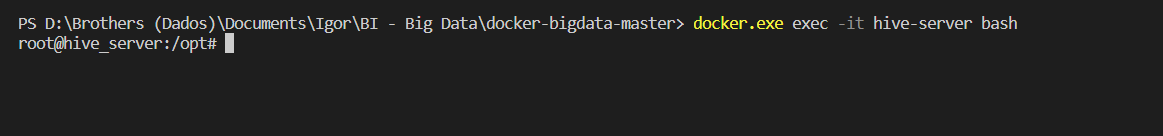
**1. ROLLOUT**

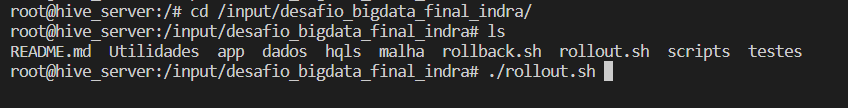
O “rollout.sh” é um arquivo Shell script para criação do ambiente necessário para o funcionamento do pipeline. Os comandos executados irão criar todos os diretórios necessários dentro do HDFS e ao termino irá executar outro arquivo Shell Script “create\_tables.sh”. O arquivo “create\_tables.sh” é responsável por criar o banco de dados “desafio\_final” e todas as tabelas externas e internas necessárias.

O arquivo “create\_tables.sh” executa arquivos “hqls”, onde cada tabela externa e cada tabela interna possui um arquivo no formato .hql. Os arquivos estão armazenados no diretório “hqls”.

Para executar o arquivo “rollout.sh” é necessário está dentro do ambiente Hive-server. Utilizando o comando Docker é possível ativar o ambiente:



Acessando o diretório dentro do hive-serve onde se encontra o arquivo “rollback.sh” é possível executado:



Através do terminal é possível visualizar os processos de criação dos diretórios, criação do banco de dados “desafio\_final” e das tabelas internas e externas para cada um dos arquivos (CLIENTES.csv, DIVISAO.csv, ENDERECO.csv, REGIAO.csv e VENDAS.csv) durante o a execução do “rollout.sh”. Todas as tabelas terão um prefixo “TBL\_” e as tabelas externas terão o sufixo “\_STG”.

TABELA TBL\_CLIENTES\_STG:

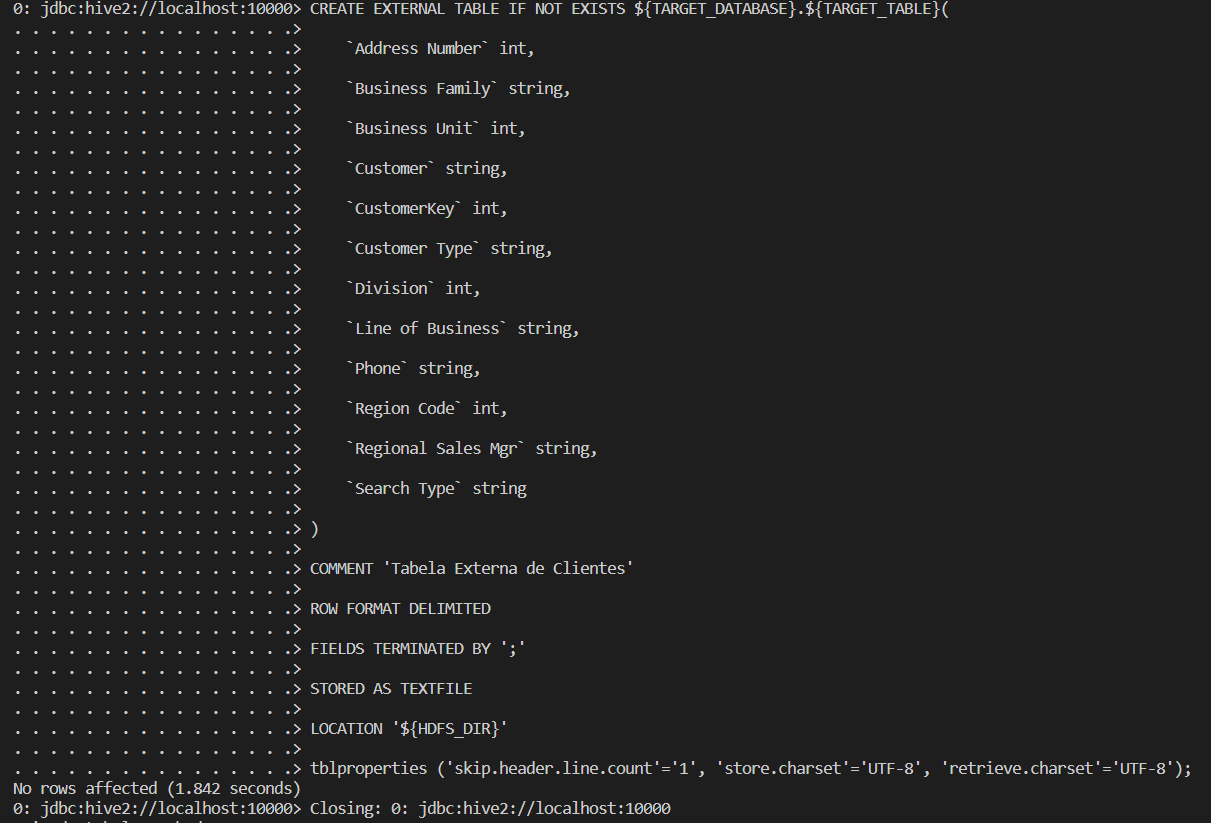


TABELA TBL\_CLIENTES:

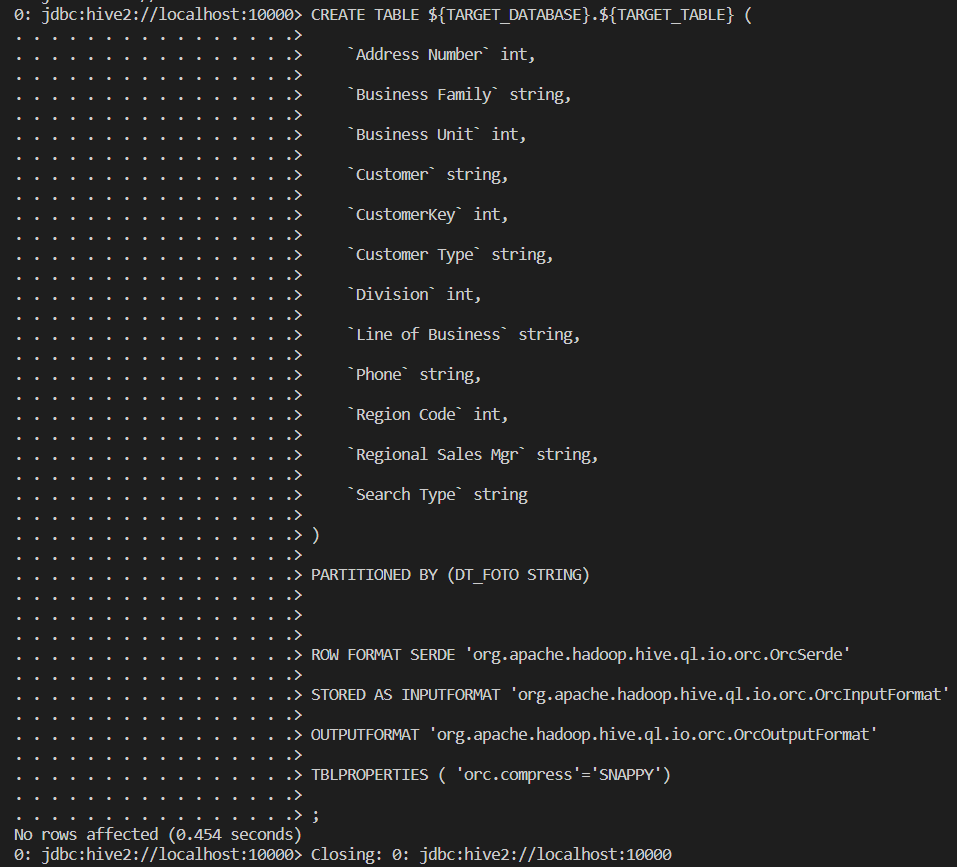


Tabela TBL\_DIVISAO\_STG:

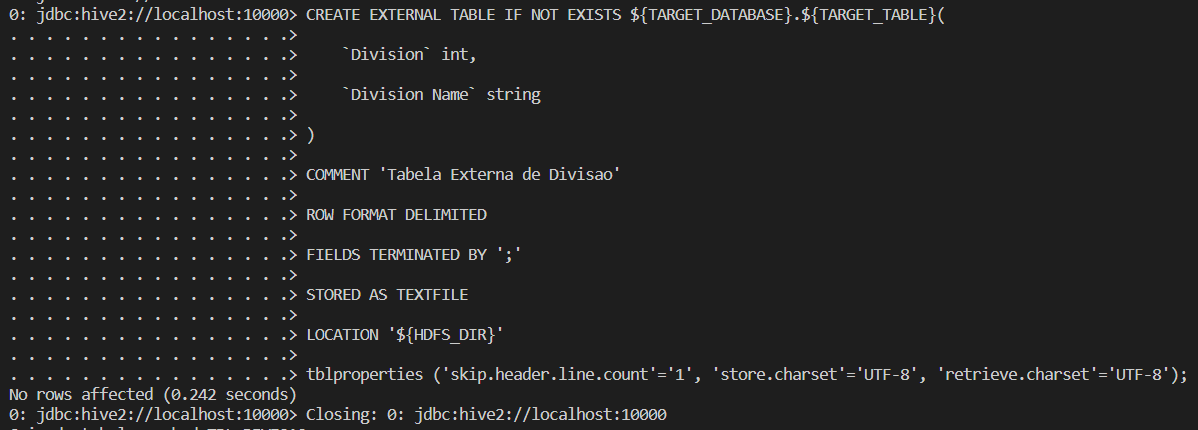


Tabela TBL\_DIVISAO:

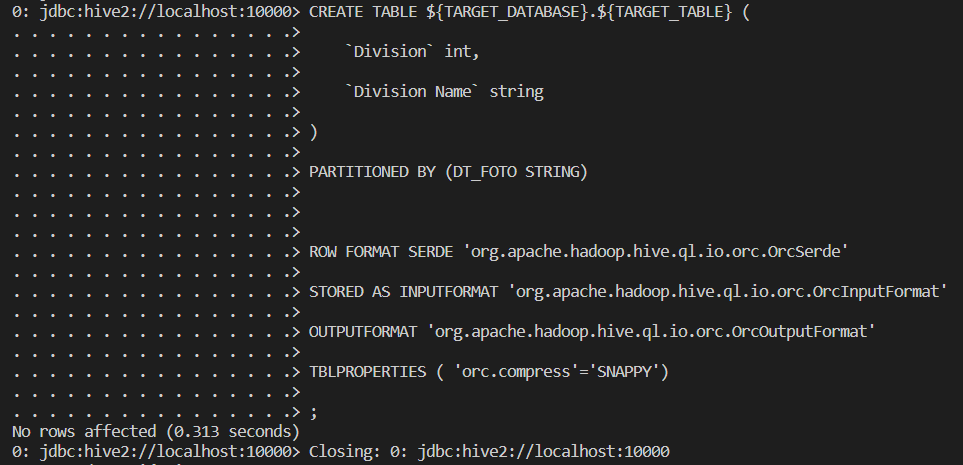


Tabela TBL\_ENDERECO\_STG:

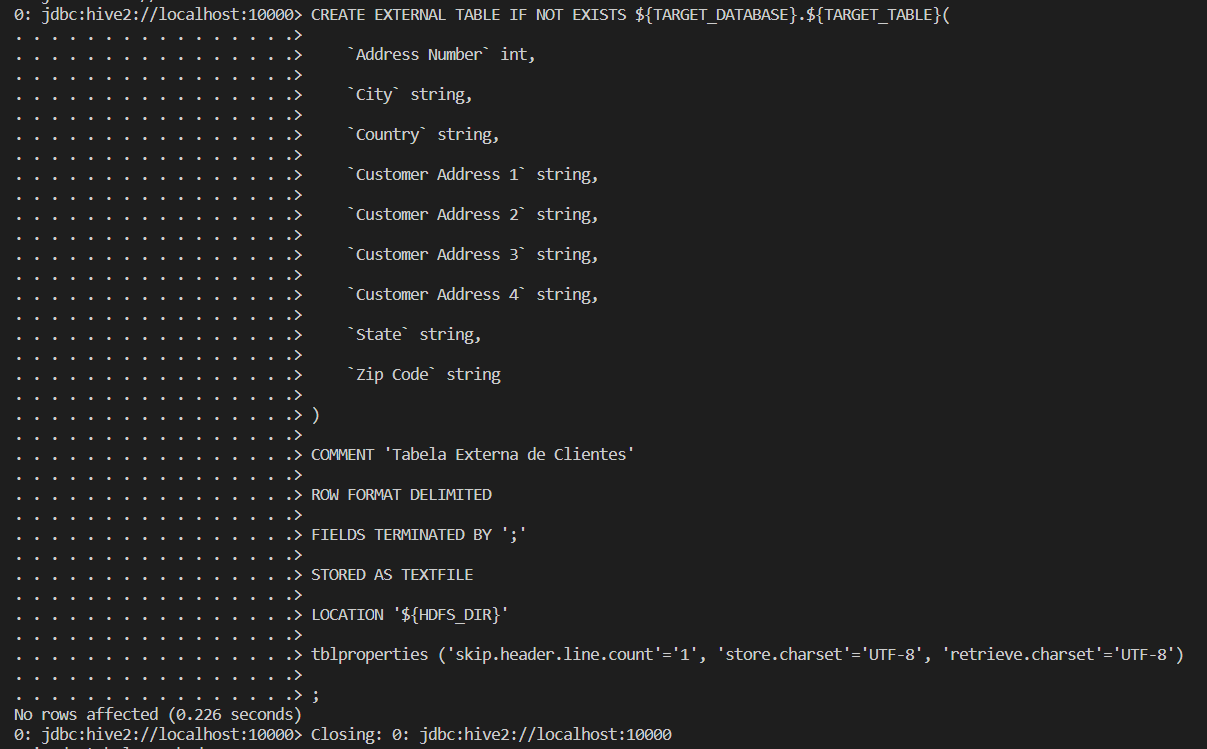


Tabela TBL\_ENDERECO:

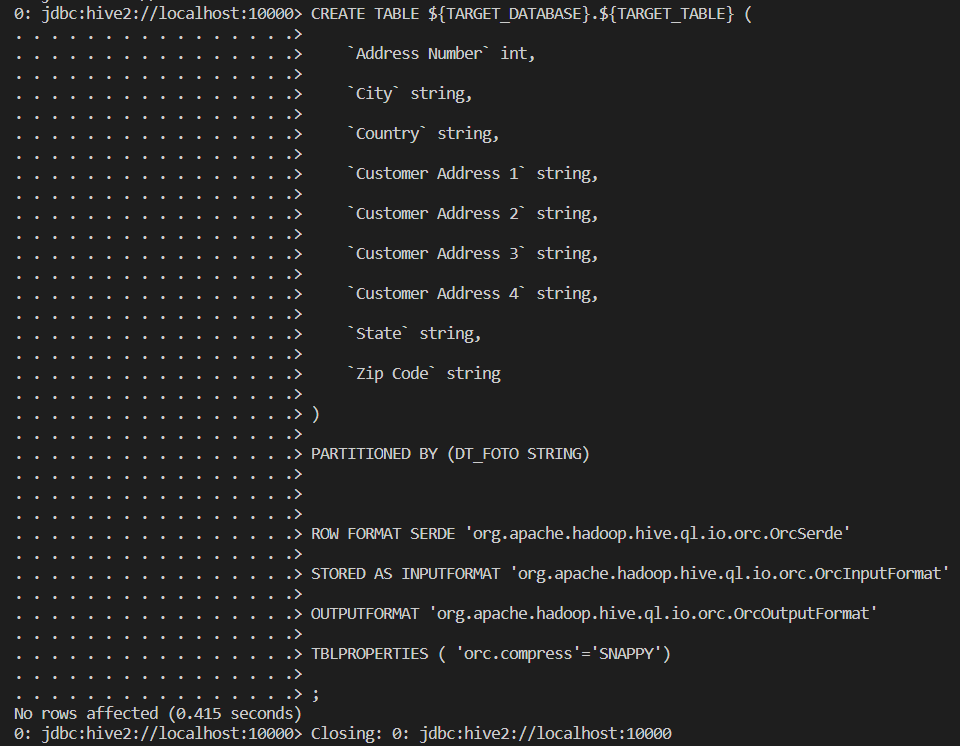


Tabela TBL\_REGIAO\_STG:

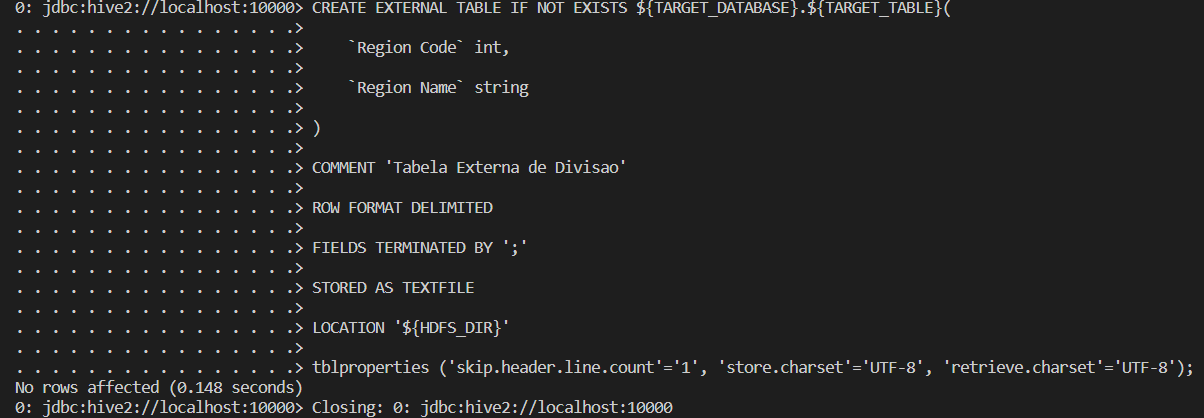


Tabela TBL\_REGIAO:

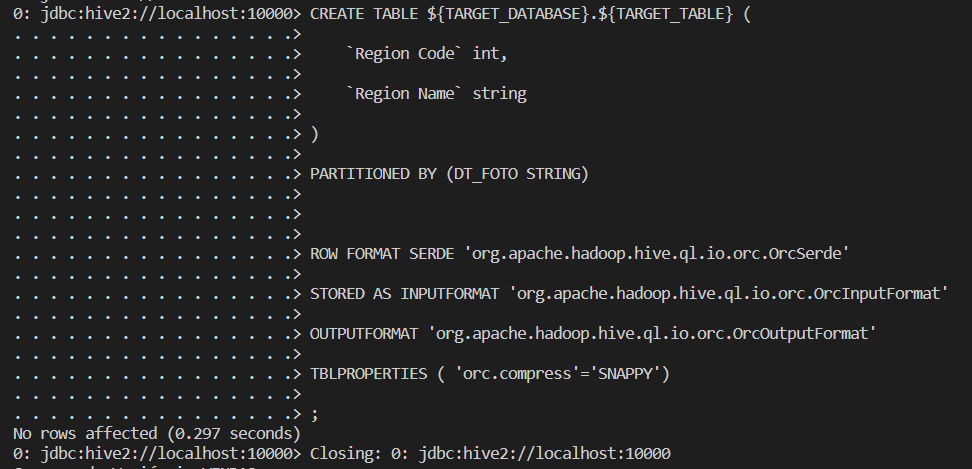
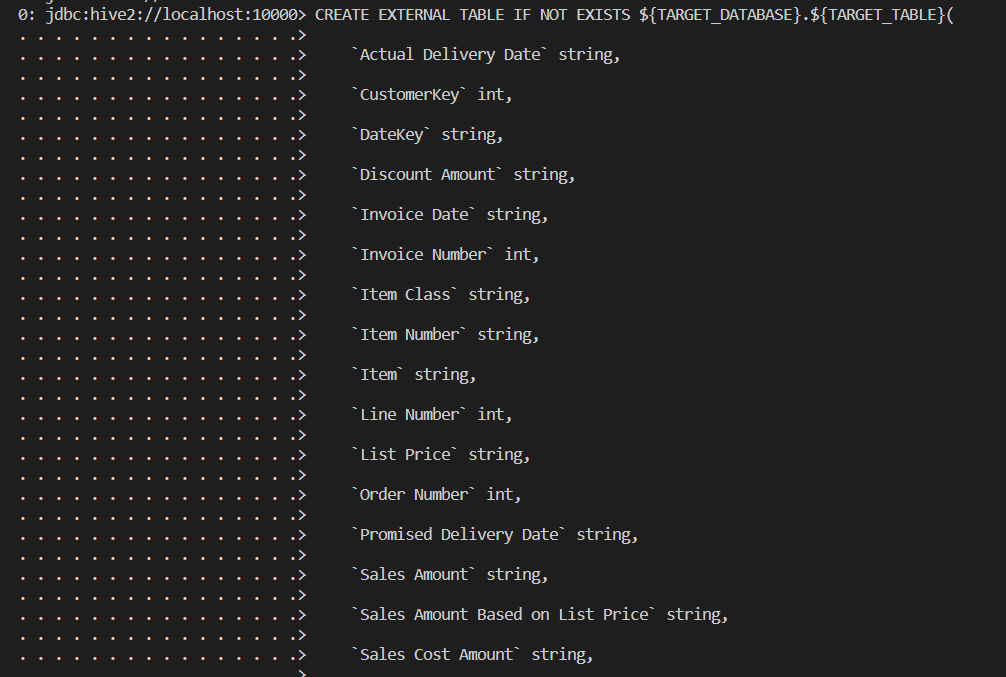


Tabela TBL\_VENDAS\_STG:



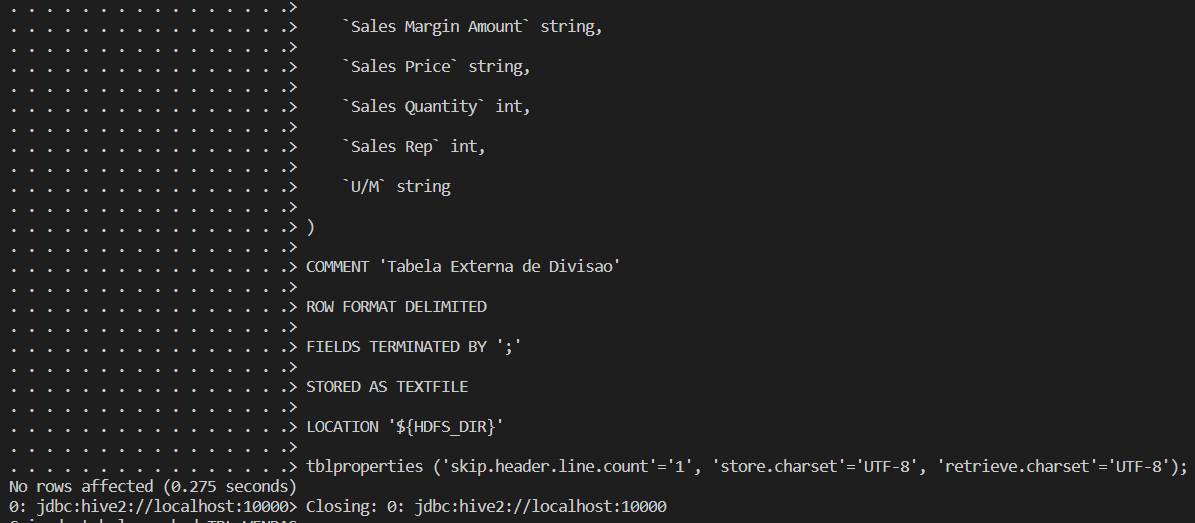
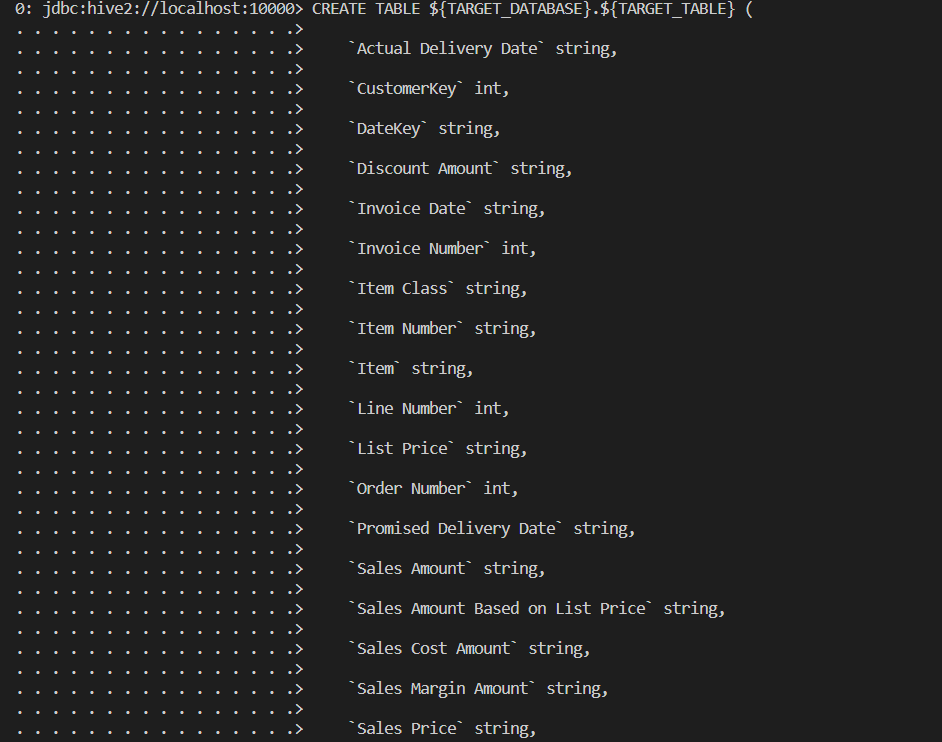
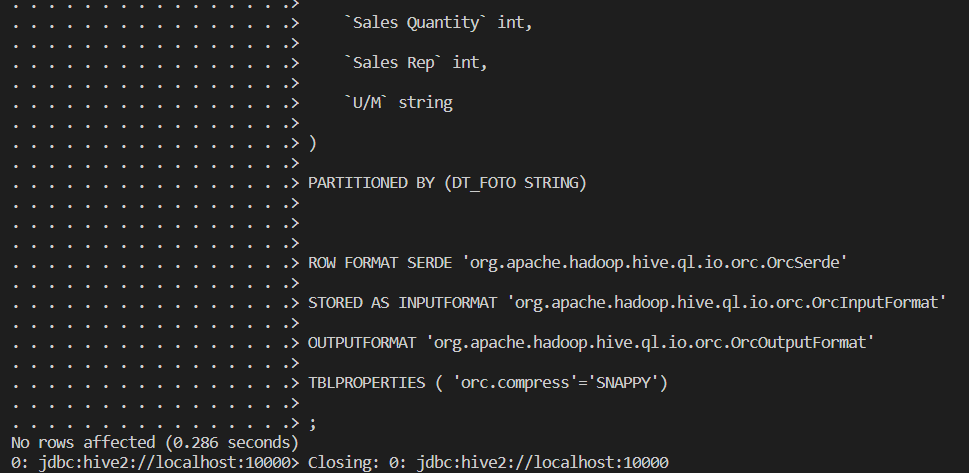
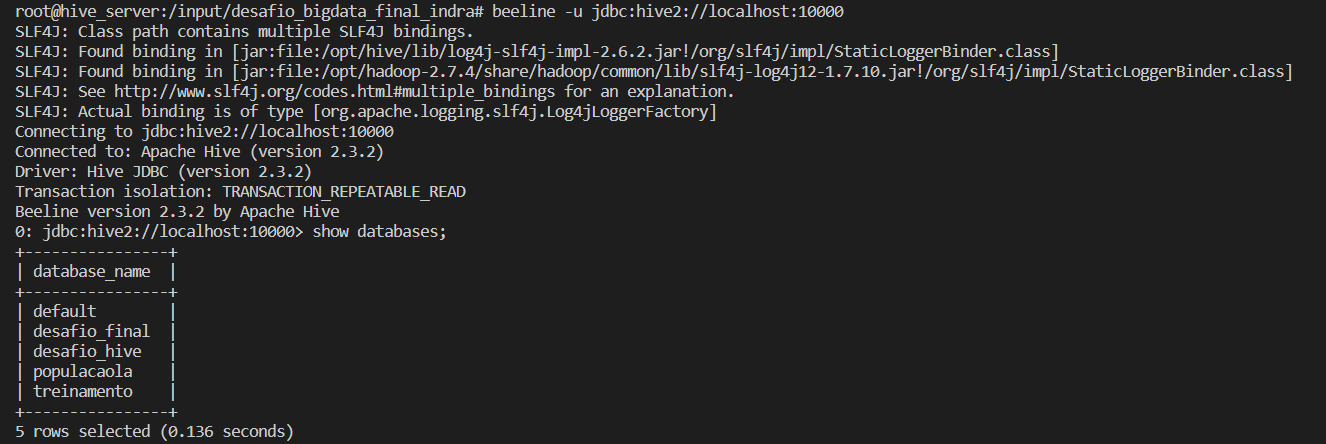


Tabela TBL\_VENDAS\_STG:





Entrando no beeline do hive-serve é possível visualizar a base de dados criadas e dentro dele todas as tabelas internas e externas:



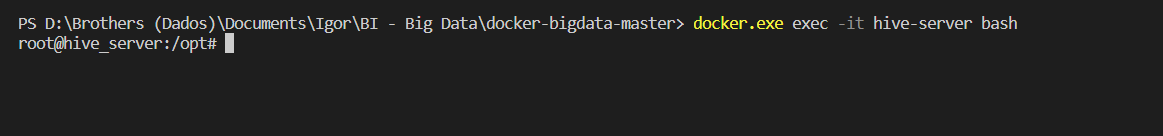
**2. JOBS\_HIVE**

O “jobs\_hive.sh” é um arquivo Shell script utilizado para ingestão das tabelas internas e externas do hive-server. As tabelas externas são atualizadas com dados vindo de arquivos no formato “.csv” que estão no diretório /dados/dados\_entradas, já os dados para as tabelas internas são inseridos a partir das tabelas externas.

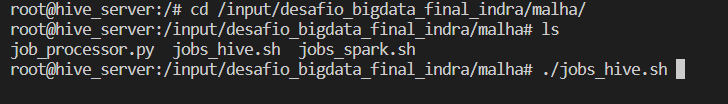
Para o processo de ingestão das tabelas externas o arquivo “jobs\_hive.sh” executa um arquivo denominado “update\_data\_external\_table.sh” passando como parâmetro o nome do arquivo do formato .csv (CLIENTES, DIVISAO, ENDERECO, REGIAO e VENDAS).

Para a ingestão das tabelas internas o arquivo “jobs\_hive.sh” executa um arquivo denominado “insert\_data\_worked\_table.sh” passando como parâmetro o nome da tabela (TBL\_CLIENTES, TBL\_DIVISAO, TBL\_ENDERECO, TBL\_REGIAO e TBL\_VENDAS) e o nome do arquivo “hql” (insert-table-clientes-wrk, insert-table-divisao-wrk, insert-table-endereco-wrk, insert-table-regiao-wrk, insert-table-vendas-wrk).

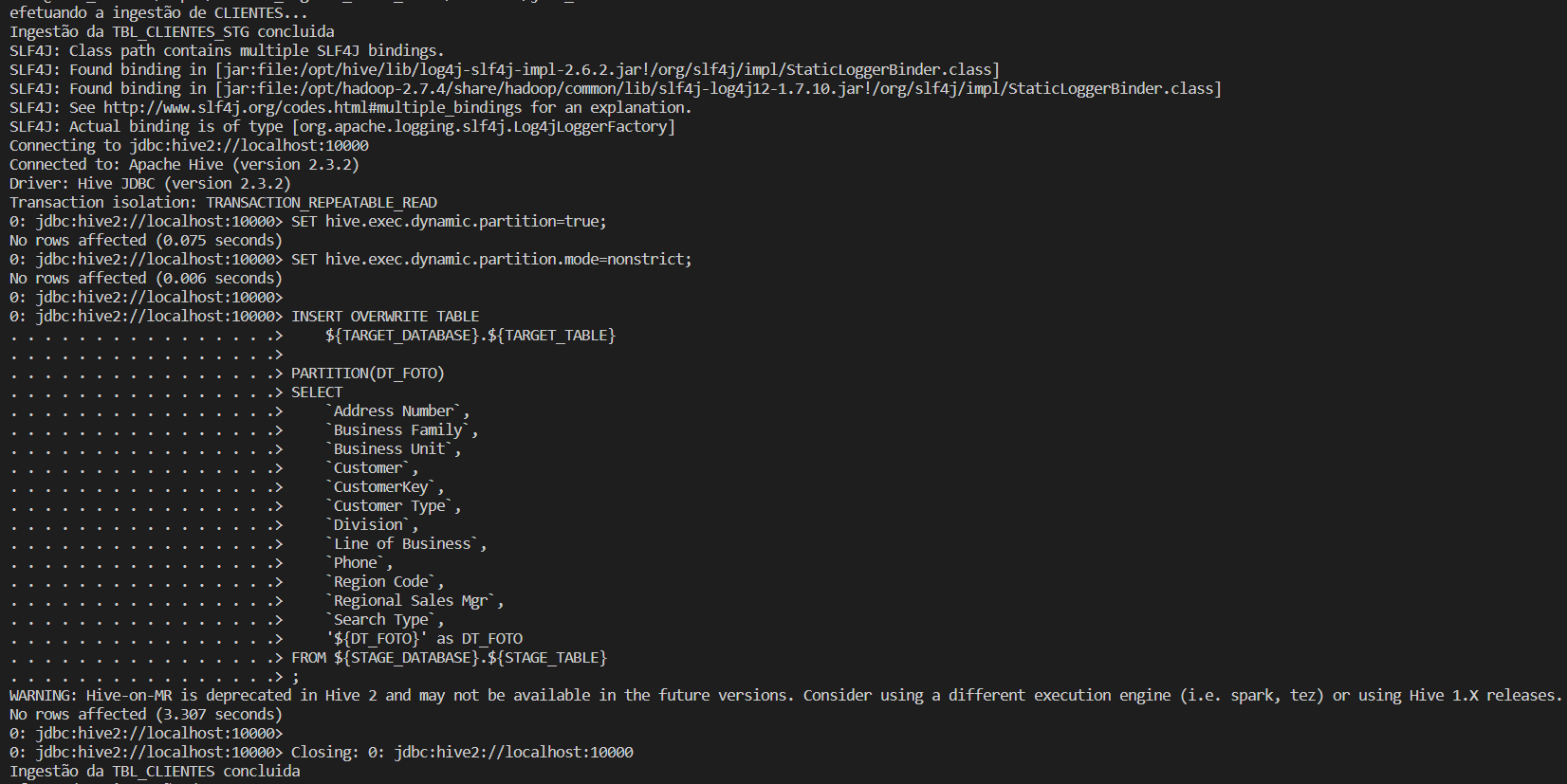
Para executar o arquivo “rollback.sh” é necessário está dentro do ambiente Hive-server. Utilizando o comando Docker é possível ativar o ambiente:



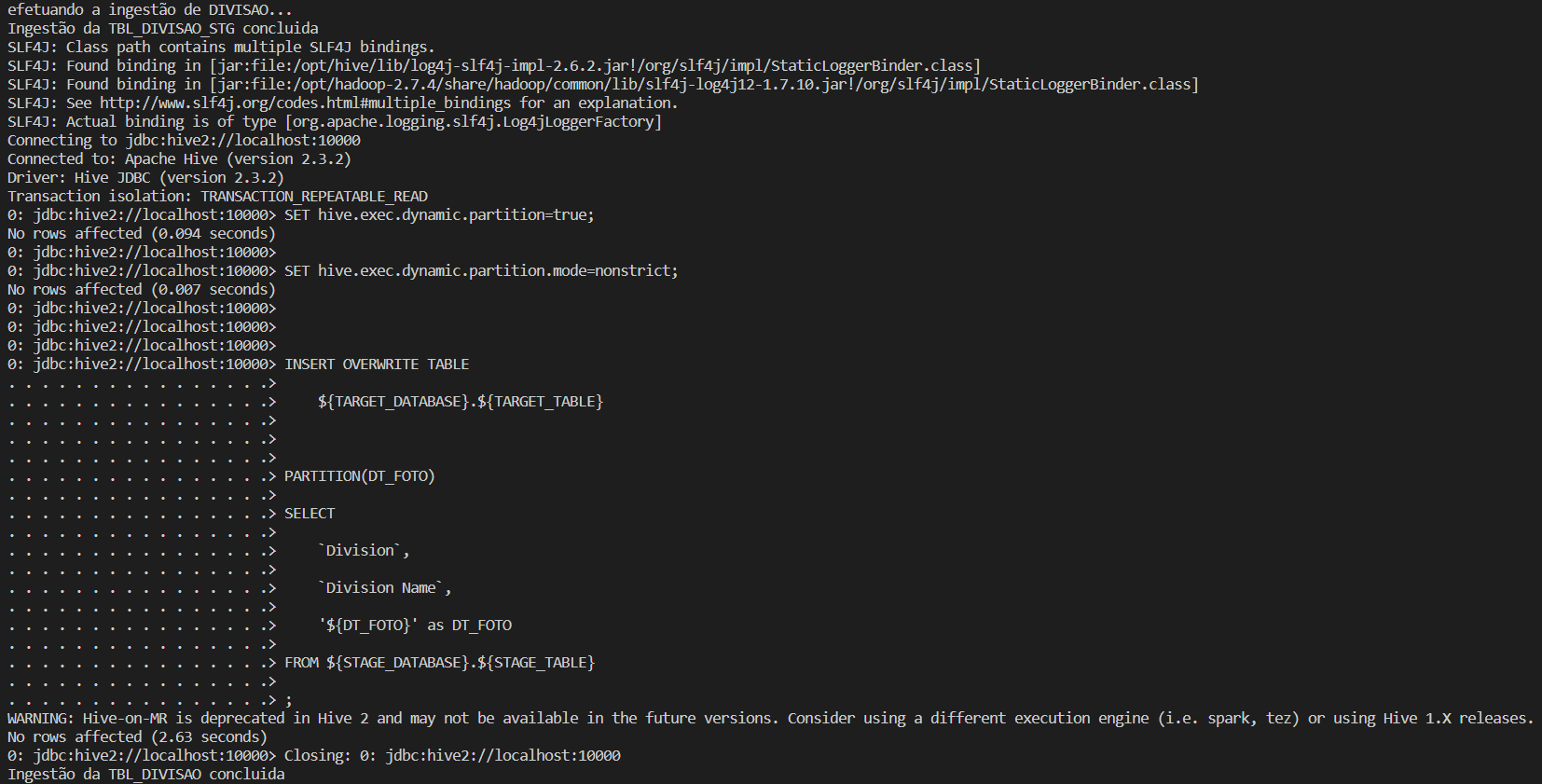
Acessando o diretório dentro do hive-serve onde se encontra o arquivo “jobs\_hive.sh” é possível executado:



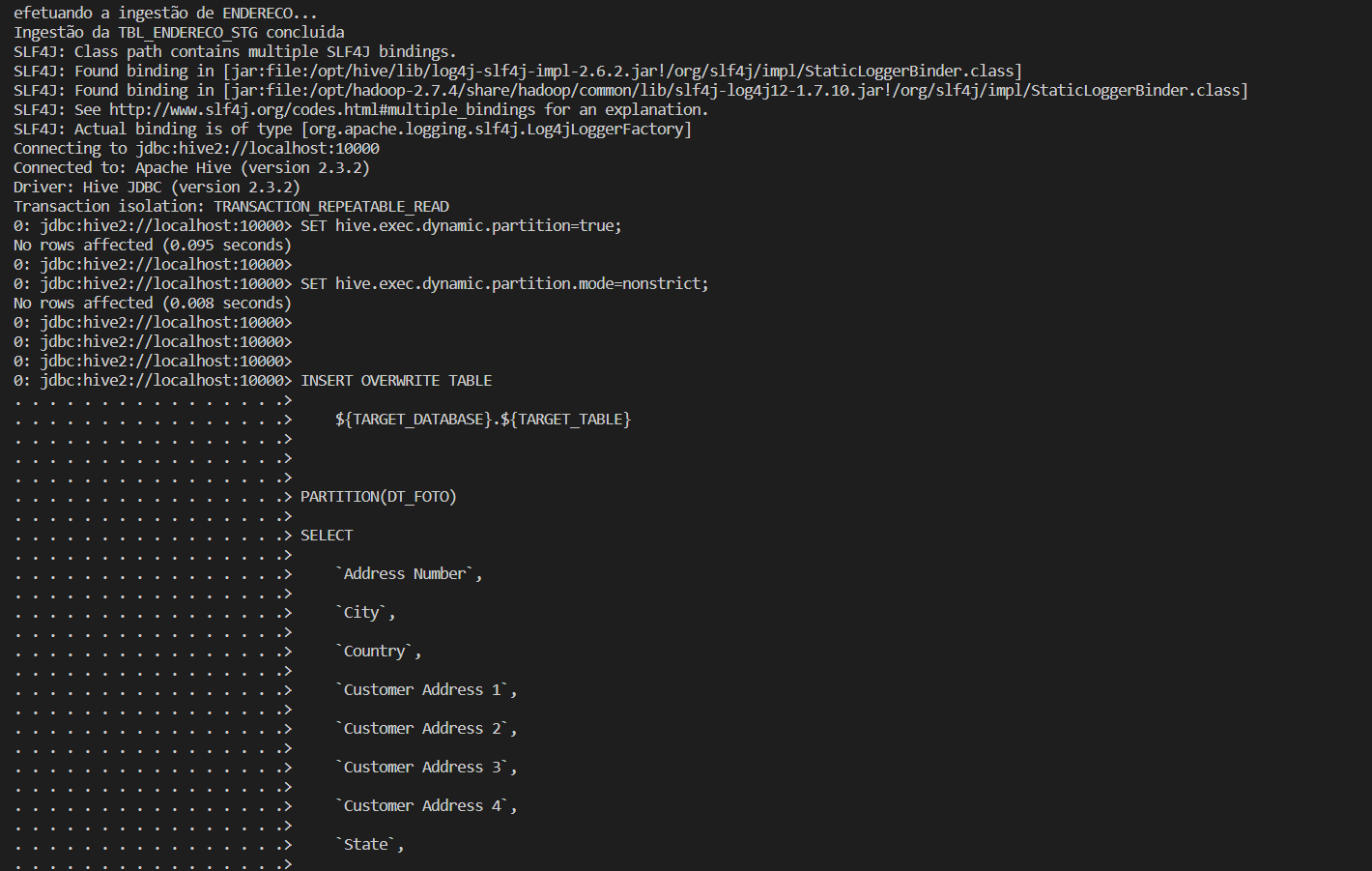
Ingestão das tabelas externas e internas de CLIENTES:

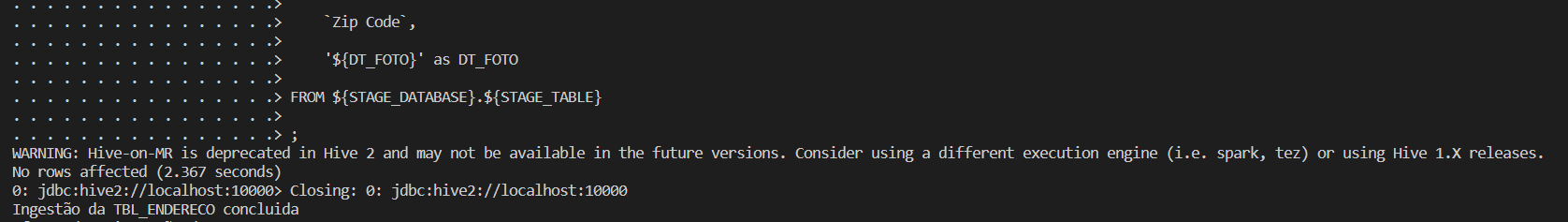


Ingestão das tabelas externas e internas de DIVISAO:

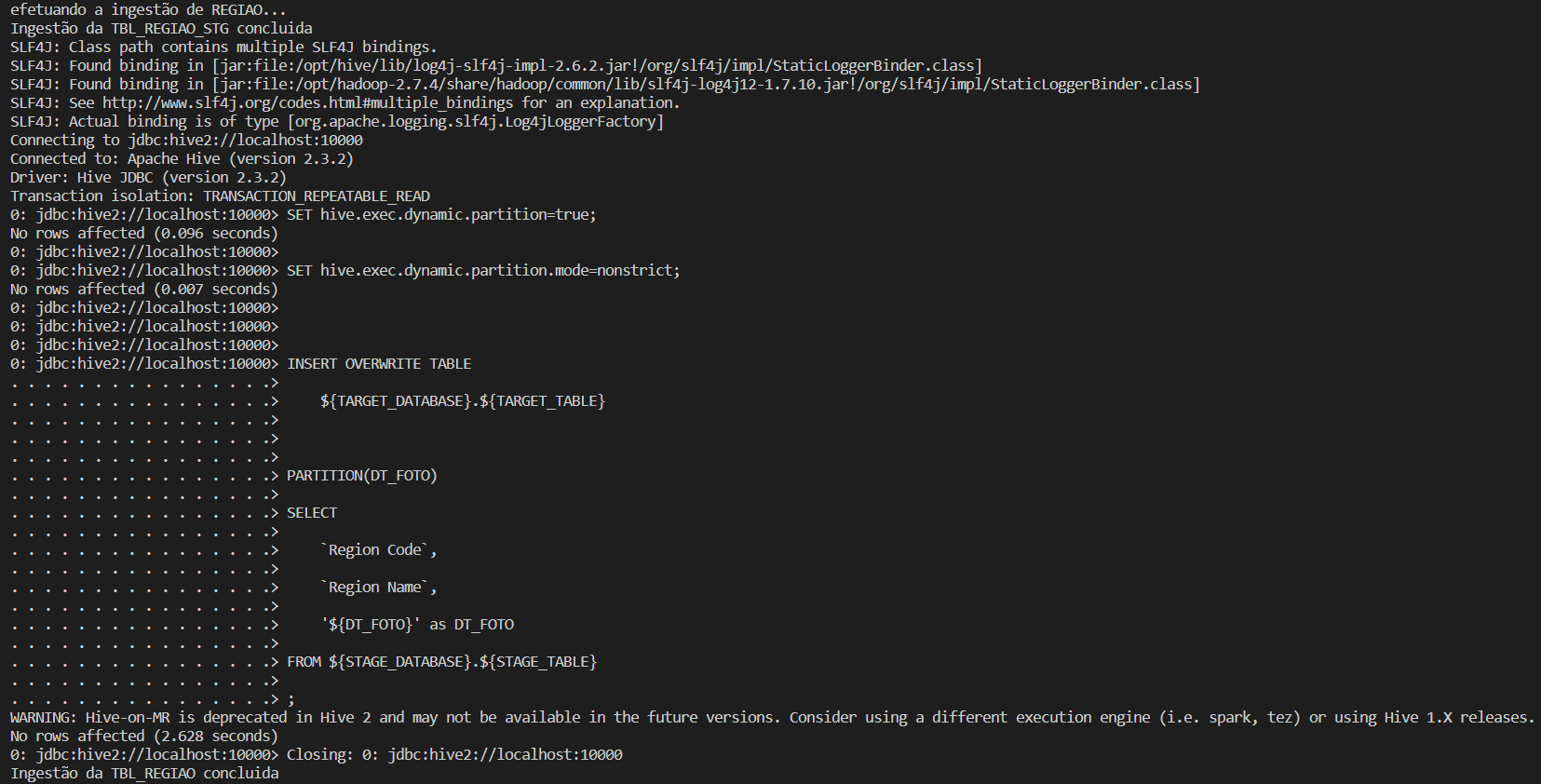


Ingestão das tabelas externas e internas de ENDERECO:

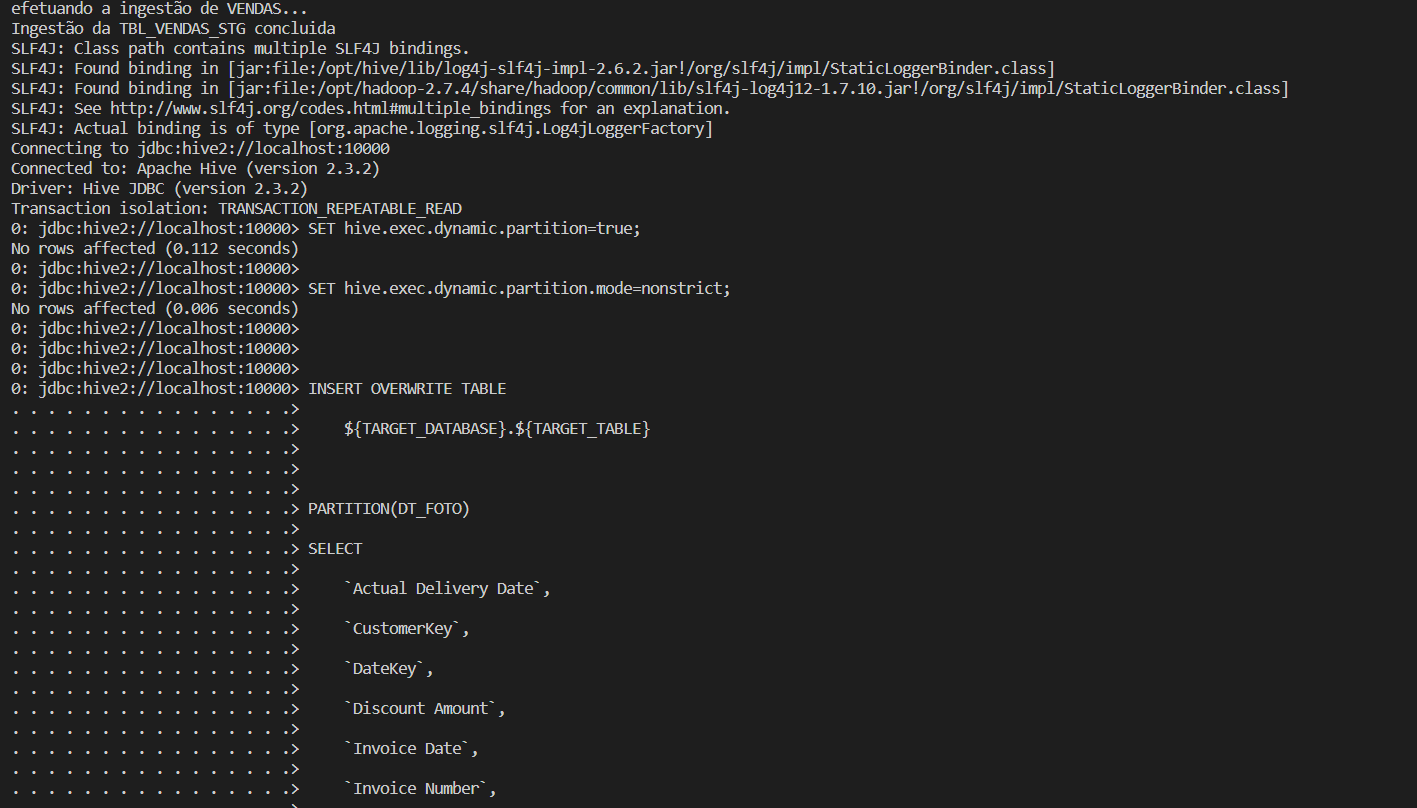


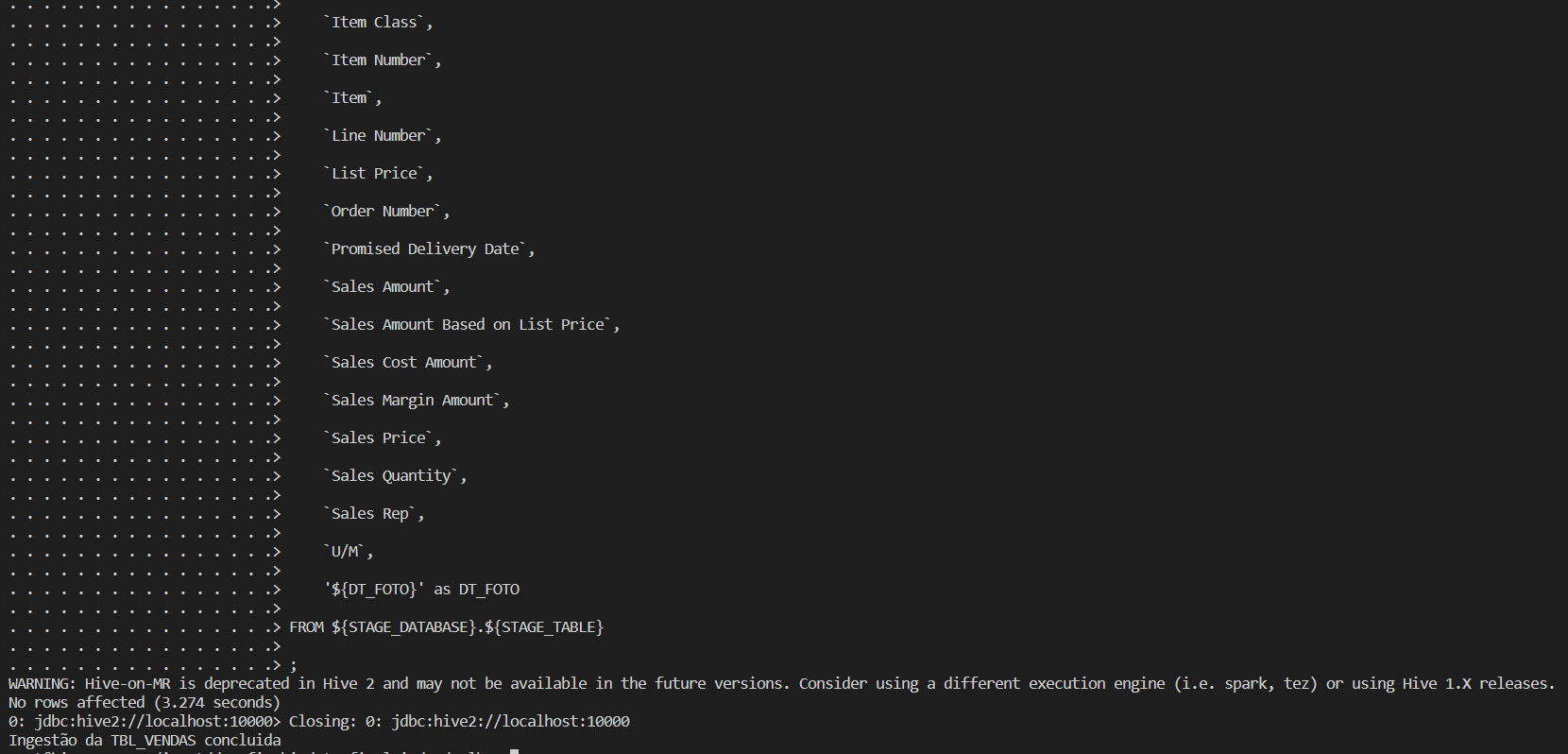


Ingestão das tabelas externas e internas de REGIAO:

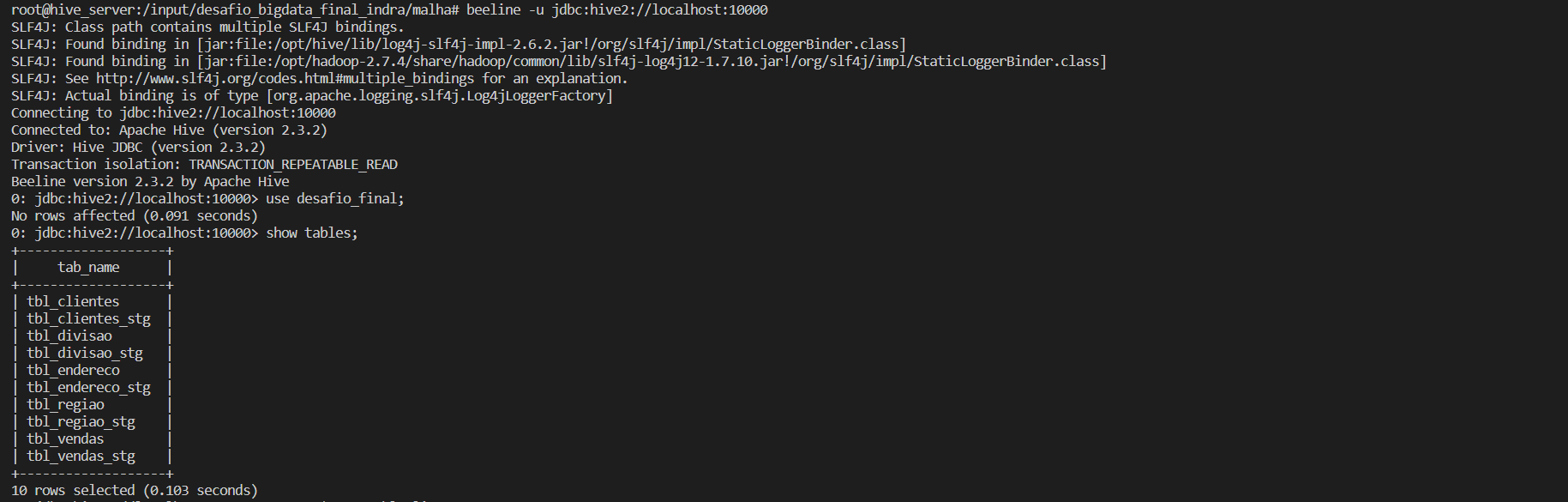


Ingestão das tabelas externas e internas de VENDAS:

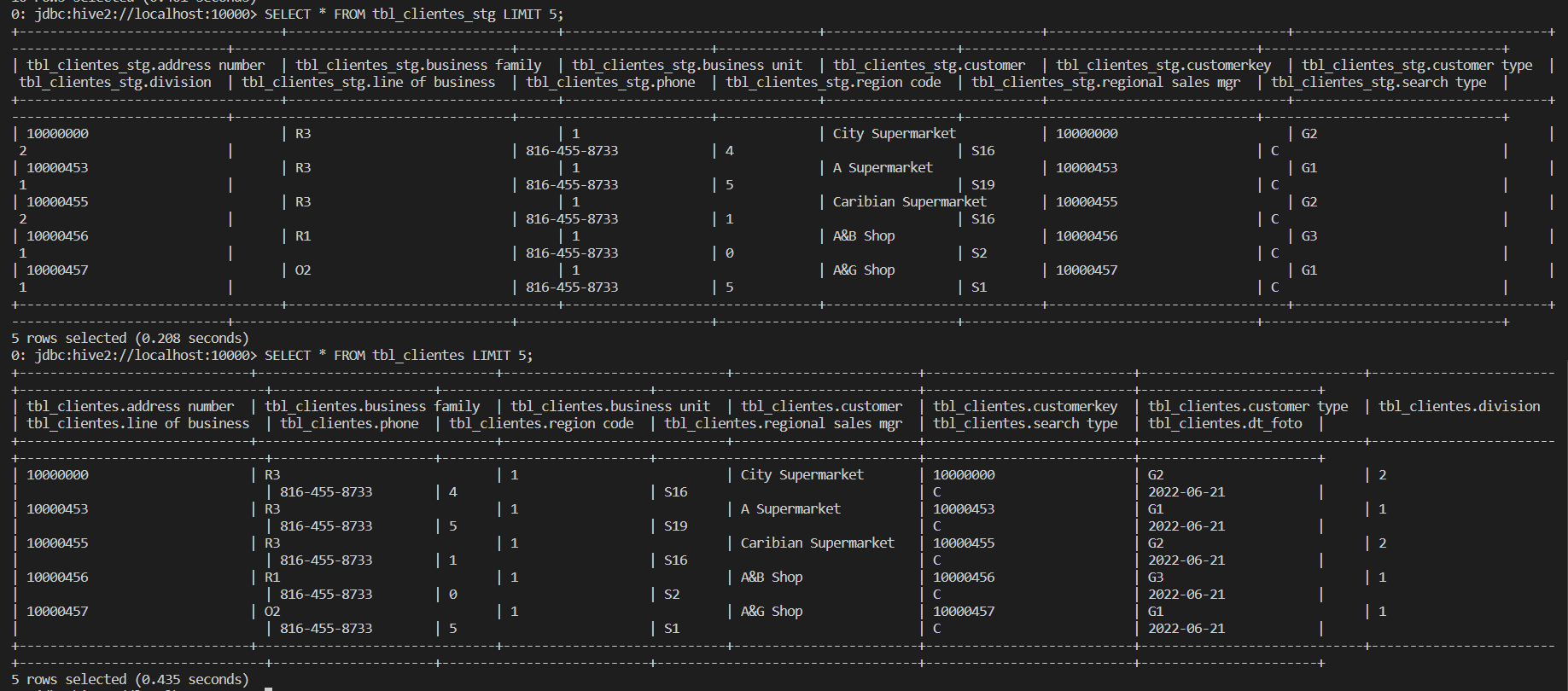




Entrando no beeline do hive-serve é possível conferir se os dados foram inseridos corretamente dentro das tabelas internas e externas:



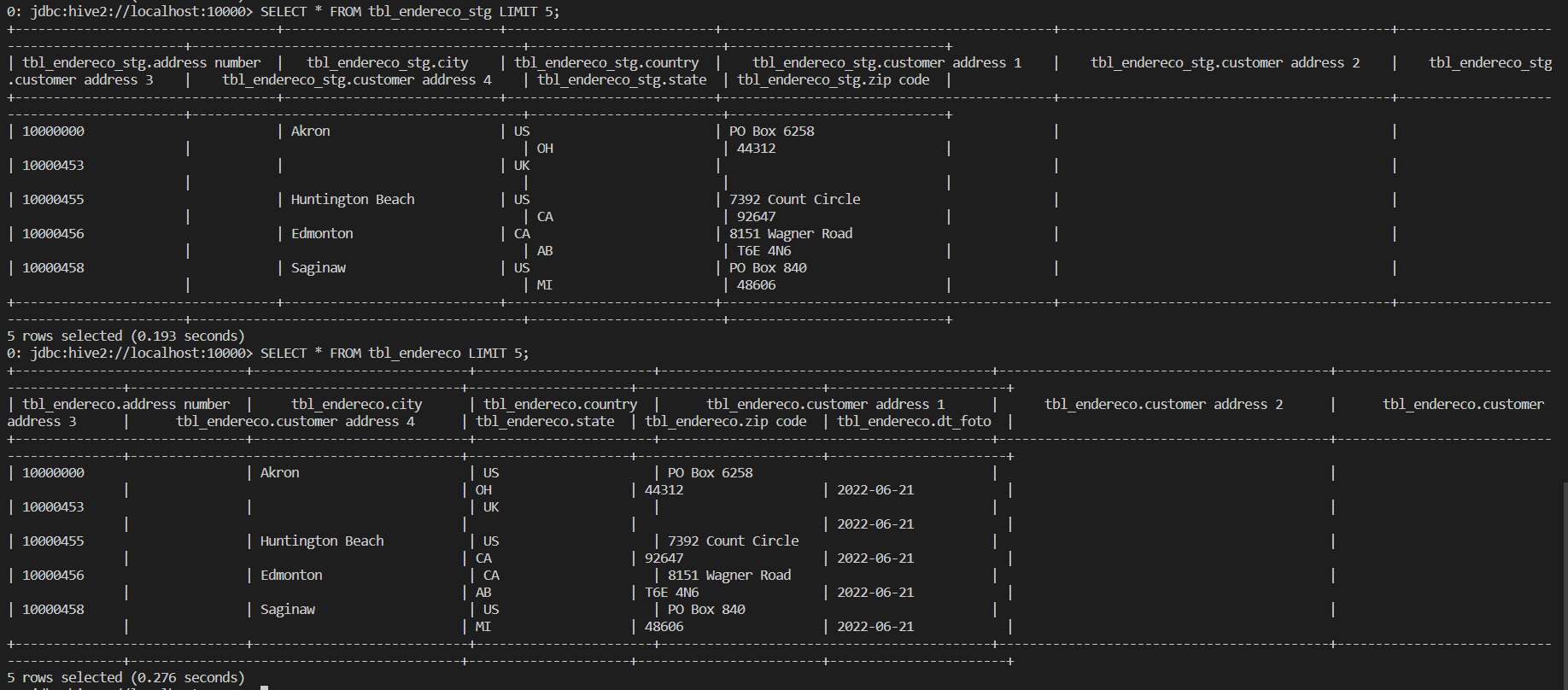
Tabelas de CLIENTES:



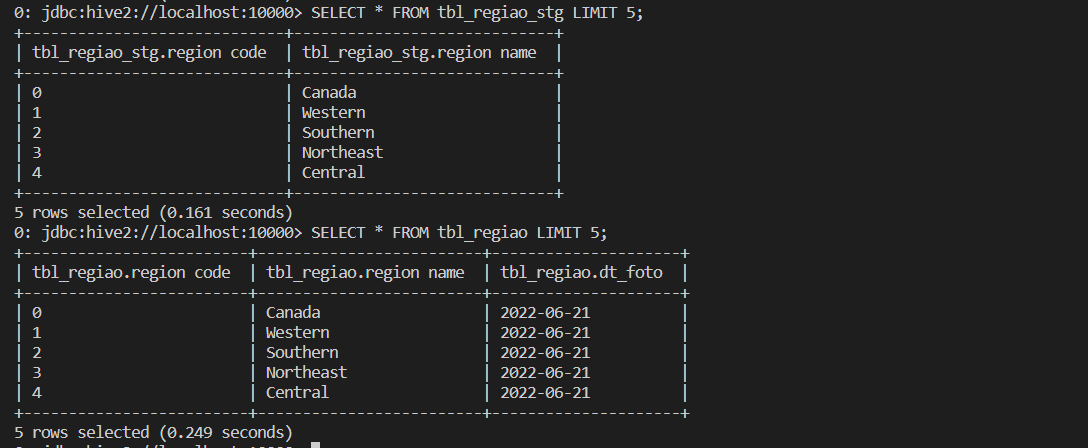
Tabelas de DIVISAO:



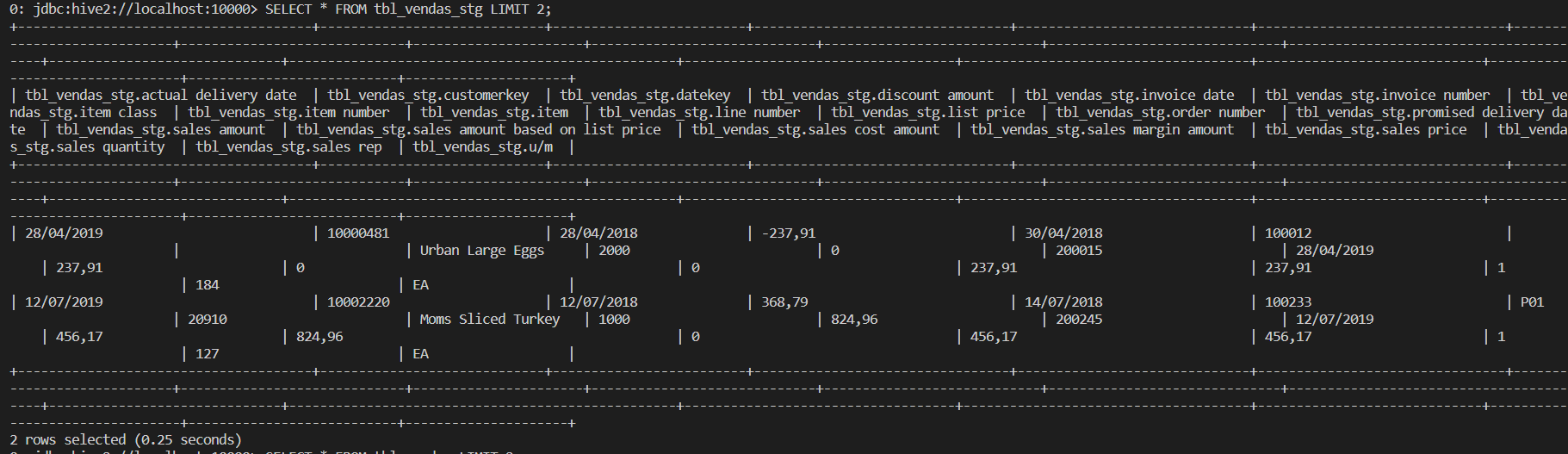
Tabelas de ENDERECO:

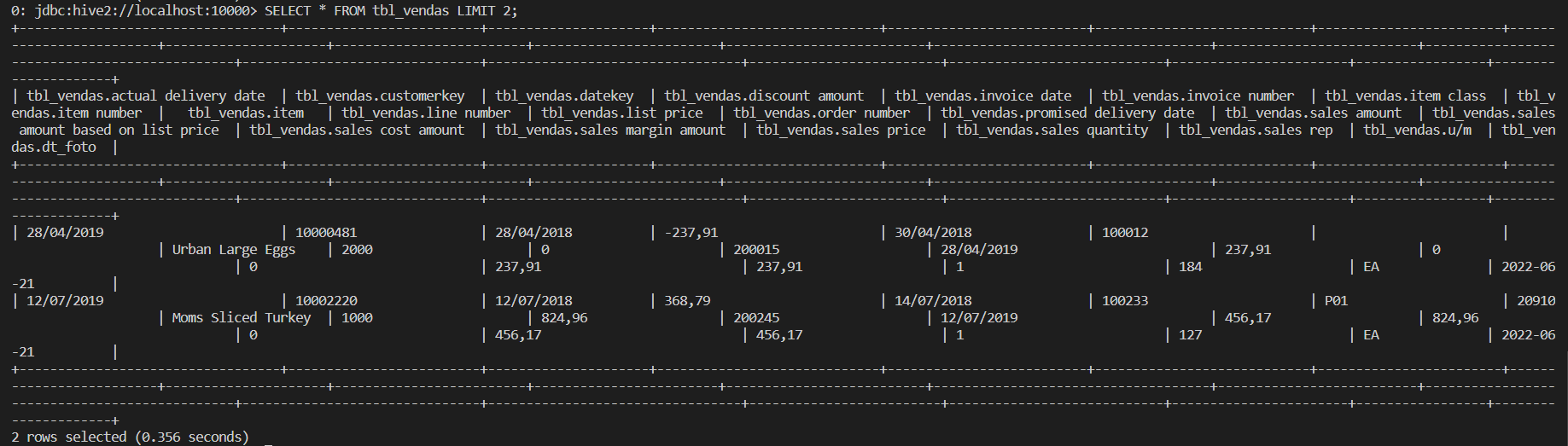


Tabelas de REGIAO:



Tabelas de VENDAS:

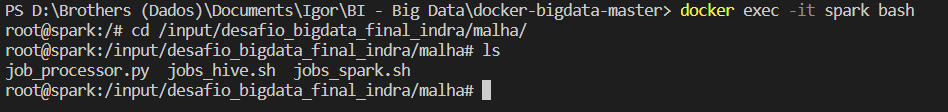




**3. JOBS\_SPARK**

O “jobs\_spark.sh” é um arquivo Shell script utilizado para tratamento dos dados que estão armazenados dentro das tabelas internas criadas no hive-serve. O arquivo irá executar o “job\_processor.py” que um arquivo Python que contem todo o código responsável por fazer limpeza dos dados e transformar o modelo transacional em dimensional STAR. Após o tratamento e transformação o modelo dimensional é salvo no HDFS. A partir dos arquivos salvos no HDFS, o jobs\_spark renomeia e transfere os arquivos para UNIX, no diretório /dados/dados\_saida.

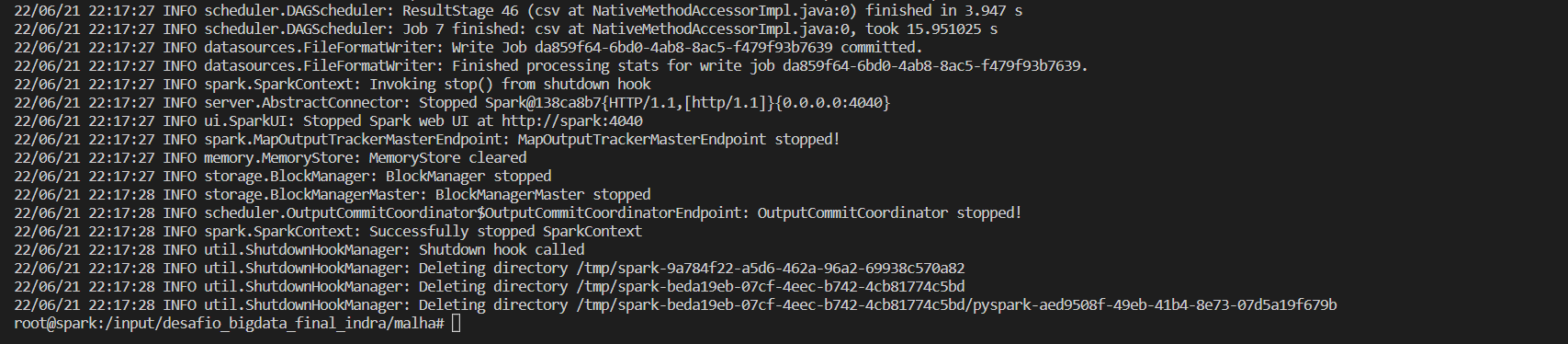
Para executar o arquivo “jobs\_spark.sh” é necessário está dentro do ambiente Spark. Utilizando o comando Docker é possível ativar o ambiente:



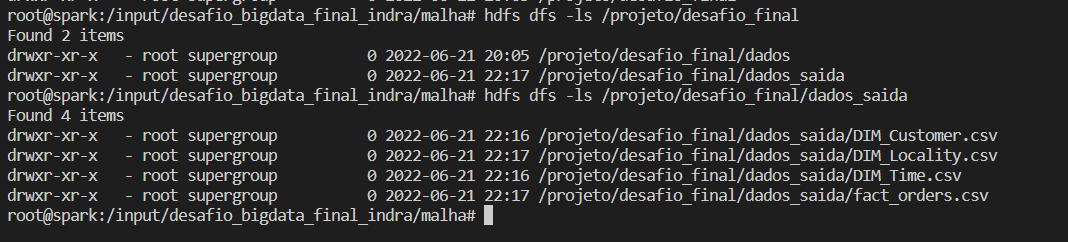
Executado “jobs\_hive.sh”:



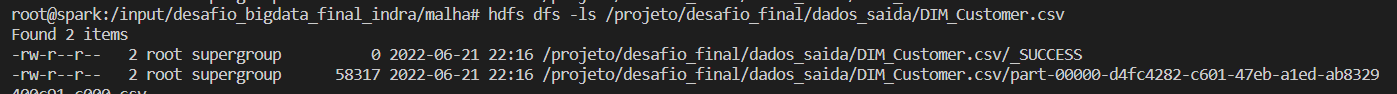
Finalizada a execução sem nem um tipo de erro:



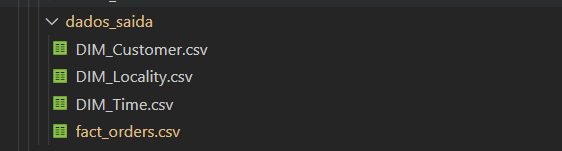
No HDFS é criado no diretório /projeto/desafio\_final/dados\_saida quatro diretórios: três dimensões e um fato.



Dentro de cada diretório é encontrado a seguinte estrutura:



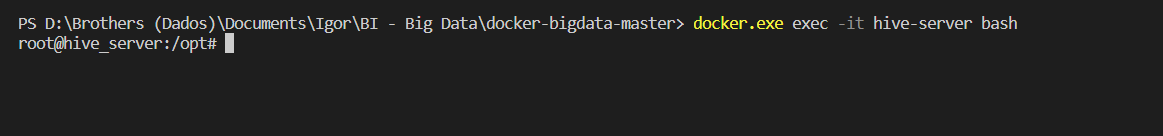
Com objetivo de facilitar a utilização dos arquivos foi transferido apenas os arquivos que realmente contém os dados no formado .csv para o UNIX devidamente renomeados. Logo, no diretório /dados/dados\_saida é possível encontrar os arquivos:



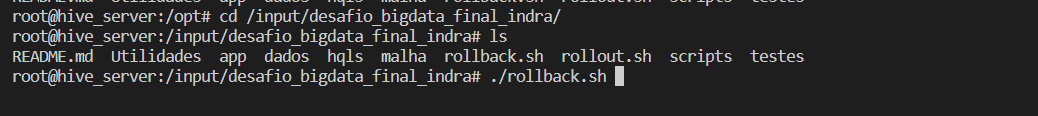
**4. ROLLBACK**

O “rollback.sh” é um arquivo Shell script para desfazer todas as alterações realizadas durante a criação do ambiente do pipeline. Os comandos executados irão deletar todos os diretórios criados e as tabelas internas e externas dentro do Hive.

Para executar o arquivo “rollback.sh” é necessário está dentro do ambiente Hive-server. Utilizando o comando Docker é possível ativar o ambiente:



Acessando o diretório dentro do hive-serve onde se encontra o arquivo “rollback.sh” é possível executado:

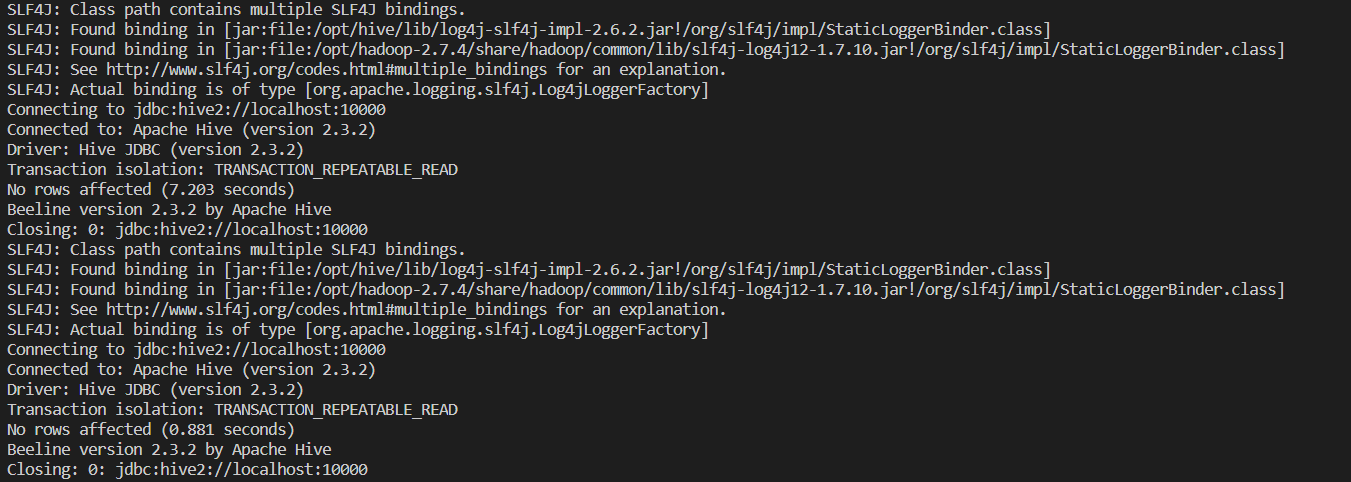


Através do terminal é possível visualizar os processos de remoção das tabelas internas e externas e o banco de dados que foram criadas durante o a execução do “rollout.sh”. Além dos diretórios do HDFS que foram removidos de forma recursiva.

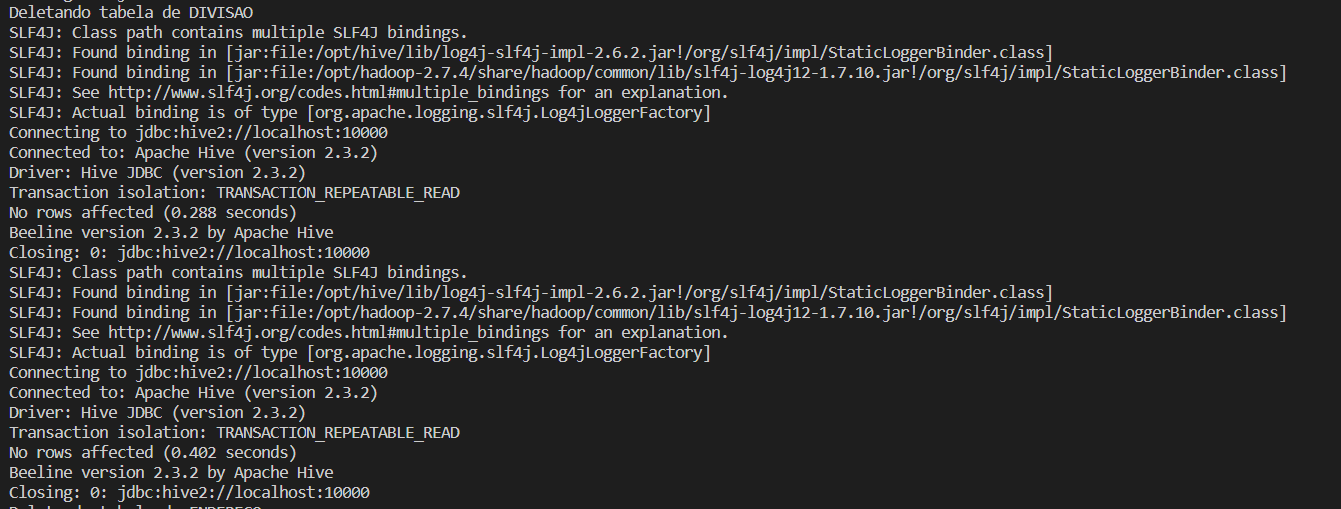
Diretórios:



Tabelas de CLIENTES:



Tabelas de DIVISÃO:



Tabelas de ENDERECO:

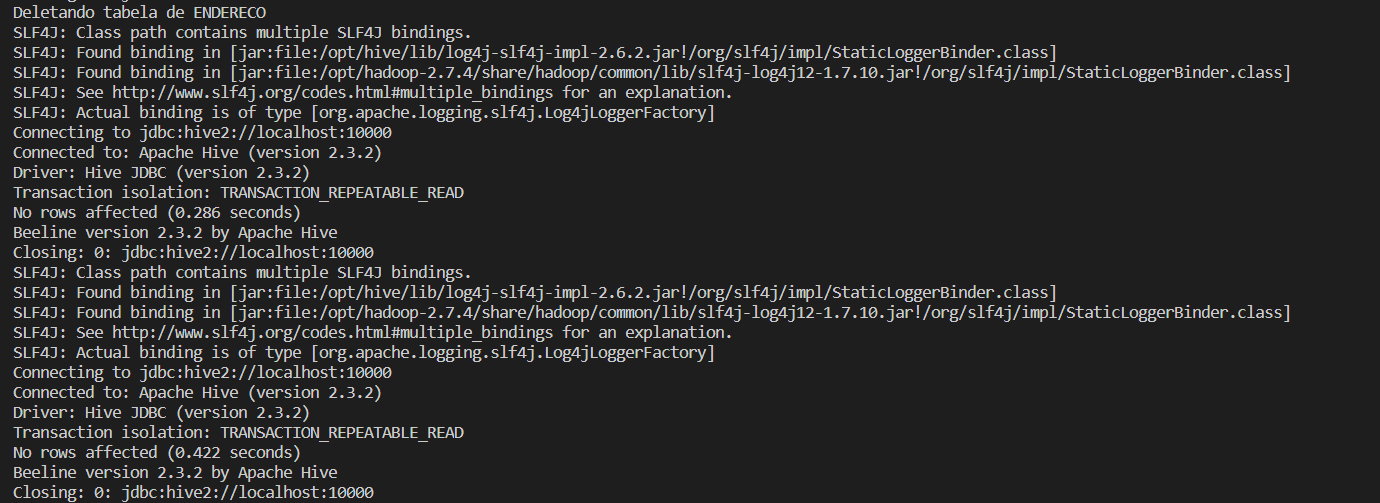


Tabela de REGIAO:

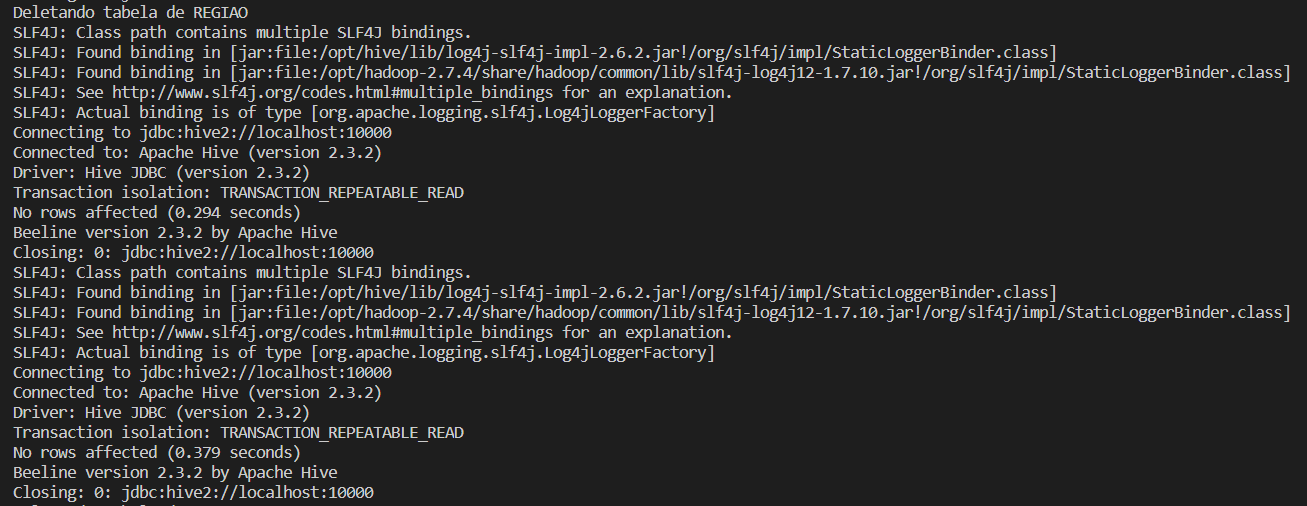
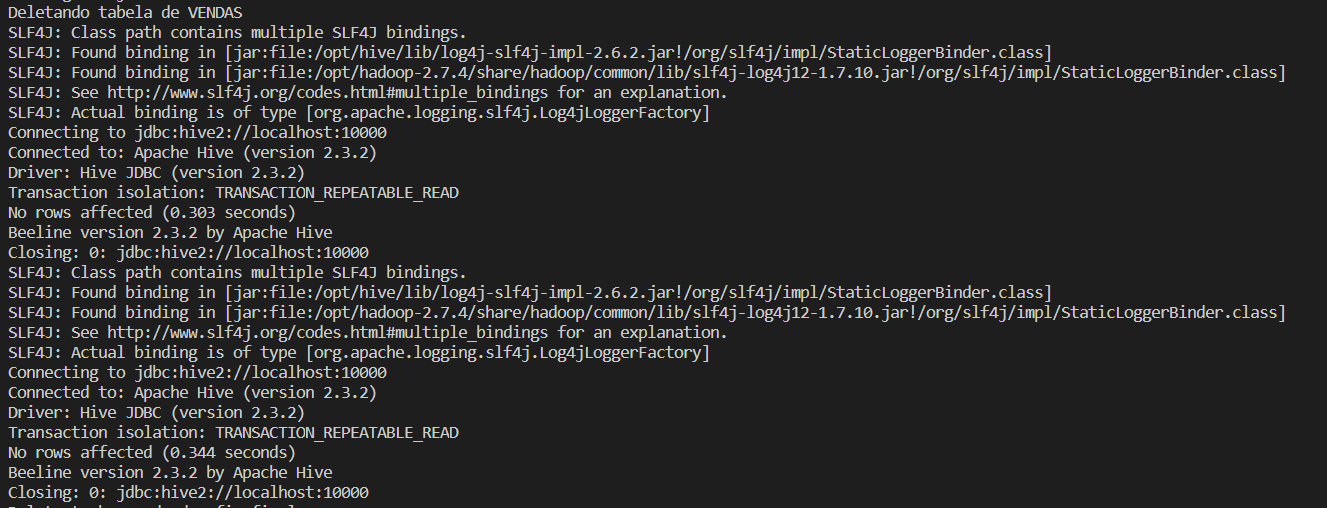
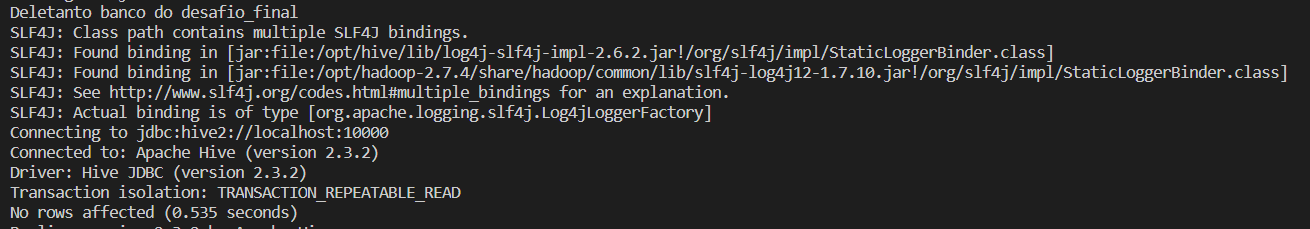


Tabela de VENDAS:



Banco de dados desafio\_final:



Entrando no beeline do hive-serve é possível visualizar que o banco de dados desafio\_final não existe mais no ambiente.