O sistema do DockerHub removeu a aba do Dockerfile disponibilizado anteriormente durante a gravação da aula do professor Takeshi. Além disso algumas atualizações foram feitas na imagem e que podem apresentar alguns problemas durante a instalação dela.

Fizemos um fork do repositório original e correções no Dockerfile. Algumas correções foram realizadas devido a alterações nos repositórios do Debian durante o processo de update dos pacotes. Outro problema corrigido é em relação a expiração do certificado da página do PostgreSQL durante o download do driver JDBC do PostgreSQL, o que gerava erro ao baixar. Foi incluída a flag --no-check-certificate no Dockerfile para desabilitar a checagem do certificado e o download ser realizado.

Assim, elaboramos esse tutorial para que todos possam executar sem maiores problemas as práticas do professor Takeshi.

A principal orientação é pular o comando "docker pull bde2020/hive" do professor Takeshi e, antes do "docker build -t "hive:hive". ", copiar localmente (com o comando git clone) o repositório que possui o Dockerfile corrigido, localizado em https://github.com/tiagoferreto/docker-hive.git.

Depois, pode iniciar o build da imagem com o docker build e seguir o roteiro.

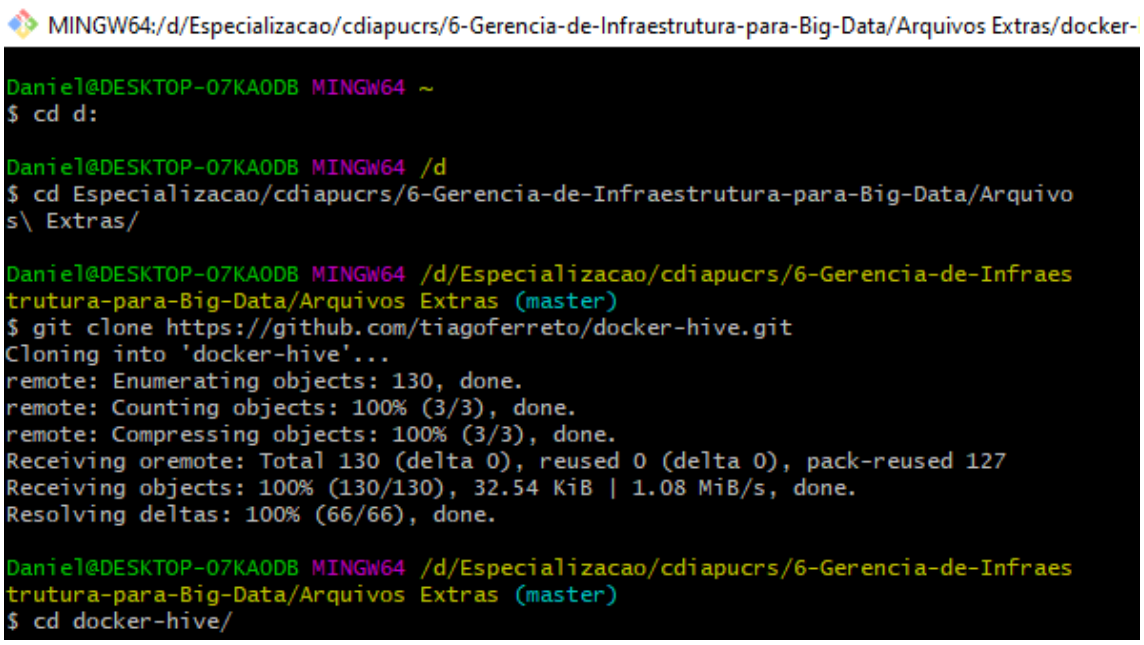
Abaixo é apresentado um detalhamento das etapas.

1) Abra um terminal e faça o clone\download do repositório que possui os arquivos para executar o Hive usando o Docker.

git clone https://github.com/tiagoferreto/docker-hive.git

2) Antes de efetuar o build, entre dentro da pasta que acabamos de clonar

cd docker-hive



3) Em seguida execute o comando de build para montar as imagens necessárias

docker build -t "hive:hive".

```
Daniel@DESKTOP-O7KAODB MINGW64 /d/Especializacao/cdiapucrs/6-Gerencia-de-Infraestrutura-para-Big-Data/Arquivos Extras/docker-hive (master)
$ docker build -t "hive:hive" .
#2 [internal] load .dockerignore
#2 sha256:ff726069f46b12ca39eee1d922d826d124967800622252cf216fe37ed1f93b6f
#2 transferring context: 2B done
#2 ...

#1 [internal] load build definition from Dockerfile
#1 sha256:d72e53a20b2836221cd33051cd84aec8a3b4da4034d0826f356c10a2b8d786af
#1 transferring dockerfile: 2.24kB 0.1s done
#1 DONE 0.3s

#2 [internal] load .dockerignore
#2 sha256:ff726069f46b12ca39eee1d922d826d124967800622252cf216fe37ed1f93b6f
#2 DONE 0.3s

#3 [internal] load metadata for docker.io/bde2020/hadoop-base:2.0.0-hadoop2.7.4-java8
#3 sha256:b7a0feab1d9dc0888fa955473b9ba2993b77b13b49438c2769b1f11c8bb929a5
#3 ...

#4 [auth] bde2020/hadoop-base:pull token for registry-1.docker.io
#4 sha256:9575a4674a66130c87a5cb969d624eb07c96ef075a6544179cc4eab43455e203
#4 DONE 0.0s

#3 [internal] load metadata for docker.io/bde2020/hadoop-base:2.0.0-hadoop2.7.4-java8
#3 sha256:b7a0feab1d9dc0888fa955473b9ba2993b77b13b49438c2769b1f11c8bb929a5
#3 DONE 7.4s

#5 [ 1/18] FROM docker.io/bde2020/hadoop-base:2.0.0-hadoop2.7.4-java8@sha256:db63b313d57fc230287f6beed3e083e19e413a93ea2619a3a26f6073be7ff70
#5 sha256:5e131d147595cd95ca8a12a8888a1aa49baa4ac5c2edc55df4a8dec4be3c7bf9
#5 DONE 0.0s

#6 [ 2/18] WORKDIR /opt
#6 sha256:74269f686afb568f37d47a9ad84173593f372479b38ccb6ce56c73c76f7b4fb9
#6 CACHED

#12 [internal] load build context
#12 sha256:008b797bed2867138e630d442ecd44a25b4d059584eb76952975f65510a368e4
#12 transferring context: 21.34kB 0.1s done
#12 DONE 0.6s
```

3.1) É possível listar todas as imagens contidas no Docker através do seguinte comando:

docker image ls

```
Daniel@DESKTOP-O7KAODB MINGW64 /d/Especializacao/cdiapucrs/6-Gerencia-de-Infraestrutura-para-Big-Data/Arquivos Extras/docker-hive (master)
$ docker image ls
REPOSITORY   TAG       IMAGE ID       CREATED          SIZE
hive         hive      c126a9cea856   49 seconds ago   1.42GB
hadoopimg    latest    897a1aaba254   7 months ago     1.82GB
ubuntu       18.04     39a8cfeef173   7 months ago     63.1MB
```

4) Para iniciar os containers execute o comando abaixo:

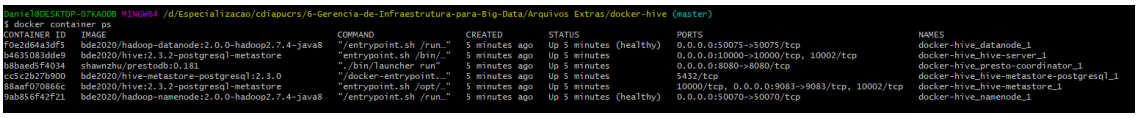
docker-compose up -d

```
Daniel@DESKTOP-O7KAODB MINGW64 /d/Especializacao/cdiapucrs/6-Gerencia-de-Infraestrutura-para-Big-Data/Arquivos Extras/docker-hive (master)
$ docker image ls
REPOSITORY   TAG       IMAGE ID       CREATED          SIZE
hive         hive      c126a9cea856   49 seconds ago   1.42GB
hadoopimg    latest    897a1aaba254   7 months ago     1.82GB
ubuntu       18.04     39a8cfeef173   7 months ago     63.1MB

Daniel@DESKTOP-O7KAODB MINGW64 /d/Especializacao/cdiapucrs/6-Gerencia-de-Infraestrutura-para-Big-Data/Arquivos Extras/docker-hive (master)
$ docker-compose up -d
Creating network "docker-hive_default" with the default driver
Pulling namenode (bde2020/hadoop-namenode:2.0.0-hadoop2.7.4-java8)...
2.0.0-hadoop2.7.4-java8: Pulling from bde2020/hadoop-namenode
Digest: sha256:54a9482c51d4e701e530f15ef2e01ca2d3a15545760c42ea6ad0e65e8196c335
Status: Downloaded newer image for bde2020/hadoop-namenode:2.0.0-hadoop2.7.4-java8
Pulling datanode (bde2020/hadoop-datanode:2.0.0-hadoop2.7.4-java8)...
2.0.0-hadoop2.7.4-java8: Pulling from bde2020/hadoop-datanode
Digest: sha256:5623fca5e36d890983cdc6cfd29744d1d65476528117975b3af6a80d99b3c62f
Status: Downloaded newer image for bde2020/hadoop-datanode:2.0.0-hadoop2.7.4-java8
Pulling hive-server (bde2020/hive:2.3.2-postgresql-metastore)...
2.3.2-postgresql-metastore: Pulling from bde2020/hive
Digest: sha256:620267768985bb57e52a86db9263a354e92d0202319d8356788852539b21e0895
Status: Downloaded newer image for bde2020/hive:2.3.2-postgresql-metastore
Pulling hive-metastore-postgresql (bde2020/hive-metastore-postgresql:2.3.0)...
2.3.0: Pulling from bde2020/hive-metastore-postgresql
Digest: sha256:9ab91699d15131b874829e6572006cd9d9f1cca413f438b6f21c65b412152bf1
Status: Downloaded newer image for bde2020/hive-metastore-postgresql:2.3.0
Pulling presto-coordinator (shawnzhu/prestodb:0.181)...
0.181: Pulling from shawnzhu/prestodb
Digest: sha256:74237b72a6f058dafc282e535dde93f470ac6e4e729fa0fe5342e494d5c4b7c4
Status: Downloaded newer image for shawnzhu/prestodb:0.181
Creating docker-hive_hive-metastore-postgresql_1 ...
Creating docker-hive_hive-server_1               ...
Creating docker-hive_hive-metastore_1            ...
Creating docker-hive_datanode_1                  ...
Creating docker-hive_namenode_1                  ...
Creating docker-hive_presto-coordinator_1        ...
Creating docker-hive_hive-metastore-postgresql_1 ... done
Creating docker-hive_namenode_1                  ... done
Creating docker-hive_hive-server_1               ... done
Creating docker-hive_datanode_1                  ... done
Creating docker-hive_hive-metastore_1            ... done
Creating docker-hive_presto-coordinator_1        ... done
```

5) Após todos os containers estiverem com status “done”, podemos listar todos os containers em execução com o comando:

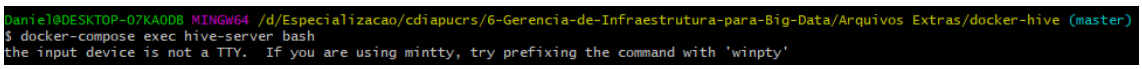
docker container ps



6) Para acessar o container principal (hive-server) é utilizado o seguinte comando:

docker-compose exec hive-server bash

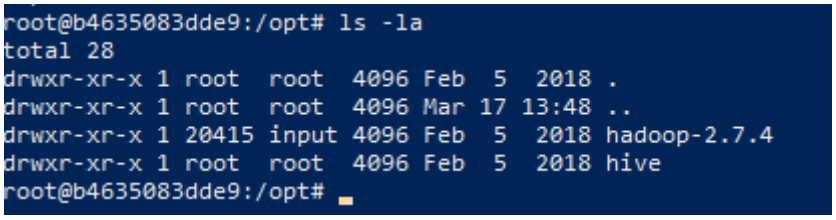
Dependendo do terminal que possa estar usando, ele pode indicar uma mensagem de erro



Caso essa mensagem apareça, utilize outro terminal, como por exemplo o Power Shell

6.1) Após conseguir acessar o container, é possível listar os arquivos que estão dentro do container utilizando o comando:

ls -la



6.2) Para sair do container, basta executar o comando:

exit

7) Prática 1 – Prof. Takeshi

Agora testamos alguns conteúdos da aula prática do professor Takeshi para o tutorial:

7.1) Desligando os containers: irá parar todos os containers, mas não os deletará.

docker-compose down

```
PS D:\Especializacao\cdiapucrs\6-Gerencia-de-Infraestrutura-para-Big-Data\Arquivos Extras\docker-hive> docker-compose down
Stopping docker-hive_datanode_1                 ... done
Stopping docker-hive_hive-server_1              ... done
Stopping docker-hive_presto-coordinator_1       ... done
Stopping docker-hive_hive-metastore-postgresql_1 ... done
Stopping docker-hive_hive-metastore_1           ... done
Stopping docker-hive_namenode_1                 ... done
Removing docker-hive_datanode_1                 ... done
Removing docker-hive_hive-server_1              ... done
Removing docker-hive_presto-coordinator_1       ... done
Removing docker-hive_hive-metastore-postgresql_1 ... done
Removing docker-hive_hive-metastore_1           ... done
Removing docker-hive_namenode_1                 ... done
Removing network docker-hive_default
PS D:\Especializacao\cdiapucrs\6-Gerencia-de-Infraestrutura-para-Big-Data\Arquivos Extras\docker-hive> docker container ps
CONTAINER ID   IMAGE     COMMAND   CREATED   STATUS    PORTS     NAMES
PS D:\Especializacao\cdiapucrs\6-Gerencia-de-Infraestrutura-para-Big-Data\Arquivos Extras\docker-hive>
```

7.2) Listar documentos, criar diretórios, mover arquivos, abrir arquivos e alterar permissões de acesso de arquivos no Linux:

Windows PowerShell

```
root@2aba4f3ec24b:/opt# ls -lha /opt/hive/examples/files/kv1
kv1.seq                kv1.string-sorted.txt  kv1.txt                kv1.val.sorted.txt     kv10.txt
root@2aba4f3ec24b:/opt# ls -lha /opt/hive/examples/files/kv1
kv1.seq                kv1.string-sorted.txt  kv1.txt                kv1.val.sorted.txt     kv10.txt
root@2aba4f3ec24b:/opt# ls -lha /opt/hive/examples/files/kv1.txt
-rw-r--r-- 1 root staff 5.7K Nov  9  2017 /opt/hive/examples/files/kv1.txt
root@2aba4f3ec24b:/opt# pwd
/opt
root@2aba4f3ec24b:/opt# mkdir dreyes
root@2aba4f3ec24b:/opt# ls -la
total 32
drwxr-xr-x 1 root  root  4096 Mar 17 14:39 .
drwxr-xr-x 1 root  root  4096 Mar 17 14:35 ..
drwxr-xr-x 2 root  root  4096 Mar 17 14:39 dreyes
drwxr-xr-x 1 20415 input 4096 Feb  5  2018 hadoop-2.7.4
drwxr-xr-x 1 root  root  4096 Feb  5  2018 hive
root@2aba4f3ec24b:/opt# cp /opt/hive/examples/files/kv1.txt ./dreyes/
root@2aba4f3ec24b:/opt# ls -lha
total 32K
drwxr-xr-x 1 root  root  4.0K Mar 17 14:39 .
drwxr-xr-x 1 root  root  4.0K Mar 17 14:35 ..
drwxr-xr-x 2 root  root  4.0K Mar 17 14:40 dreyes
drwxr-xr-x 1 20415 input 4.0K Feb  5  2018 hadoop-2.7.4
drwxr-xr-x 1 root  root  4.0K Feb  5  2018 hive
root@2aba4f3ec24b:/opt# ls -lha ./dreyes/
total 20K
drwxr-xr-x 2 root root 4.0K Mar 17 14:40 .
drwxr-xr-x 1 root root 4.0K Mar 17 14:39 ..
-rw-r--r-- 1 root root 5.7K Mar 17 14:40 kv1.txt
root@2aba4f3ec24b:/opt# cd dreyes/
root@2aba4f3ec24b:/opt/dreyes# mv kv1.txt kv2.txt
root@2aba4f3ec24b:/opt/dreyes# ls -la
total 20
drwxr-xr-x 2 root root 4096 Mar 17 14:41 .
drwxr-xr-x 1 root root 4096 Mar 17 14:39 ..
-rw-r--r-- 1 root root 5812 Mar 17 14:40 kv2.txt
root@2aba4f3ec24b:/opt/dreyes# cat kv2.txt | head
238@val_238
86@val_86
311@val_311
27@val_27
165@val_165
409@val_409
255@val_255
278@val_278
98@val_98
484@val_484
root@2aba4f3ec24b:/opt/dreyes# ls
kv2.txt
root@2aba4f3ec24b:/opt/dreyes# ls -h
kv2.txt
root@2aba4f3ec24b:/opt/dreyes# ls -hl
total 8.0K
-rw-r--r-- 1 root root 5.7K Mar 17 14:40 kv2.txt
root@2aba4f3ec24b:/opt/dreyes# chmod 775 kv2.txt
root@2aba4f3ec24b:/opt/dreyes# ls -l
total 8
-rwxrwxr-x 1 root root 5812 Mar 17 14:40 kv2.txt
root@2aba4f3ec24b:/opt/dreyes#
```

8) Prática 2 – Prof. Takeshi

Alguns comandos para utilizar o HDFS

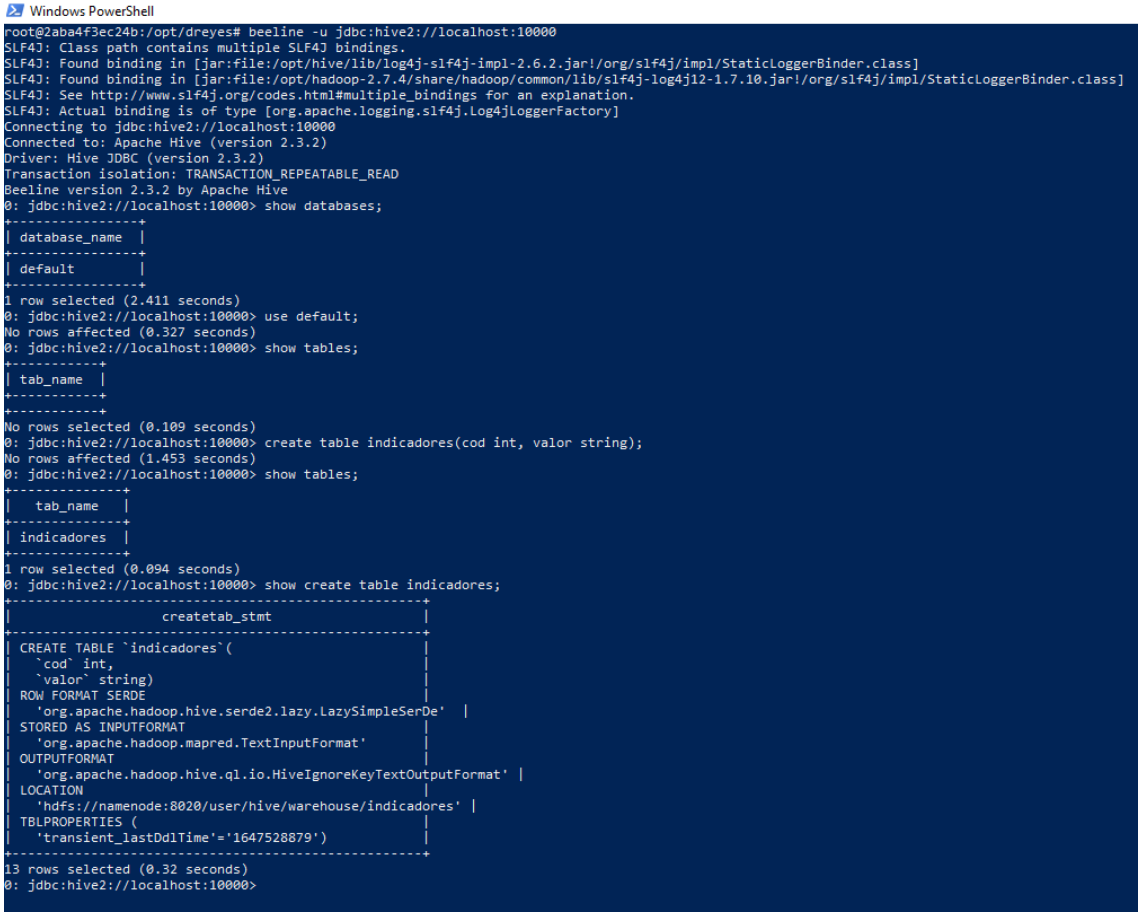
Listar arquivos no HDFS, criar arquivos, inserir arquivos no HDFS, abrir, copiar e mover arquivos no HDFS.

Windows PowerShell

```
root@2aba4f3ec24b:/opt/dreyes# hdfs dfs -ls /
Found 2 items
drwxrwxr-x   - root supergroup          0 2022-03-16 19:49 /tmp
drwxr-xr-x   - root supergroup          0 2022-03-16 19:49 /user
root@2aba4f3ec24b:/opt/dreyes# hdfs dfs -ls /user
Found 1 items
drwxr-xr-x   - root supergroup          0 2022-03-16 19:49 /user/hive
root@2aba4f3ec24b:/opt/dreyes# hdfs dfs -mkdir /user/dreyes
root@2aba4f3ec24b:/opt/dreyes# hdfs dfs -ls /user
Found 2 items
drwxr-xr-x   - root supergroup          0 2022-03-17 14:44 /user/dreyes
drwxr-xr-x   - root supergroup          0 2022-03-16 19:49 /user/hive
root@2aba4f3ec24b:/opt/dreyes# hdfs dfs -put ./kv2.txt /user/dreyes/kv3.txt
root@2aba4f3ec24b:/opt/dreyes# hdfs dfs -ls /user/dreyes
Found 1 items
-rw-r--r--   3 root supergroup       5812 2022-03-17 14:45 /user/dreyes/kv3.txt
root@2aba4f3ec24b:/opt/dreyes# hdfs dfs -cat /user/takeshi/kv3.txt | head
cat: `/user/takeshi/kv3.txt': No such file or directory
root@2aba4f3ec24b:/opt/dreyes# hdfs dfs -cat /user/dreyes/kv3.txt | head
238@val_238
86@val_86
311@val_311
27@val_27
165@val_165
409@val_409
255@val_255
278@val_278
98@val_98
484@val_484
root@2aba4f3ec24b:/opt/dreyes# hdfs dfs -mkdir /user/dreyes/teste
root@2aba4f3ec24b:/opt/dreyes# hdfs dfs -ls /user/dreyes
Found 2 items
-rw-r--r--   3 root supergroup       5812 2022-03-17 14:45 /user/dreyes/kv3.txt
drwxr-xr-x   - root supergroup          0 2022-03-17 14:47 /user/dreyes/teste
root@2aba4f3ec24b:/opt/dreyes# hdfs dfs -cp /user/dreyes/kv3.txt /user/dreyes/teste/
root@2aba4f3ec24b:/opt/dreyes# hdfs dfs -ls /user/dreyes/teste
Found 1 items
-rw-r--r--   3 root supergroup       5812 2022-03-17 14:48 /user/dreyes/teste/kv3.txt
root@2aba4f3ec24b:/opt/dreyes# hdfs dfs -ls /user/dreyes
Found 2 items
-rw-r--r--   3 root supergroup       5812 2022-03-17 14:45 /user/dreyes/kv3.txt
drwxr-xr-x   - root supergroup          0 2022-03-17 14:48 /user/dreyes/teste
root@2aba4f3ec24b:/opt/dreyes#
```

9) Prática 3 – Prof. Takeshi

Comandos via Beeline e Hive


Windows PowerShell
root@2aba4f3ec24b:/opt/dreyes# beeline -u jdbc:hive2://localhost:10000
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/opt/hive/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/opt/hadoop-2.7.4/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Connecting to jdbc:hive2://localhost:10000
Connected to: Apache Hive (version 2.3.2)
Driver: Hive JDBC (version 2.3.2)
Transaction isolation: TRANSACTION_REPEATABLE_READ
Beeline version 2.3.2 by Apache Hive
0: jdbc:hive2://localhost:10000> show databases;
+----------------+
| database_name |
+----------------+
| default |
+----------------+
1 row selected (2.411 seconds)
0: jdbc:hive2://localhost:10000> use default;
No rows affected (0.327 seconds)
0: jdbc:hive2://localhost:10000> show tables;
+-----------+
| tab_name |
+-----------+
+-----------+
No rows selected (0.109 seconds)
0: jdbc:hive2://localhost:10000> create table indicadores(cod int, valor string);
No rows affected (1.453 seconds)
0: jdbc:hive2://localhost:10000> show tables;
+--------------+
| tab_name |
+--------------+
| indicadores |
+--------------+
1 row selected (0.094 seconds)
0: jdbc:hive2://localhost:10000> show create table indicadores;
+--+
| createtab_stmt |
+--+
| CREATE TABLE `indicadores`(|
| `cod` int, |
| `valor` string) |
| ROW FORMAT SERDE |
| 'org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe' |
| STORED AS INPUTFORMAT |
| 'org.apache.hadoop.mapred.TextInputFormat' |
| OUTPUTFORMAT |
| 'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat' |
| LOCATION |
| 'hdfs://namenode:8020/user/hive/warehouse/indicadores' |
| TBLPROPERTIES (|
| 'transient_lastDdlTime'='1647528879') |
+--+
13 rows selected (0.32 seconds)
0: jdbc:hive2://localhost:10000>

Windows PowerShell

```
root@2aba4f3ec24b:/opt/dreyes# hdfs dfs -ls /user/hive/warehouse
Found 1 items
drwxrwxr-x   - root supergroup          0 2022-03-17 14:54 /user/hive/warehouse/indicadores
root@2aba4f3ec24b:/opt/dreyes# hdfs dfs -ls /user/hive/warehouse/indicadores
root@2aba4f3ec24b:/opt/dreyes# beeline -u jdbc:hive2://localhost:10000
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/opt/hive/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/opt/hadoop-2.7.4/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Connecting to jdbc:hive2://localhost:10000
Connected to: Apache Hive (version 2.3.2)
Driver: Hive JDBC (version 2.3.2)
Transaction isolation: TRANSACTION_REPEATABLE_READ
Beeline version 2.3.2 by Apache Hive
0: jdbc:hive2://localhost:10000> load data local inpath "/opt/dreyes/kv2.txt" overwrite into table indicadores;
No rows affected (2.13 seconds)
0: jdbc:hive2://localhost:10000> select * from indicadores limit 5;
+-----------------+-------------------+
| indicadores.cod  | indicadores.valor  |
+-----------------+-------------------+
| 238              | val_238           |
| 86               | val_86            |
| 311              | val_311           |
| 27               | val_27            |
| 165              | val_165           |
+-----------------+-------------------+
5 rows selected (3.891 seconds)
0: jdbc:hive2://localhost:10000> Closing: 0: jdbc:hive2://localhost:10000
root@2aba4f3ec24b:/opt/dreyes# hdfs dfs -ls /user/hive/warehouse/indicadores
Found 1 items
-rwxrwxr-x   3 root supergroup       5812 2022-03-17 15:03 /user/hive/warehouse/indicadores/kv2.txt
root@2aba4f3ec24b:/opt/dreyes# beeline -u jdbc:hive2://localhost:10000
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/opt/hive/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/opt/hadoop-2.7.4/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Connecting to jdbc:hive2://localhost:10000
Connected to: Apache Hive (version 2.3.2)
Driver: Hive JDBC (version 2.3.2)
Transaction isolation: TRANSACTION_REPEATABLE_READ
Beeline version 2.3.2 by Apache Hive
0: jdbc:hive2://localhost:10000> drop table indicadores;
No rows affected (2.569 seconds)
0: jdbc:hive2://localhost:10000> show tables;
+-----------+
| tab_name  |
+-----------+
+-----------+
No rows selected (0.156 seconds)
0: jdbc:hive2://localhost:10000> Closing: 0: jdbc:hive2://localhost:10000
root@2aba4f3ec24b:/opt/dreyes# hdfs dfs -ls /user/hive/warehouse/indicadores
ls: `/user/hive/warehouse/indicadores': No such file or directory
root@2aba4f3ec24b:/opt/dreyes#
```