** .

Abra o terminal do nó mestre

1. 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.

**Tarefa 2: habilitar o acesso seguro à Web ao cluster do Dataproc**

Criar uma regra de firewall restritiva usando tags de destino, endereço IP e protocolo

Crie uma regra de firewall que permita acesso somente ao nó mestre do endereço IP do seu computador. Somente as portas 8088 (Interface de trabalho do Hadoop) e 9870 (interface do Administrador do Hadoop) serão permitidas.

Verifique se a tag de rede está definida no nó mestre

Verifique se a tag de rede **hadoopaccess** está configurada no nó mestre. Isso aplicará a regra de firewall ao nó mestre, dando acesso ao seu laptop.

1. No Console, no **menu Navegação** ( 7a91d354499ac9f1.png), clique em **Compute Engine** > **Instâncias de VM** .
2. Clique no nó mestre, **cluster-dataproc-m** .
3. Verifique se, em Network Tags, ele lista **hadoopaccess** .
4. Se a tag não estiver lá, clique em **EDITAR** .
5. Em Network Tags, adicione a tag: **hadoopaccess**
6. Clique em **Salvar** .

Identifique o endereço IP do navegador

Você usará o endereço IP do navegador para permitir que seu navegador local se conecte ao cluster Dataproc.

1. Encontre o endereço IP do navegador do seu computador abrindo uma janela do navegador e visualizando <http://ip4.me/> Copie o endereço IP.

Crie a regra de firewall

1. No console, no **menu Navegação** ( 7a91d354499ac9f1.png), clique em **Rede VPC** > **Regras de firewall** .
2. Clique em **Criar regra de firewall** .
3. Especifique o seguinte e deixe as configurações restantes como seus padrões:

|  |  |
| --- | --- |
| **Propriedade** | **Valor**  (digite o valor ou selecione a opção conforme especificado) |
| **Nome** | allow-hadoop |
| **Rede** | default |
| **Prioridade** | **1000** |
| **Direção de tráfego** | Ingress |
| **Ação no jogo** | Allow |
| **Alvos** | Specified target tags |
| **Tags alvo** | hadoopaccess |
| **Intervalos de IP de origem** | <your-IP>/32 |
| **Protocolos e portas especificados** | Verifique tcpe insira o número da porta9870,8088 |

1. Clique em **Criar** .

Levará alguns minutos para que a regra de firewall se torne ativa.

**Tarefa 3: Preparar os dados para o Hive**

Copie os arquivos de amostra para o diretório inicial do nó mestre

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. Na janela do terminal SSH do Nó Mestre. Agora você deve ter o diretório / training em seu diretório inicial. E deve ter arquivos dentro dele.

ls

cd training/training-data-analyst/courses/unstructured

ls pet\*.\*

Copie o arquivo de dados para o HDFS

1. Veja os dados estruturados no arquivo de texto.

cat pet-details.txt

Os dados são delimitados por vírgulas. Está organizado em campos e linhas.

1. Coloque os dados no HDFS.

hadoop fs -mkdir /pet-details

hadoop fs -put pet-details.txt /pet-details

Visualizar os arquivos de dados usando a interface de administração do Hadoop

1. No Console, no **menu Navegação** ( 7a91d354499ac9f1.png), clique em **Compute Engine** > **Instâncias de VM** .
2. Na lista de instâncias de VMs, na linha do **cluster-dataproc-m** , destaque o **IP externo** e copie-o.
3. Abra uma nova aba ou janela do navegador e cole o IP externo. Adicione ": 9870" após o IP e pressione Enter. Exemplo:<External IP>:9870
4. Agora você deve ver a interface de administração do Hadoop. Em **Utilitários** , clique em **Navegar no sistema de arquivos** . Clique na pasta **/ detalhes do animal de estimação** .
5. Observe que o arquivo **pet-details.txt** está dentro de **/ pet-detalhes** .
6. Deixe a interface de administração do Hadoop aberta. Você retornará em etapas posteriores.

**Tarefa 4: Explorar o Hive usando a CLI interativa do Hive**

Use o HIVE para acessar os dados no HDFS como se estivesse em um banco de dados

O Hive fornece um subconjunto do SQL. A maneira de fazer isso é mantendo os metadados para definir um esquema sobre os dados. Essa é uma maneira de trabalhar com uma grande quantidade de dados distribuídos no HDFS usando a sintaxe familiar do SQL.

1. Na janela SSH do nó mestre, verifique se você está no diretório correto e inicie o interpretador Hive CLI.

hive

1. No intérprete de seção, crie um banco de dados. Isso cria o esquema de metadados para um banco de dados.

CREATE DATABASE pets;

USE pets;

1. Crie uma tabela. Observe que o delimitador de vírgula é especificado na criação da tabela, não na entrada de dados.

CREATE TABLE details (Type String, Name String, Breed String, Color String, Weight Int) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';

SHOW TABLES;

DESCRIBE pets.details;

1. Estabeleça o relacionamento entre o esquema de metadados e os dados no HDFS.

load data INPATH '/pet-details/pet-details.txt' OVERWRITE INTO TABLE details;

1. Verifique se tudo está funcionando.

SELECT \* FROM pets.details;

1. Saia do interpretador de colmeia.

quit;

Use a interface de administração do Hadoop para ver como funciona a seção

O Hive ingeriu o arquivo **pet-details.txt** em um formato de data warehouse que requer um esquema. Você usará a interface de administração do Hadoop para ver essa transformação.

1. Retorne à interface de administração do Hadoop no navegador.
2. Em **Utilitários** , clique em **Navegar no sistema de arquivos** . Clique na pasta **/ detalhes do animal de estimação** . O arquivo **pet-details.txt** sumiu.
3. Em **Utilitários** , clique em **Navegar no sistema de arquivos** . Em seguida, clique em **user> hive> warehouse> pets.db** > **details** . O arquivo **pet-details.txt** foi movido para este local.

O Hive foi projetado para trabalhos em lote e não para transações. Ele insere dados em um formato de data warehouse que requer um esquema. Ele não suporta consultas em tempo real, atualizações em nível de linha ou dados não estruturados. Algumas consultas podem ser muito mais lentas do que outras devido às transformações subjacentes que o Hive deve implementar para simular o SQL.

**Tarefa 5: Executar um trabalho do Pig**

Executar um trabalho de porco

1. Na janela SSH do nó mestre, visualize o aplicativo Pig.

cat pet-details.pig

2 Na linha 'x1', a instrução load no aplicativo cria um esquema na parte superior do arquivo de dados HDFS. Linhas de 'x2' a 'x5' realizam transformações nos dados. E a última linha armazena o resultado em uma pasta chamada **/ GroupedByType** no HDFS.

1. O aplicativo espera encontrar o arquivo de entrada no HDFS no diretório / **pet-details** . Faça outra cópia dos dados nesse local.

hadoop fs -put pet-details.txt /pet-details

1. Execute o aplicativo:

pig < pet-details.pig

1. Retorne à guia do navegador que contém a interface do Hadoop Applications e atualize-a ou reabra-a com <External-IP>:8088. Observe que Pig gerou uma tarefa Java MapReduce que está sendo executada no cluster. Clique no botão de atualização do navegador para observar a conclusão do trabalho.
2. Retorne à guia do navegador que contém a interface de administração do Hadoop e atualize-a ou reabra-a com <External-IP>:9870. Em **Utilitários** , clique em **Navegar no sistema de arquivos** . Na lista resultante, clique em **GroupedByType** . Este é o diretório de saída especificado no aplicativo Pig. O arquivo chamado **part-r-00000** é o arquivo HDFS que contém a saída. Você não pode ver o conteúdo daqui. Primeiro, você deve baixar essa parte para o sistema de arquivos local
3. Retorne ao terminal SSH no nó mestre, cloud-dataproc-m, crie um diretório de saída local e recupere os resultados do HDFS.

cd

mkdir output

cd output

hadoop fs -get /GroupedByType/part\* .

1. Agora você pode ver os resultados do trabalho Pig.

cat part-r-00000

Pig fornece primitivos SQL similares ao Hive, mas em um formato de linguagem de script mais flexível. Pig também pode lidar com dados semi-estruturados, como dados com esquemas parciais, ou para os quais o esquema ainda não é conhecido. Por esse motivo, às vezes é usado para extrair carga de transformação (ETL). Ele gera trabalhos MapReduce Java. O Pig não foi projetado para lidar com dados não estruturados.