Conteúdo

Este guia tem por objetivo realizar as seguintes atividades na plataforma GovData:

- Explorar no Laboratório de Big Data (Hue) os dados de uma tabela;
- Criar tabelas por meio da importação de arquivos CSV;
- Criar tabelas com base em consultas SQL;
- Cruzar bases de dados por meio de scripts SQL.

Pré-requisitos

Para seguir este guia é necessário ter acesso à ferramenta Hue na plataforma GovData.

Bases de dados

Serão utilizadas as seguintes bases de dados e arquivo no presente estudo:

Banco	Tabela	Descrição
tutorial	servidor	Cadastro de Servidores (dados fictícios)
	servidor_salario	Salários de Servidores (dados fictícios)

Baixar o arquivo do link abaixo para utilização nos exercícios que vem a seguir.

Arquivo	Descrição	Link
cbo.csv	Cadastro Brasileiro de Ocupação (extração parcial para fins do exercício)	http://govdata.gov.br/files/cbo.csv

Exploração dos dados

1. Acessar portal GovData

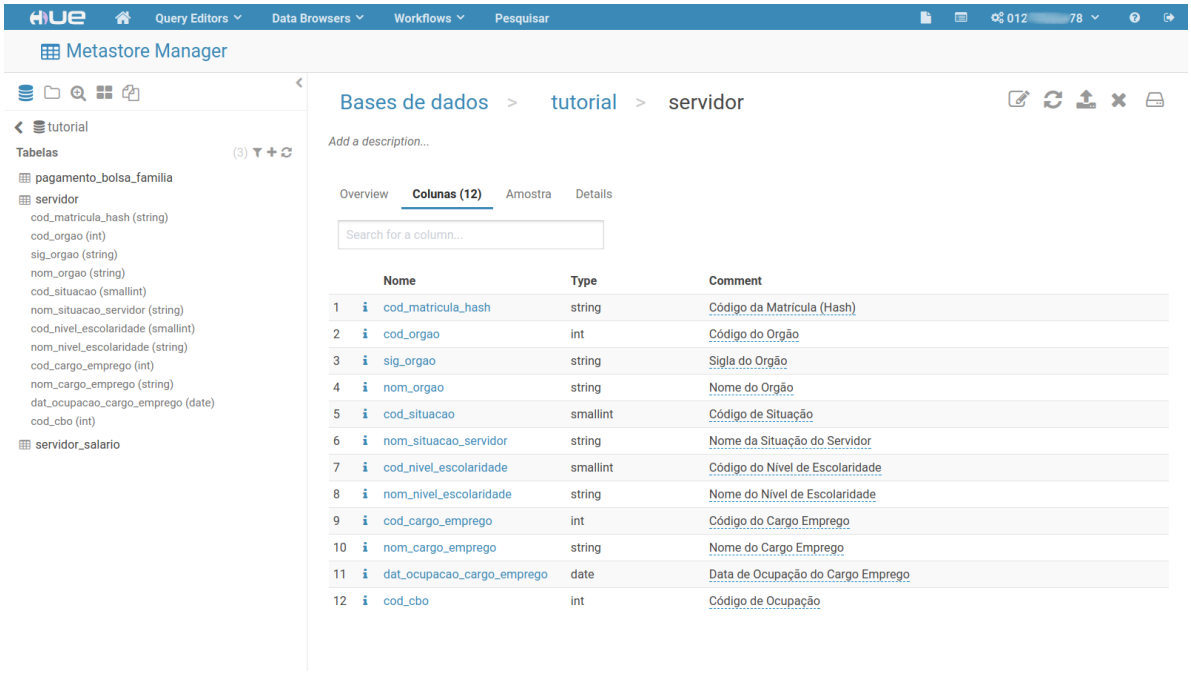
URL do portal: `http://govdata.gov.br`

2. Acessar o Laboratório de Big Data (Hue)

No portal GovData, clicar no link `Laboratório de Big Data (Hue)`. Em seguida, acessar a funcionalidade `Metastore Manager` do Hue e abrir o banco de dados `tutorial` e as tabelas `servidor` (Cadastro de Servidores) e `servidor_salario` (Folhas de Pagamento):

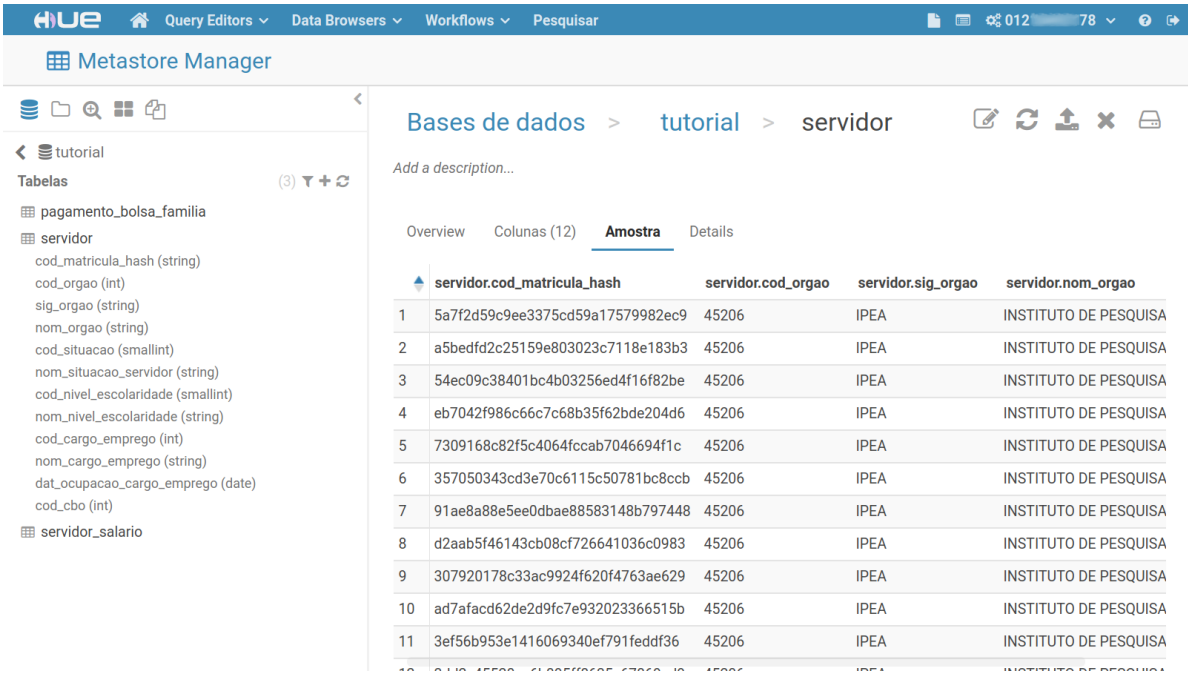
3. Conhecer os metadados

Clicar na opção `Colunas` a fim de conhecer os campos das duas tabelas e seus tipos de dados:



4. Explorar amostras de dados

Clicar na opção `Amostra` a fim de explorar exemplos de registros das tabelas:

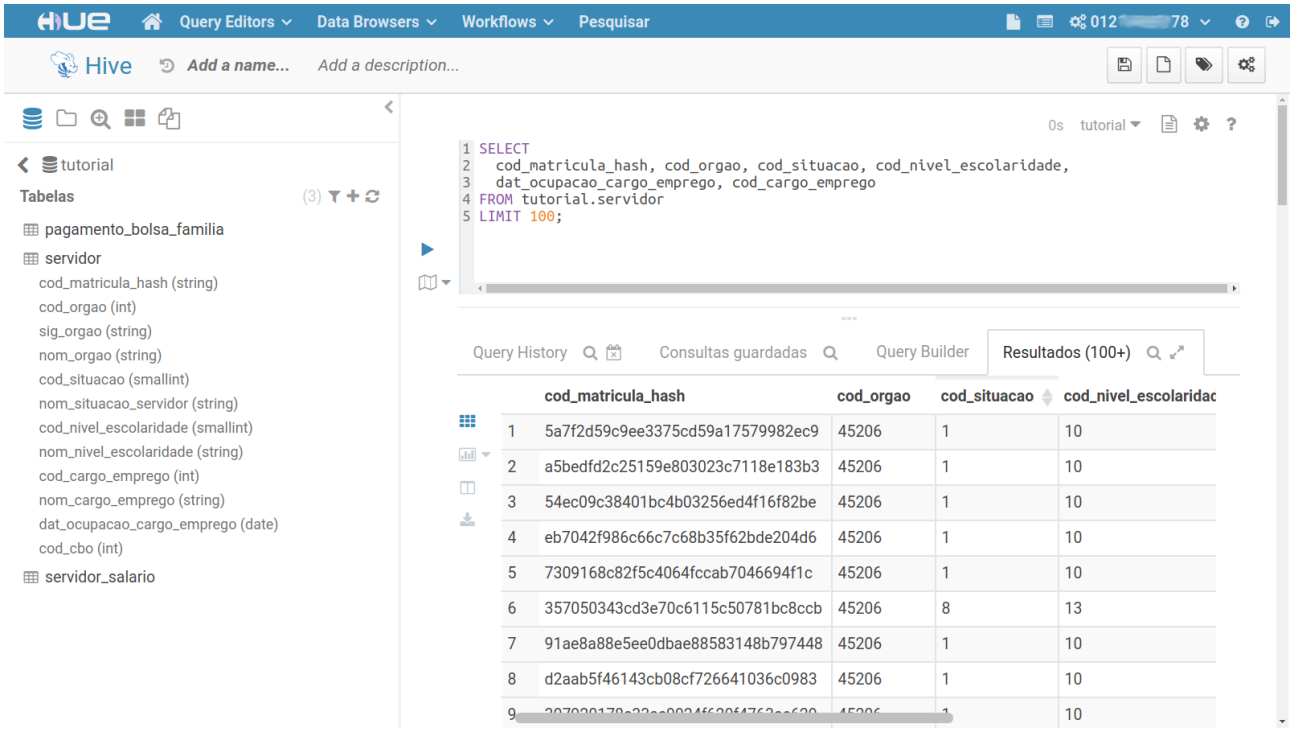


Experimentação de consultas

1. Utilizar o editor de consultas do Hive

Acessar a funcionalidade `Query Editors > Hive` no Hue. Em seguida, executar a seguinte instrução SQL:

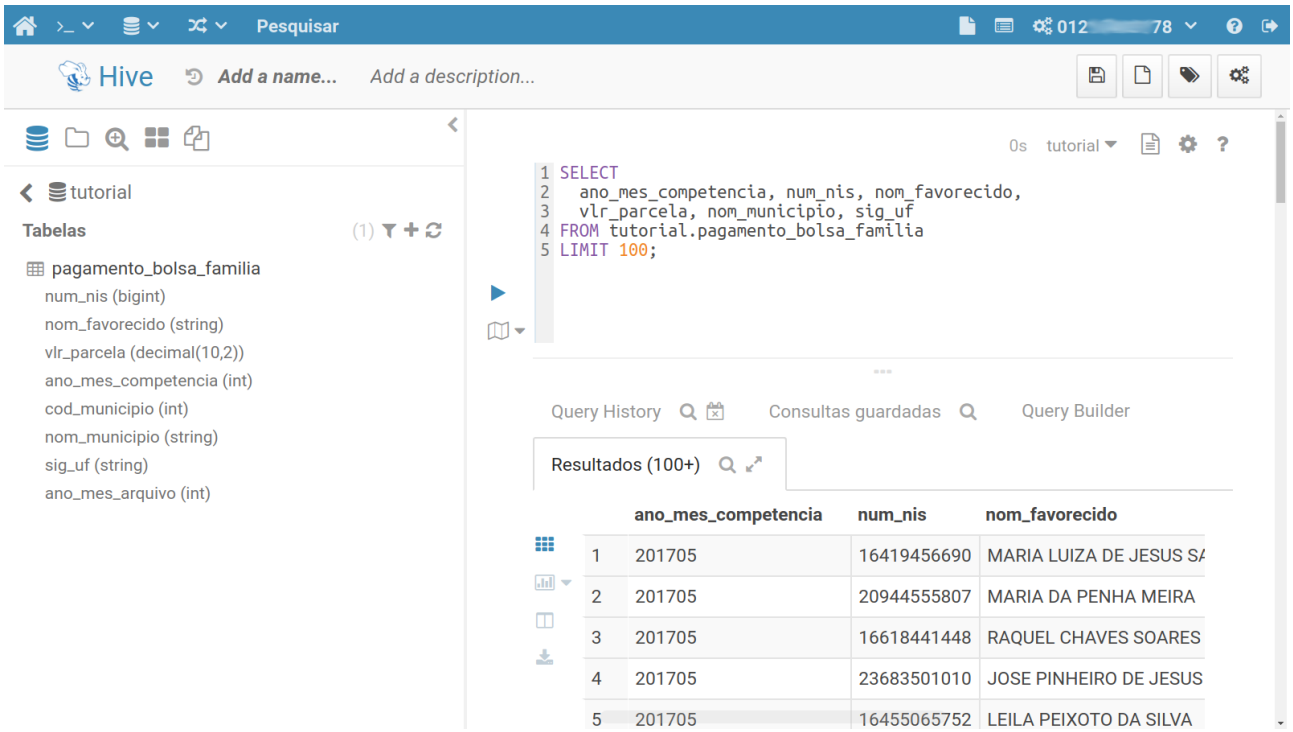
```
SELECT
  cod_matricula_hash, cod_orgao, cod_situacao, cod_nivel_escolaridade,
  dat_ocupacao_cargo_emprego, cod_cargo_emprego
FROM tutorial.servidor
LIMIT 100;
```



Use esse ambiente para avaliar o resultado das instruções SQL antes de especificá-las nas demais ferramentas do GovData.

Agora vamos executar a seguinte instrução SQL para visualizar os dados referentes a salário dos servidores:

```
SELECT cod_matricula_hash, qtd_tempo_servico, vlr_salario_medio
FROM tutorial.servidor_salario
LIMIT 100;
```



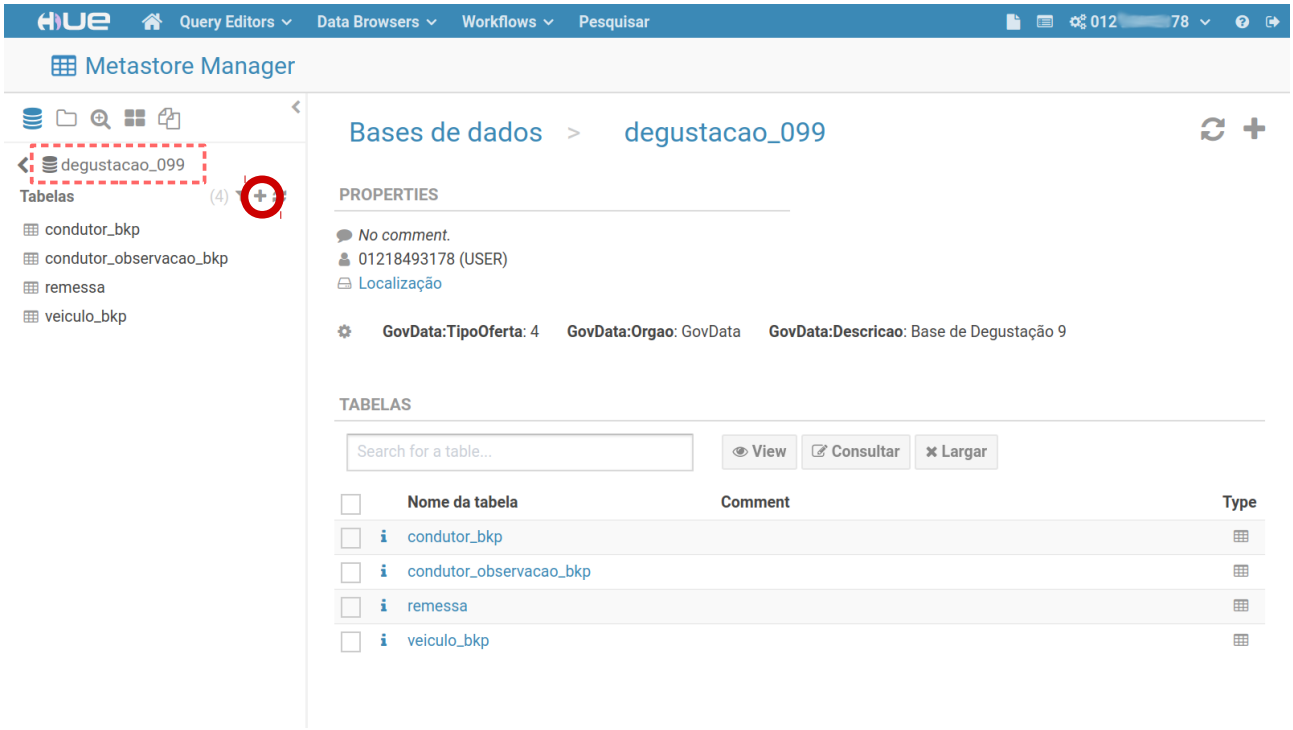
Use esse ambiente para avaliar o resultado das instruções SQL antes de especificá-las nas demais ferramentas do GovData.

Importando dados via arquivo CSV

1. Acessar o Laboratório de Big Data (Hue)

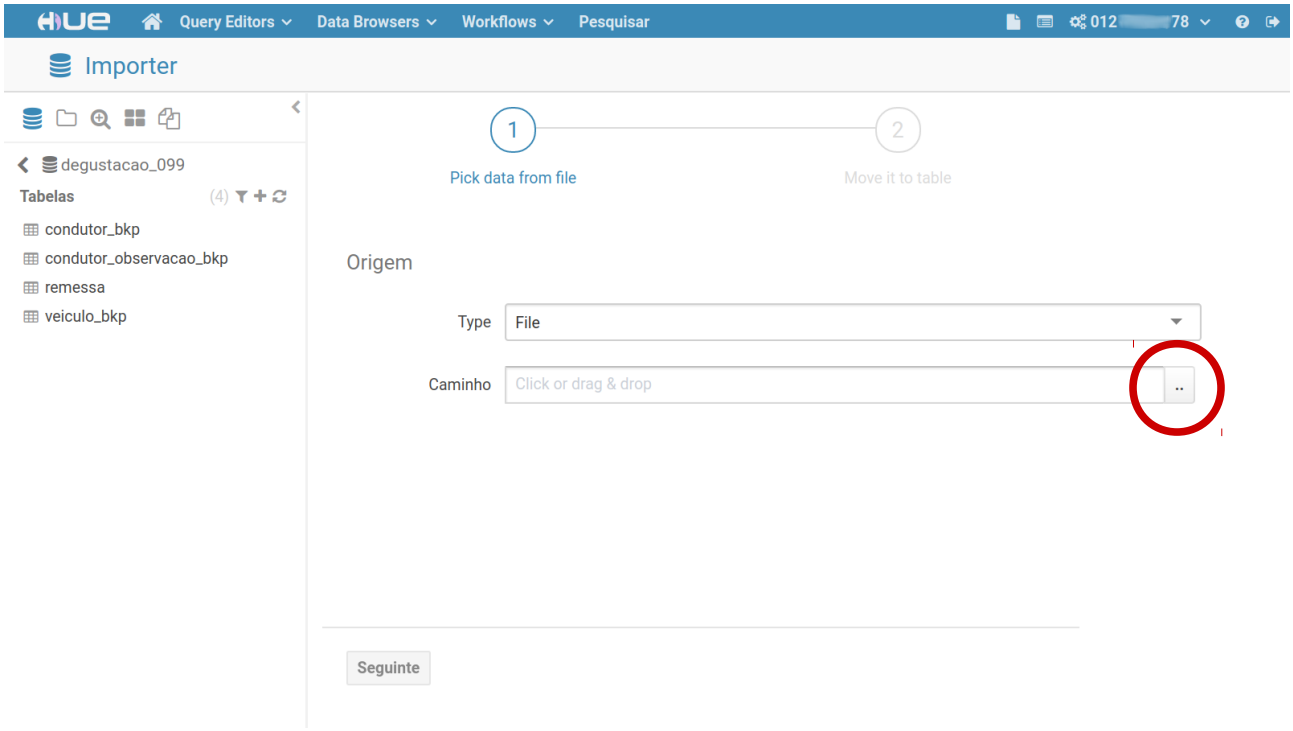
No portal GovData, clicar no link `Laboratório de Big Data (Hue)`. Em seguida, informar o usuário/senha e acessar a funcionalidade `Metastore Manager` do Hue:

Para fins ilustrativos, será utilizado o banco de dados `degustacao_099`, mas para sua prática utilize um dos seus bancos de dados com permissão de escrita.

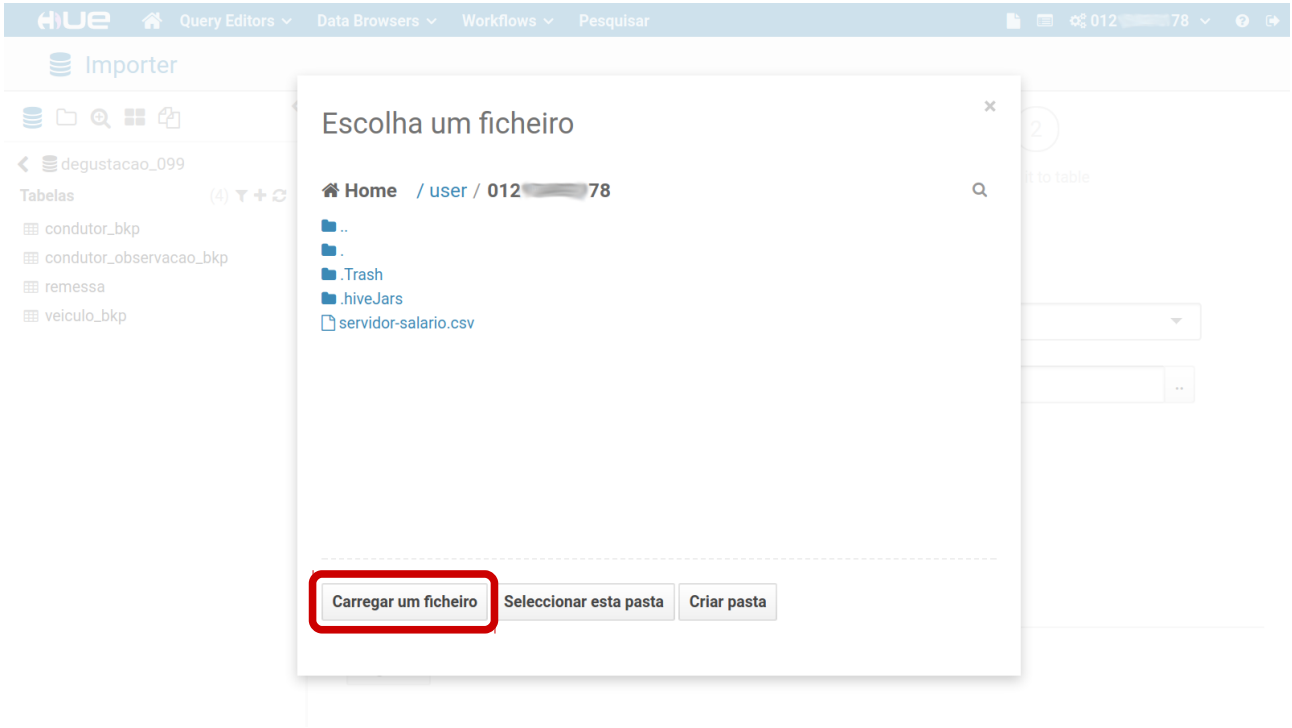


Certifique que está com a base de dados selecionada e clique no ícone "+" para começar o processo de adicionar uma nova tabela à base.

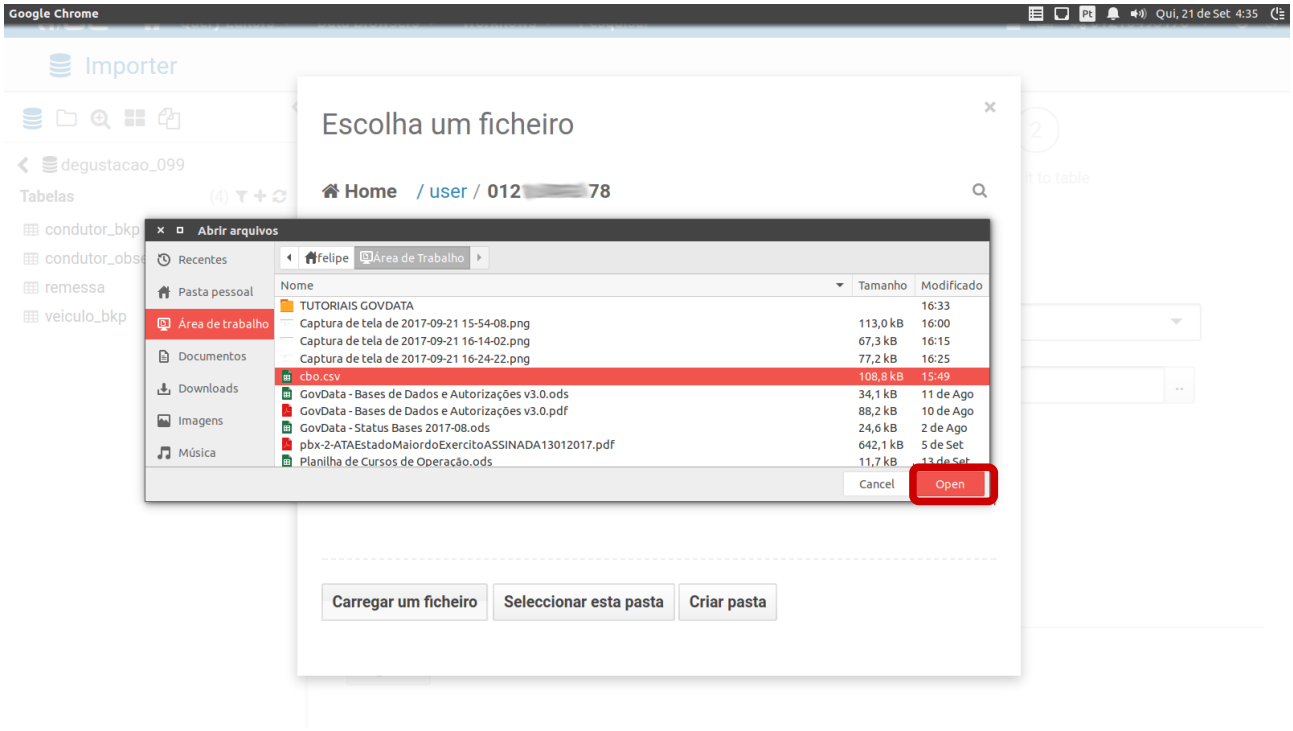
A aplicação solicitará o *upload* do aquivo que deseja importar:



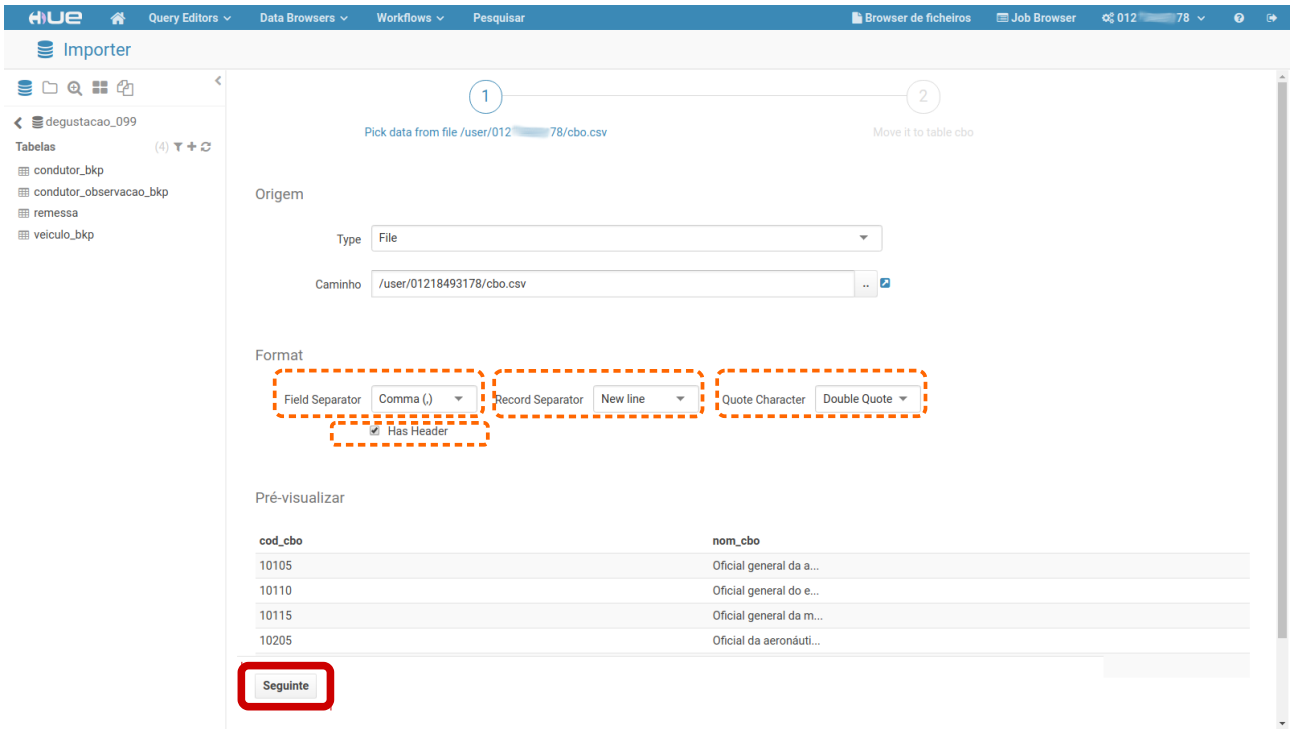
No campo "Caminho" arraste o arquivo até ele ou clique no ícone "..." para abrir uma caixa de diálogo para selecionar o aquivo a partir do sistema de arquivo do GovData:



Clique no botão "Carregar um ficheiro" para importar o arquivo "cbo.csv" localizado na sua estação de trabalho.

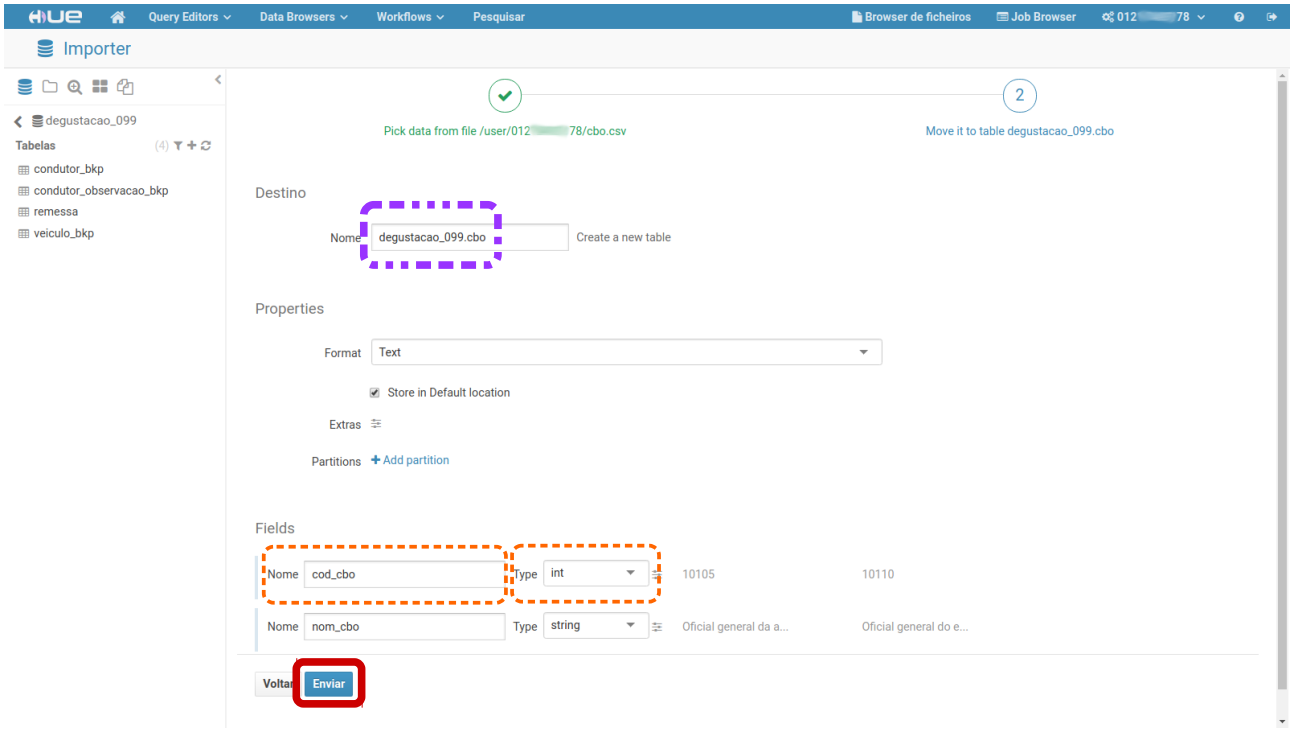


O sistema fará upload do arquivo para sua pasta do sistema de arquivo no GovData. Clique no nome do arquivo recém-importado para o sistema abrir na tela com a etapa de configuração para importação dos dados do arquivo para tabela na base selecionada.



Na **etapa 1** (acima), configuramos os parâmetros para interpretação do arquivo, identificação de cabeçalho e pré-visualização da interpretação do arquivo.

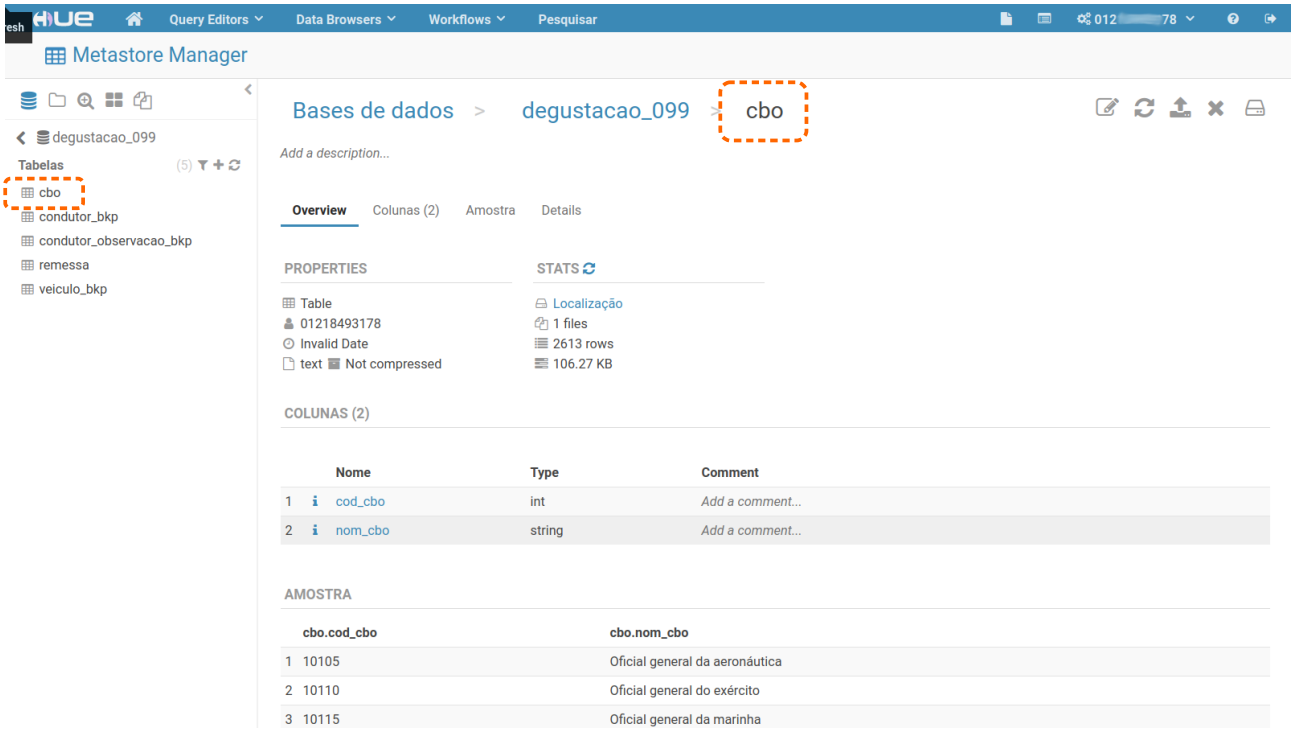
Para confirmar a conversão clique em “Seguinte”:



Na **etapa 2** (acima), configuramos o nome da base de dados e o nome da tabela que criaremos. Também estabelecemos o nome dos campos e as tipagens para cada um deles.

É **estritamente importante** que o nome do banco de dados preceda o nome da tabela a ser criada. Exemplo: “degustacao_099.cbo”. Lembrando de separá-los por “ponto”.

Concluídas as configurações, clique em “Enviar” para o sistema criar a tabela.



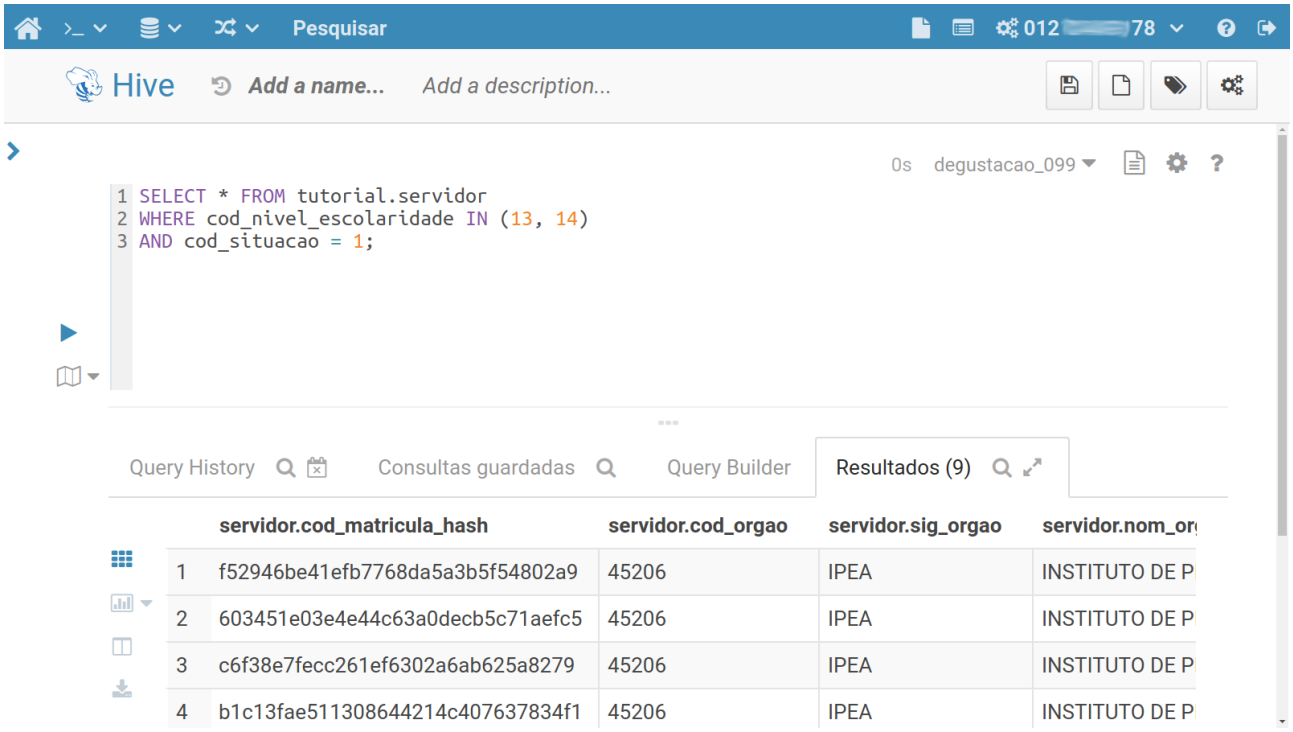
Pronto! Seus dados provenientes do arquivo CSV foram importados para a base de dados do GovData.

Criar bases por meio de consulta SQL

1. Criar consulta aos dados desejados

No `Query Editor`, criaremos uma consulta onde buscaremos só por servidores "permanentemente ativos" e que tenham nível de escolaridade de Mestrado ou Doutorado:

```
SELECT *
FROM tutorial.servidor
WHERE cod_nivel_escolaridade IN (13, 14)
  AND cod_situacao = 1;
```

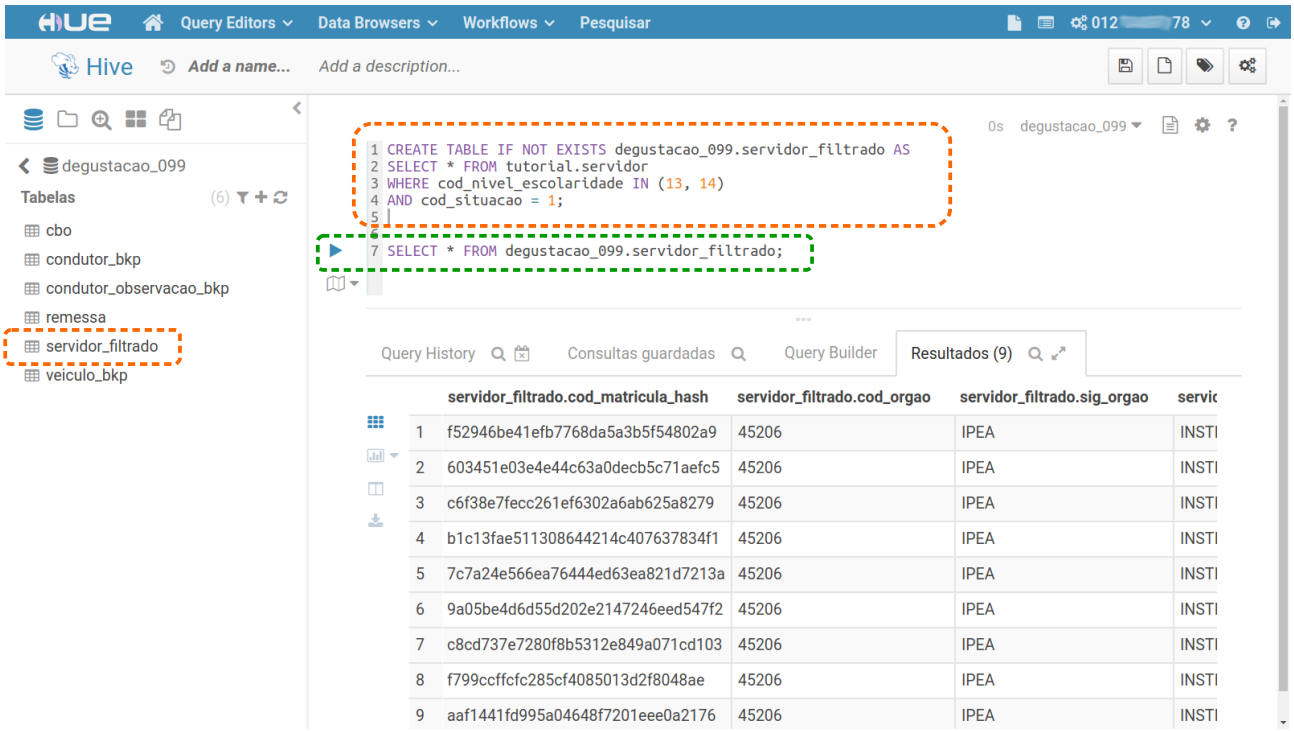


2. Criar tabela por meio do resultado da consulta

Anteriormente efetuamos uma consulta simples para trazer o resultado com os filtros desejados. Agora criaremos uma tabela com estes resultados da consulta. A *query SQL* para isso é:

```
CREATE TABLE IF NOT EXISTS degustacao_099.servidor_filtrado AS
SELECT *
FROM tutorial.servidor
WHERE cod_nivel_escolaridade IN (13, 14)
  AND cod_situacao = 1;
```

Lembre-se de **mudar o nome do banco** de dados `degustacao_099` para o nome sua base própria ou de degustação aonde tenha permissão de escrita.



Se tudo ocorrer corretamente, a lista de tabelas será atualizada com o nome da nova tabela e para fins de confirmação efetuamos uma consulta simples à tabela para ver o resultado final.

Cruzamento de bases de dados

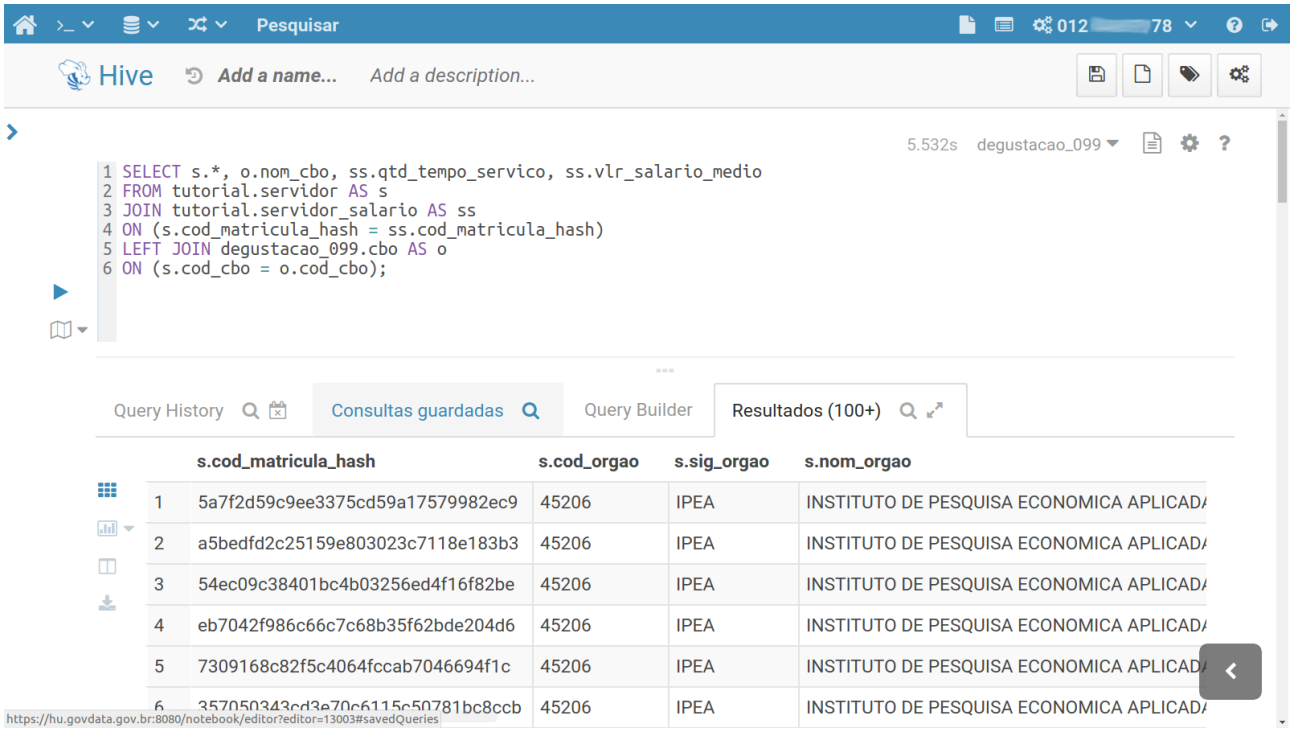
1. Unindo tabelas para agregar informações

Como exemplo de cruzamento de dados agregaremos os dados dos servidores (tabela *tutorial.servidor*), seus dados salariais (tabela *tutorial.servidor_salario*) e a informação da sua ocupação (tabela que importamos do CSV – *degustacao_099.cbo*).

A agregação destas tabelas é feita pelo comando SQL JOIN, onde utilizamos as chaves em comum entre as tabelas para fazer o batimento:

```
SELECT s.*, o.nom_cbo, ss.qtd_tempo_servico, ss.vlr_salario_medio
FROM tutorial.servidor AS s
  JOIN tutorial.servidor_salario AS ss
    ON (s.cod_matricula_hash = ss.cod_matricula_hash)
  LEFT JOIN degustacao_099.cbo AS o
    ON (s.cod_cbo = o.cod_cbo);
```

Nesse caso, utilizamos o campo *cod_matricula_hash* para fazer o batimento do servidor com seus dados salariais e o campo *cod_cbo* para o batimento do nome da ocupação do servidor.



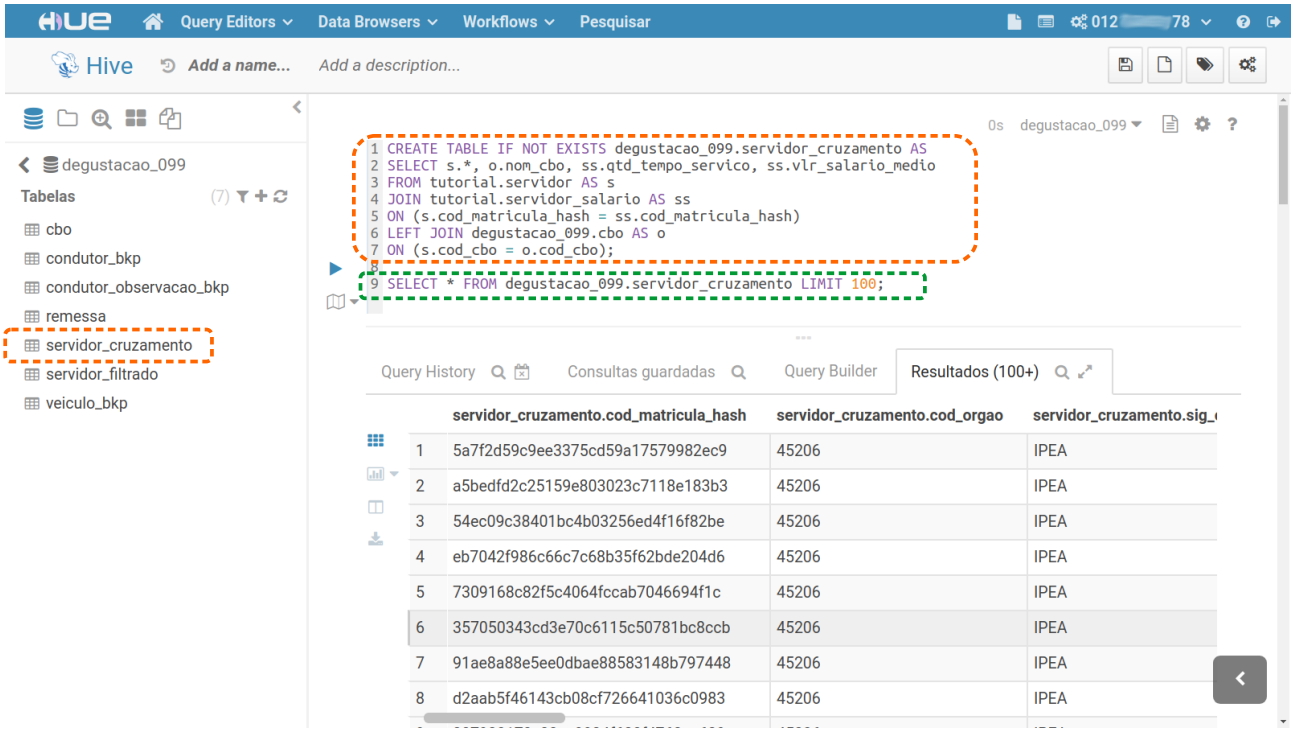
A cláusula LEFT JOIN no cruzamento da ocupação que significa que, caso não haja batimento para o valor da chave à esquerda (tabela *servidor.cod_cbo*), os dados do servidor ainda serão considerados, porém a respectiva coluna “nom_cbo” ficará nula.

2. Criando uma nova tabela com o resultado do cruzamento

Para gerar uma nova tabela com o resultado do cruzamento acima é só colocar o mesmo comando ensinado anteriormente:

```
CREATE TABLE IF NOT EXISTS degustação_099.servidor_cruzamento AS
SELECT s.*, o.nom_cbo, ss.qtd_tempo_servico, ss.vlr_salario_medio
FROM tutorial.servidor AS s
  JOIN tutorial.servidor_salario AS ss
    ON (s.cod_matricula_hash = ss.cod_matricula_hash)
  LEFT JOIN degustacao_099.cbo AS o
    ON (s.cod_cbo = o.cod_cbo);
```

Lembre-se de **mudar o nome do banco** de dados `degustacao_099` para o nome sua base própria ou de degustação aonde tenha permissão de escrita.



Executada a consulta SQL, será criada uma nova tabela com base no cruzamento! Executamos uma consulta simples no final para verificar os dados. Você pode conferir via ferramenta *Data Browser → Tabelas Metastore.*