

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:

Metastore Manager

Bases de dados > tutorial > servidor

Overview **Colunas (12)** Amostra Details

Search for a column...

	Nome	Type	Comment
1	i cod_matricula_hash	string	Código da Matrícula (Hash)
2	i cod_orgao	int	Código do Orgão
3	i sig_orgao	string	Sigla do Orgão
4	i nom_orgao	string	Nome do Orgão
5	i cod_situacao	smallint	Código de Situação
6	i nom_situacao_servidor	string	Nome da Situação do Servidor
7	i cod_nivel_escolaridade	smallint	Código do Nível de Escolaridade
8	i nom_nivelEscolaridade	string	Nome do Nível de Escolaridade
9	i cod_cargo_emplo	int	Código do Cargo Emprego
10	i nom_cargo_emplo	string	Nome do Cargo Emprego
11	i dat_ocupacao_cargo_emplo	date	Data de Ocupação do Cargo Emprego
12	i cod_cbo	int	Código de Ocupação

4. Explorar amostras de dados

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

Metastore Manager

Bases de dados > tutorial > servidor

Overview Colunas (12) **Amostra** Details

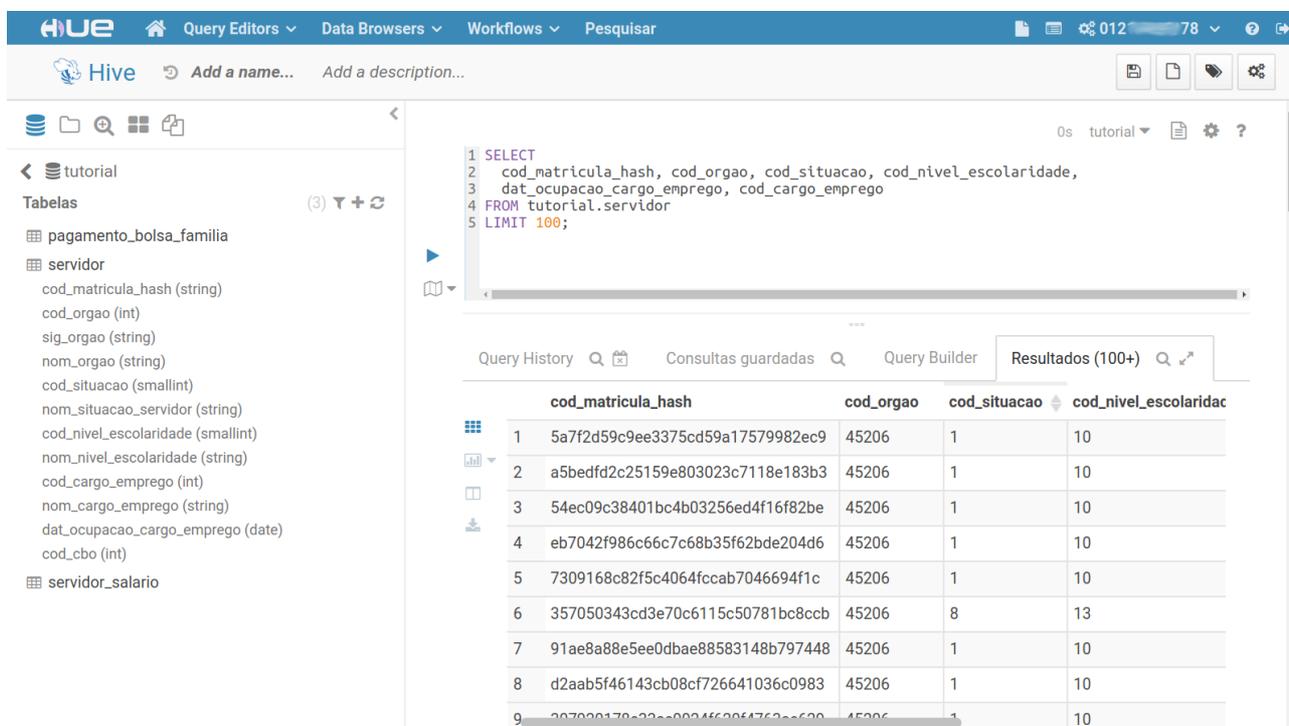
	servidor.cod_matricula_hash	servidor.cod_orgao	servidor.sig_orgao	servidor.nom_orgao
1	5a7f2d59c9ee3375cd59a17579982ec9	45206	IPEA	INSTITUTO DE PESQUISA
2	a5bedfd2c25159e803023c7118e183b3	45206	IPEA	INSTITUTO DE PESQUISA
3	54ec09c38401bc4b03256ed4f16f82be	45206	IPEA	INSTITUTO DE PESQUISA
4	eb7042f986c66c7c68b35f62bde204d6	45206	IPEA	INSTITUTO DE PESQUISA
5	7309168c82f5c4064fccab7046694f1c	45206	IPEA	INSTITUTO DE PESQUISA
6	357050343cd3e70c6115c50781bc8ccb	45206	IPEA	INSTITUTO DE PESQUISA
7	91ae8a88e5ee0dbae88583148b797448	45206	IPEA	INSTITUTO DE PESQUISA
8	d2aab5f46143cb08cf726641036c0983	45206	IPEA	INSTITUTO DE PESQUISA
9	307920178c33ac9924f620f4763ae629	45206	IPEA	INSTITUTO DE PESQUISA
10	ad7afacd62de2d9fc7e932023366515b	45206	IPEA	INSTITUTO DE PESQUISA
11	3ef56b953e1416069340ef791feddf36	45206	IPEA	INSTITUTO DE PESQUISA

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;
```



The screenshot shows the Hue Query Editor interface. The top navigation bar includes 'HUE', 'Query Editors', 'Data Browsers', 'Workflows', and 'Pesquisar'. The main area displays a SQL query in a text editor, which is the same query shown in the previous block. Below the editor, the 'Query History' and 'Consultas guardadas' sections are visible. The 'Query Builder' section shows the results of the query, labeled 'Resultados (100+)'. The results are displayed in a table with the following columns: 'cod_matricula_hash', 'cod_orgao', 'cod_situacao', and 'cod_nivel_escolaridade'. The table contains 9 rows of data.

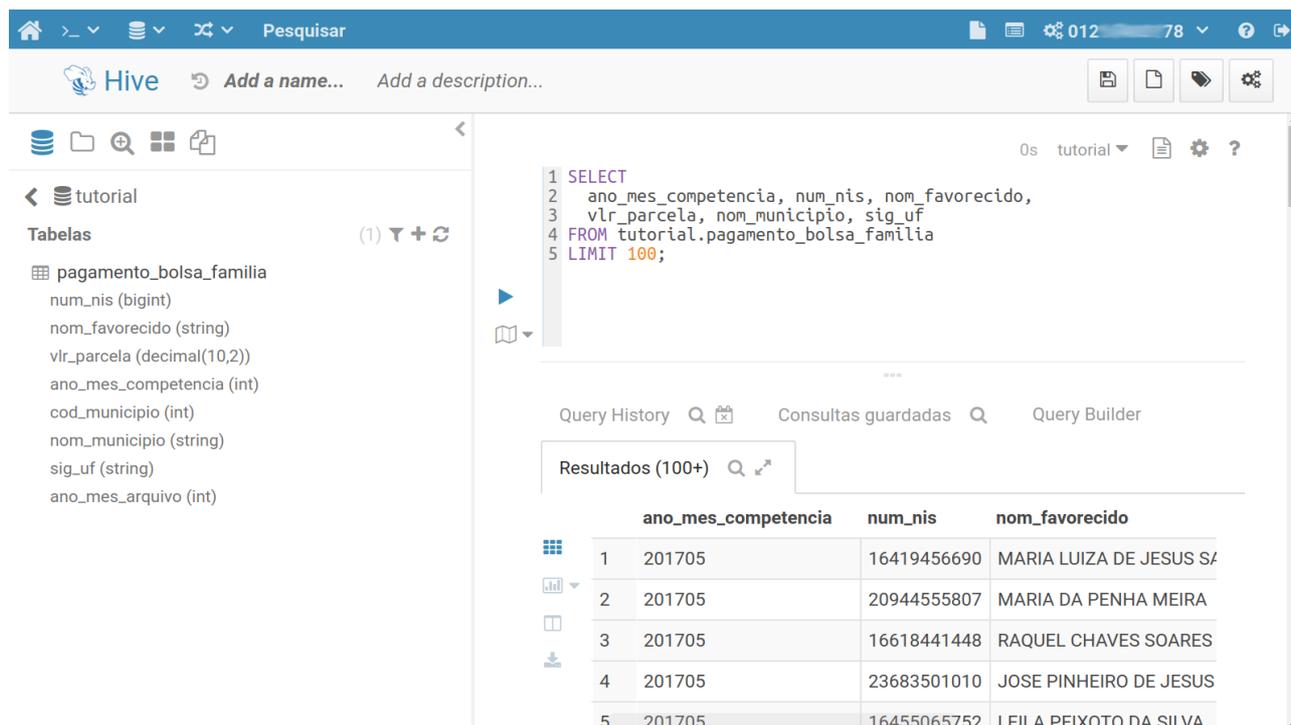
	cod_matricula_hash	cod_orgao	cod_situacao	cod_nivel_escolaridade
1	5a7f2d59c9ee3375cd59a17579982ec9	45206	1	10
2	a5bedfd2c25159e803023c7118e183b3	45206	1	10
3	54ec09c38401bc4b03256ed4f16f82be	45206	1	10
4	eb7042f986c66c7c68b35f62bde204d6	45206	1	10
5	7309168c82f5c4064fccab7046694f1c	45206	1	10
6	357050343cd3e70c6115c50781bc8ccb	45206	8	13
7	91ae8a88e5ee0dbae88583148b797448	45206	1	10
8	d2aab5f46143cb08cf726641036c0983	45206	1	10
9	207000170-22e-000466064762e-620	45206	1	10



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;
```



The screenshot shows the Hue Query Editor interface. The top bar includes navigation icons and the text 'Pesquisar'. Below the bar, the 'Hive' logo and 'Add a name...' and 'Add a description...' fields are visible. The left sidebar shows a tree view with 'tutorial' selected, listing tables: 'pagamento_bolsa_familia' and its columns: 'num_nis (bigint)', 'nom_favorecido (string)', 'vlr_parcela (decimal(10,2))', 'ano_mes_competencia (int)', 'cod_municipio (int)', 'nom_municipio (string)', 'sig_uf (string)', and 'ano_mes_arquivo (int)'. The main editor area contains the following SQL query:

```
1 SELECT  
2   ano_mes_competencia, num_nis, nom_favorecido,  
3   vlr_parcela, nom_municipio, sig_uf  
4 FROM tutorial.pagamento_bolsa_familia  
5 LIMIT 100;
```

Below the query editor, there are buttons for 'Query History', 'Consultas guardadas', and 'Query Builder'. The results section shows 'Resultados (100+)' and a table with the following data:

	ano_mes_competencia	num_nis	nom_favorecido
1	201705	16419456690	MARIA LUIZA DE JESUS SA
2	201705	20944555807	MARIA DA PENHA MEIRA
3	201705	16618441448	RAQUEL CHAVES SOARES
4	201705	23683501010	JOSE PINHEIRO DE JESUS
5	201705	16455065752	LEILA PEIXOTO DA SILVA



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

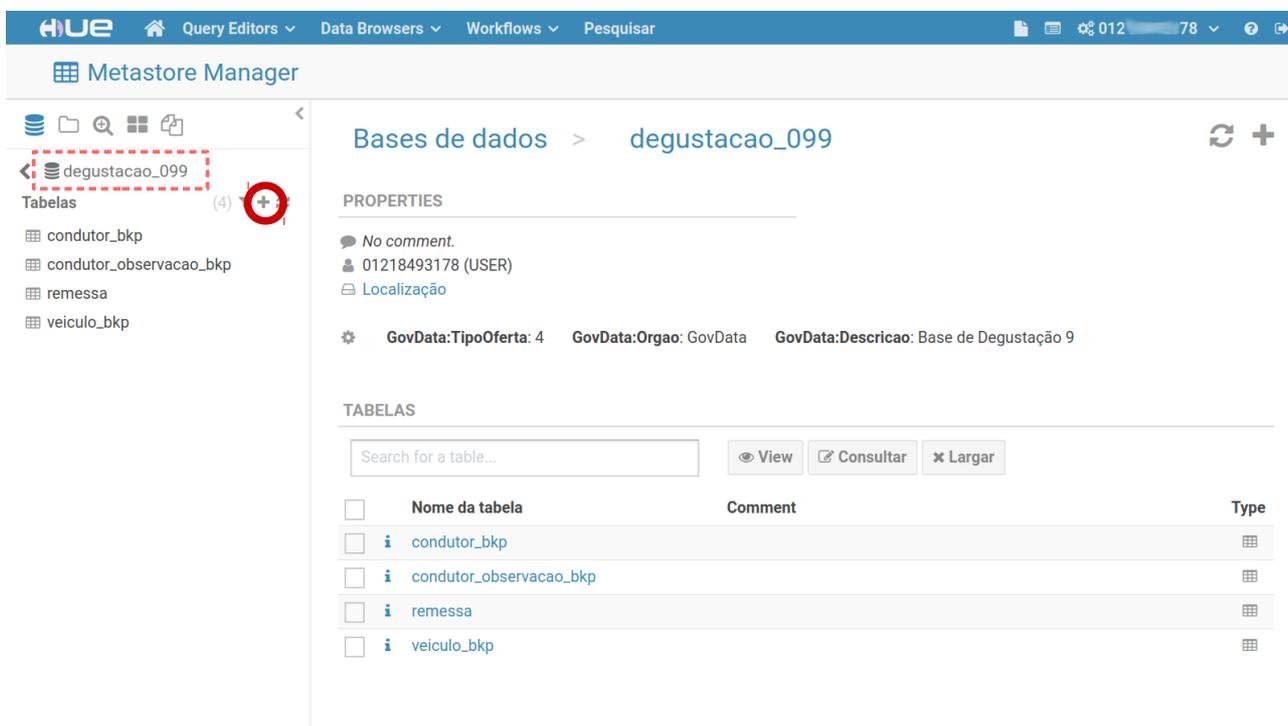
Importando dados via arquivo CSV

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:

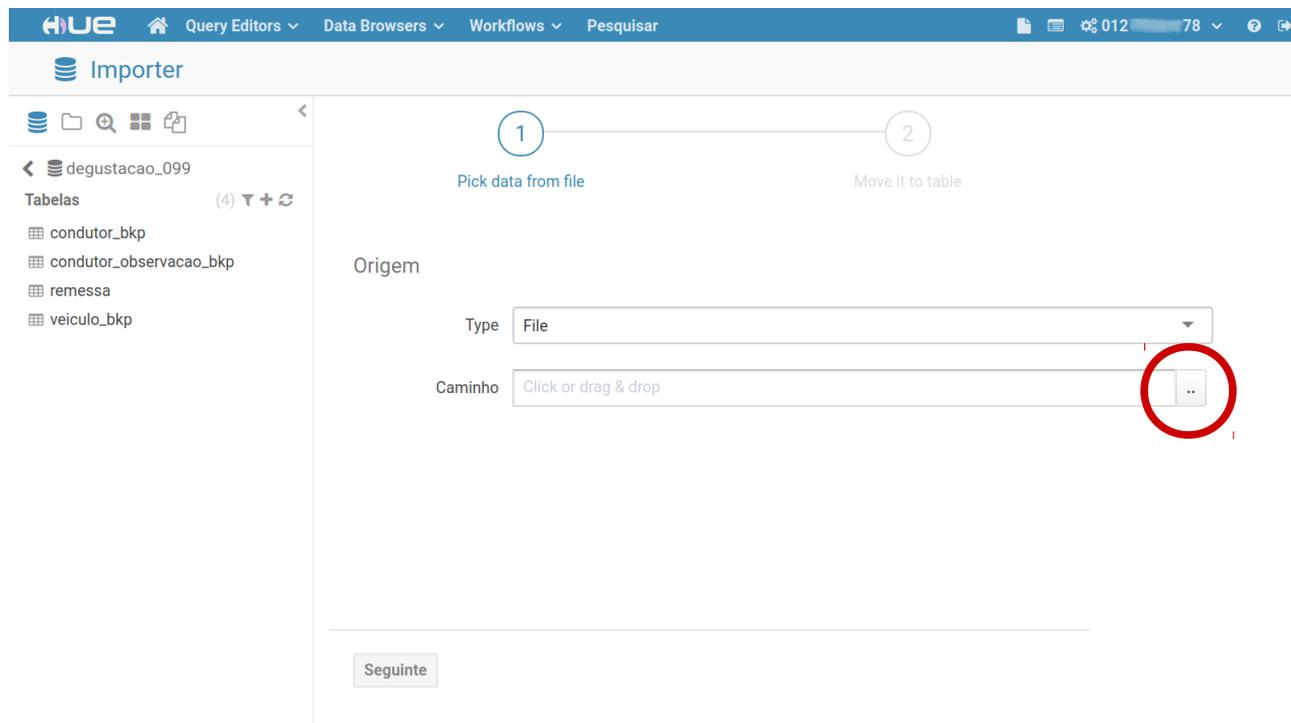


Para fins ilustrativos, será utilizado o banco de dados degustacao_099, mas para sua prática utilize um dos seus bancos de dados com permissão de escrita.

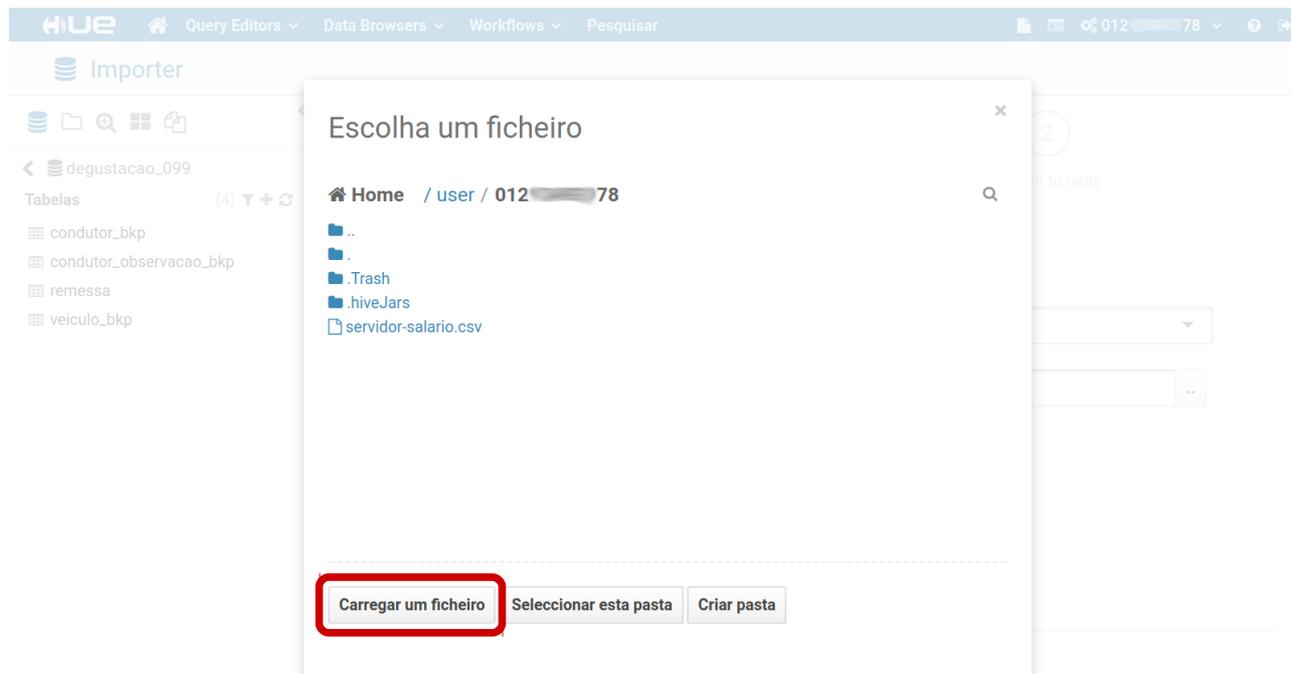


Certifique que está com a base de dados selecionada e clique no ícone “+” para começar o processo de adicionar uma nova tabela à base.

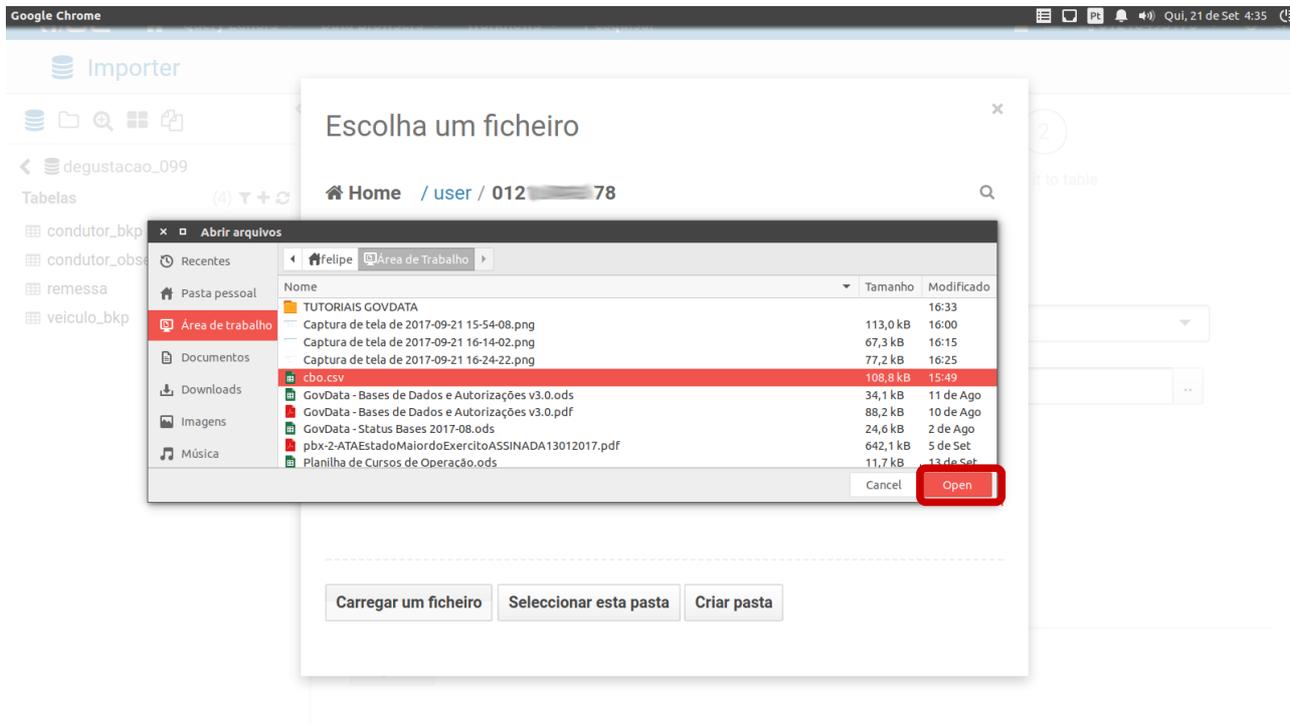
A aplicação solicitará o *upload* do arquivo que deseja importar:



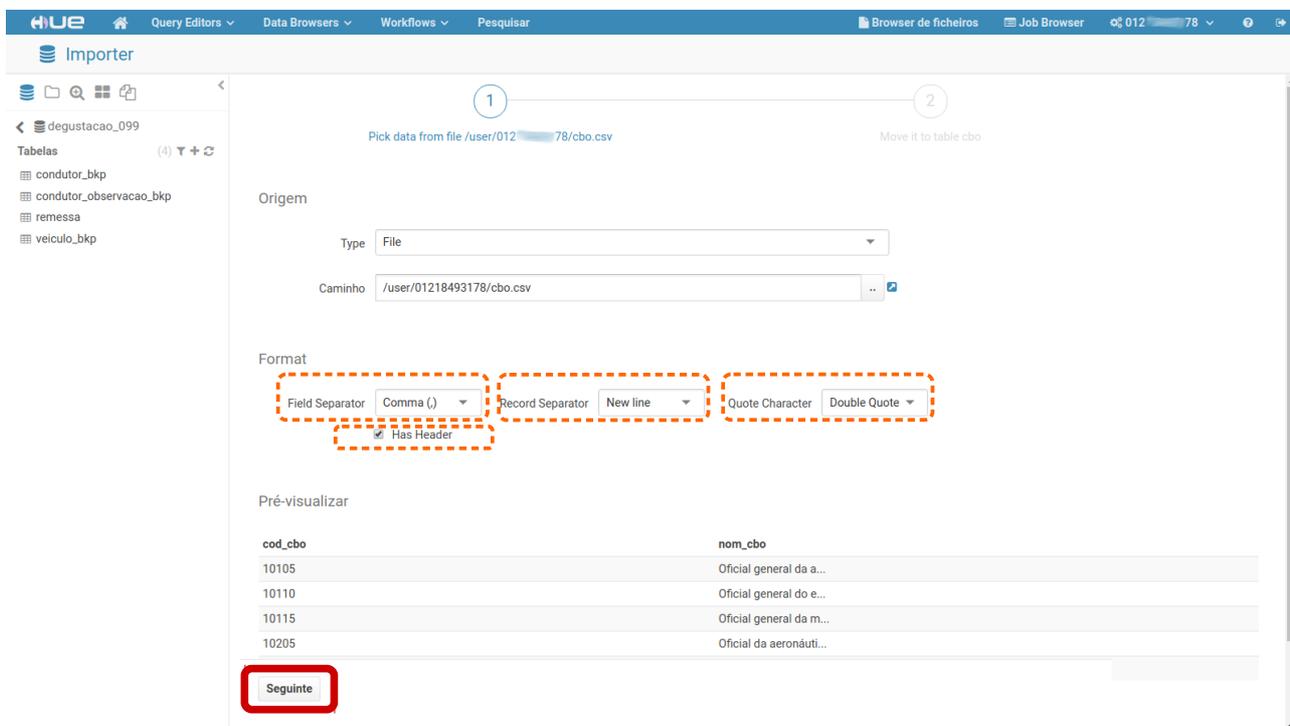
No campo “Caminho” arraste o arquivo até ele ou clique no ícone “...” para abrir uma caixa de diálogo para seleccionar o arquivo a partir do sistema de arquivo do GovData:



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”:

The screenshot shows the Hue 'Importer' interface. The 'Destino' section has a 'Nome' field containing 'degustacao_099.cbo' with a 'Create a new table' button next to it. The 'Properties' section shows 'Format' set to 'Text', 'Store in Default location' checked, and 'Partitions' set to 'Add partition'. The 'Fields' section shows two fields: 'cod_cbo' with type 'int' and 'nom_cbo' with type 'string'. At the bottom, there are 'Voltar' and 'Enviar' buttons, with 'Enviar' highlighted in a red box.

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.

The screenshot shows the Hue Metastore Manager interface. The breadcrumb navigation indicates the path: Bases de dados > degustacao_099 > cbo. The table 'cbo' is selected in the left sidebar. The main panel displays the following information:

- Overview:** Columns (2), Amostra, Details
- PROPERTIES:**
 - Table: 01218493178
 - Invalid Date
 - text, Not compressed
 - Localização: 1 files
 - 2613 rows
 - 106.27 KB
- COLUMNAS (2):**

	Nome	Type	Comment
1	<i>i</i> cod_cbo	int	Add a comment...
2	<i>i</i> nom_cbo	string	Add a comment...

AMOSTRA

	cbo.cod_cbo	cbo.nom_cbo
1	10105	Oficial general da aeronáutica
2	10110	Oficial general do exército
3	10115	Oficial general da marinha

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;
```

The screenshot shows the Hue Query Editor interface. At the top, there is a navigation bar with a search icon and the text "Pesquisar". Below this, the Hive logo is visible along with "Add a name..." and "Add a description..." options. The main area contains a SQL query editor with the following text:

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

Below the query editor, there is a section for "Resultados (9)" which displays a table with the following data:

	servidor.cod_matricula_hash	servidor.cod_orgao	servidor.sig_orgao	servidor.nom_or
1	f52946be41efb7768da5a3b5f54802a9	45206	IPEA	INSTITUTO DE P
2	603451e03e4e44c63a0decb5c71aefc5	45206	IPEA	INSTITUTO DE P
3	c6f38e7fecc261ef6302a6ab625a8279	45206	IPEA	INSTITUTO DE P
4	b1c13fae511308644214c407637834f1	45206	IPEA	INSTITUTO DE P

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.

The screenshot shows the Hue Query Editor interface. The SQL query is entered in the editor and executed. The results are displayed in a table with 9 rows and 4 columns: `servidor_filtrado.cod_matricula_hash`, `servidor_filtrado.cod_orgao`, `servidor_filtrado.sig_orgao`, and `servic`.

	<code>servidor_filtrado.cod_matricula_hash</code>	<code>servidor_filtrado.cod_orgao</code>	<code>servidor_filtrado.sig_orgao</code>	<code>servic</code>
1	f52946be41efb7768da5a3b5f54802a9	45206	IPEA	INSTI
2	603451e03e4e44c63a0decb5c71aefc5	45206	IPEA	INSTI
3	c6f38e7fecc261ef6302a6ab625a8279	45206	IPEA	INSTI
4	b1c13fae511308644214c407637834f1	45206	IPEA	INSTI
5	7c7a24e566ea76444ed63ea821d7213a	45206	IPEA	INSTI
6	9a05be4d6d55d202e214724eed547f2	45206	IPEA	INSTI
7	c8cd737e7280f8b5312e849a071cd103	45206	IPEA	INSTI
8	f799ccfffc285cf4085013d2f8048ae	45206	IPEA	INSTI
9	aaf1441fd995a04648f7201eee0a2176	45206	IPEA	INSTI

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

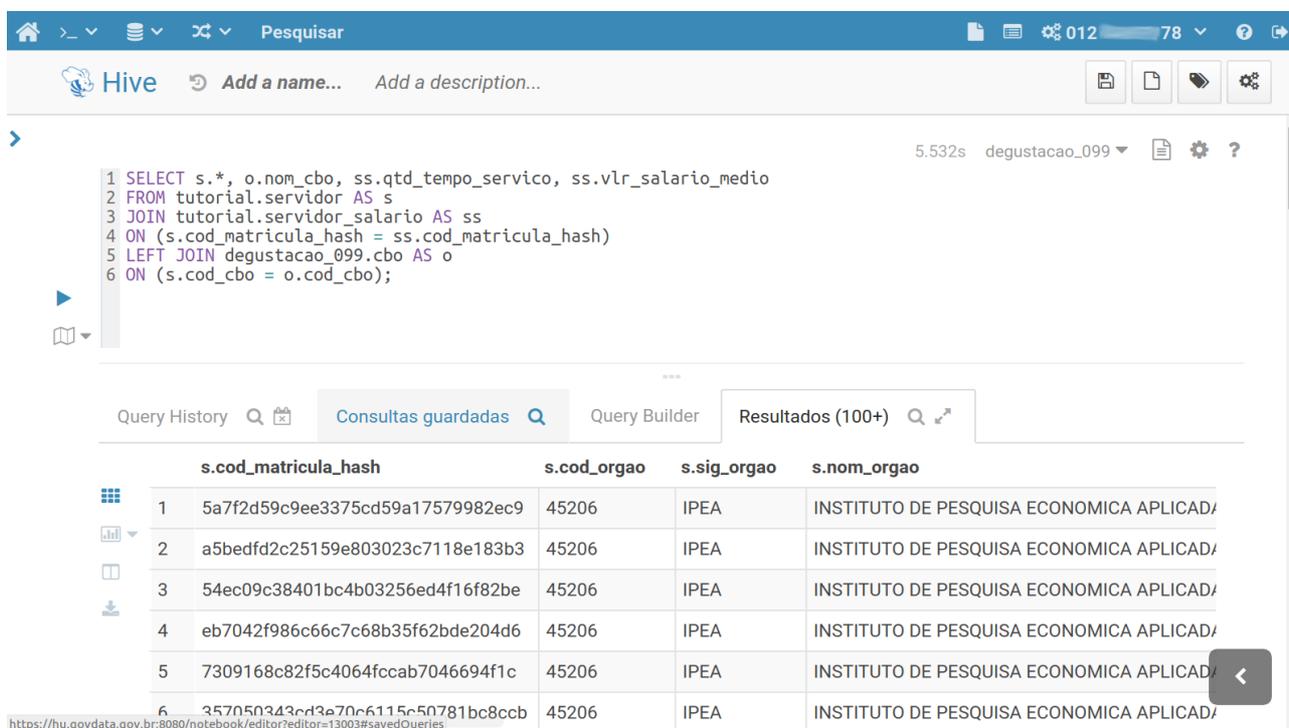
1. Unindo tabelas para agregar informações

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.



The screenshot shows the Hue Query Editor interface. At the top, there is a navigation bar with a search icon and the text "Pesquisar". Below that, the Hive logo and "Add a name..." and "Add a description..." buttons are visible. The main area displays a SQL query with line numbers 1 through 6. The query is: `1 SELECT s.*, o.nom_cbo, ss.qtd_tempo_servico, ss.vlr_salario_medio`, `2 FROM tutorial.servidor AS s`, `3 JOIN tutorial.servidor_salario AS ss`, `4 ON (s.cod_matricula_hash = ss.cod_matricula_hash)`, `5 LEFT JOIN degustacao_099.cbo AS o`, `6 ON (s.cod_cbo = o.cod_cbo);`. Below the query, there is a "Query History" section with "Consultas guardadas" and "Query Builder" tabs. The "Resultados (100+)" tab is active, showing a table with 6 rows and 4 columns: *s.cod_matricula_hash*, *s.cod_orgao*, *s.sig_orgao*, and *s.nom_orgao*. The first row shows values: 1, 5a7f2d59c9ee3375cd59a17579982ec9, 45206, IPEA, INSTITUTO DE PESQUISA ECONOMICA APLICAD/.

	s.cod_matricula_hash	s.cod_orgao	s.sig_orgao	s.nom_orgao
1	5a7f2d59c9ee3375cd59a17579982ec9	45206	IPEA	INSTITUTO DE PESQUISA ECONOMICA APLICAD/
2	a5bedfd2c25159e803023c7118e183b3	45206	IPEA	INSTITUTO DE PESQUISA ECONOMICA APLICAD/
3	54ec09c38401bc4b03256ed4f16f82be	45206	IPEA	INSTITUTO DE PESQUISA ECONOMICA APLICAD/
4	eb7042f986c66c7c68b35f62bde204d6	45206	IPEA	INSTITUTO DE PESQUISA ECONOMICA APLICAD/
5	7309168c82f5c4064fccab7046694f1c	45206	IPEA	INSTITUTO DE PESQUISA ECONOMICA APLICAD/
6	357050343cd3e70c6115c50781bc8ccb	45206	IPEA	INSTITUTO DE PESQUISA ECONOMICA APLICAD/

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.

The screenshot shows the Hue Query Editor interface. The SQL query is displayed in the editor, with the first part (CREATE TABLE and JOIN clauses) highlighted in orange and the final SELECT statement highlighted in green. The results table is shown below the query, displaying columns for cod_matricula_hash, cod_orgao, and sig_t.

	servidor_cruzamento.cod_matricula_hash	servidor_cruzamento.cod_orgao	servidor_cruzamento.sig_t
1	5a7f2d59c9ee3375cd59a17579982ec9	45206	IPEA
2	a5bedfd2c25159e803023c7118e183b3	45206	IPEA
3	54ec09c38401bc4b03256ed4f16f82be	45206	IPEA
4	eb7042f986c66c7c68b35f62bde204d6	45206	IPEA
5	7309168c82f5c4064fccab7046694f1c	45206	IPEA
6	357050343cd3e70c6115c50781bc8ccb	45206	IPEA
7	91ae8a88e5ee0dbae88583148b797448	45206	IPEA
8	d2aab5f46143cb08cf726641036c0983	45206	IPEA

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*.