Criando Tabelas com arquivos em formatos diferentes

1. Faça login no Hue e vá para o editor de consultas Impala.
2. Faça o seguinte para criar uma tabela, preencha-a com uma linha de dados e observe o arquivo resultante no HDFS:
   a) Execute a seguinte instrução CREATE TABLE:

```
CREATE TABLE jobs_txt
      (id INT, title STRING, salary INT, posted TIMESTAMP)
    STORED AS TEXTFILE;
```

   b) Carregue uma linha de dados executando a instrução a seguir.

```
INSERT INTO jobs_txt
    VALUES (1,'Data Analyst',135000,'2016-12-21 15:52:03');
```

   c) Use o Browser File [Navegador de arquivos] ou o data source panel [painel de fonte de dados] (escolhendo o ícone de arquivos em vez do ícone do banco de dados) e localize o diretório **/user/hive/warehouse/jobs_txt**. Se você não vir o subdiretório **jobs_txt**, atualize a exibição clicando no botão atualizar (duas setas curvas). Encontre um arquivo com um nome que seja apenas uma sequência de letras e números e clique nesse arquivo.

   d) Você pode ver o conteúdo do arquivo no painel principal. Observe que você pode ver claramente cada um dos valores adicionados à tabela.

3. Agora crie outra tabela usando um formato diferente e veja que o arquivo resultante parece diferente:
   a) Execute a seguinte instrução **CREATE TABLE**, que configura a tabela para armazenar dados no formato **PARQUET**:

```
CREATE TABLE jobs_parquet
       (id INT, title STRING, salary INT, posted TIMESTAMP)
   STORED AS PARQUET;
```

   b) Carregue uma linha de dados executando a instrução a seguir.

```
INSERT INTO jobs_parquet
   VALUES (1,'Data Analyst',135000,'2016-12-21 15:52:03');
```

   c) Use o Browser File [Navegador de Arquivos] ou o Data source panel [painel de origem de dados] (escolhendo o ícone de arquivos em vez do ícone do banco de dados) e localize o diretório **/user/hive/warehouse/jobs_parquet**. Se você não vir o subdiretório **jobs_parquet**, atualize a exibição clicando no botão atualizar (duas setas curvas). Encontre um arquivo com um nome que seja apenas uma sequência de letras e números e clique nesse arquivo. Você receberá uma mensagem de erro informando que o Hue não pode ler o arquivo.

   d) Abra uma janela do Terminal. (Você pode fazer isso clicando no ícone na barra de menus que se parece com um computador.) Digite e execute o seguinte comando, que mostrará o conteúdo do arquivo Parquet. (Não inclua o $; esse é o prompt para indicar que este é um comando shell de linha de comando, não uma consulta.) Observe que a saída inclui muitos caracteres não ASCII, portanto, você não pode realmente ler a maior parte.

```
$ hdfs dfs -cat /user/hive/warehouse/jobs_parquet/*
```

4. Elimine ambas as tabelas (**jobs_txt** e **jobs_parquet**), pois você não precisará de nenhuma delas novamente.

5. Agora tente criar uma tabela usando dados de um arquivo Parquet existente. Quando terminar, guarde esta tabela, porque você a usará novamente mais tarde. (Se você usar a palavra-chave **EXTERNAL** conforme indicado abaixo, então, descartar a tabela não excluirá os dados, portanto, você pode eliminá-la agora e voltar e recriar a tabela mais tarde, se desejar.)

   a) Uma versão Parquet dos dados dos **investors** também é armazenada no HDFS, em **/user/hive/warehouse/investors_parquet** (que será o local padrão para uma tabela chamada **default.investors_parquet**). Examine o arquivo da mesma forma que você examinou o arquivo Parquet de jobs: Na janela Terminal, emita o comando

```
hdfs dfs -cat /user/hive/warehouse/investors_parquet/investors.parq
```

   Novamente, você verá que não está realmente em formato legível por humanos.

   b) Agora crie a tabela a partir do editor de consultas:

```
CREATE EXTERNAL TABLE default.investors_parquet
      (name STRING, amount INT, share DECIMAL(4,3))
  STORED AS PARQUET;
```

   c) Use o painel de origem de dados ou execute uma consulta **`SELECT *`** para verificar se o conteúdo da nova tabela está correto. (Deve ser idêntica à tabela de outros **`investors`** que você criou na leitura "A cláusula ROW FORMAT".)