**Instruções**

**1) Baixar os arquivos das seguintes URLs:**

Jul 01 to Jul 31, ASCII format, 20.7 MB gzip

Aug 04 to Aug 31, ASCII format, 21.8 MB gzip

**2) Criar diretório no HFDS**

hdfs dfs -mkdir -p /user/HOME/teste/

**3) Copiar os arquivos baixados para o diretório HFS criado no passo 2**

hdfs dfs -put NASA\_access\_log\_Jul95.gz /user/HOME/teste/

hdfs dfs -put NASA\_access\_log\_Aug95.gz /user/HOME/teste/

**4) Gerar tabela externa, utilizando o diretório HDFS do passo 3, conforme segue:**

drop table owner.HOME\_teste;

create external table onwer.HOME\_teste (

coluna\_1 string,

coluna\_2 string,

coluna\_3 string,

coluna\_4 string,

coluna\_5 string,

coluna\_6 string,

coluna\_7 string,

coluna\_8 string,

coluna\_9 string,

coluna\_10 string

) ROW FORMAT DELIMITED FIELDS TERMINATED BY ' ' STORED AS textfile LOCATION '/user/HOME/teste';

**4) Gerar tabela HIVE com as coluna que serão utilizadas na solução, conforme segue:**

create table owner.HOME\_base as

select coluna\_1 as host,

concat(coluna\_4, coluna\_5) as dt\_requisicao,

concat(coluna\_6, coluna\_7, coluna\_8) as desc\_requisicao,

coluna\_9 as retorno\_HTTP,

cast(coluna\_10 as BIGINT) as tot\_bytes\_ret

from owner.HOME\_teste;

**5) Entrar no ambiente Spark.**

**6) Executar o script Teste.py para obter nos logs as respostas do desafio.**

****

**7) Logs obtidos na execução do Teste.py**

****