

DATABRICKS

Grimaldo Oliveira



O que é o DATABRICS

1. Uma poderosa plataforma de colaboração entre os profissionais da área de dados. É uma plataforma fácil de usar para aqueles que desejam executar consultas em seu Data Lake.
2. Podem ser criados vários tipos de visualização para explorar resultados de consultas de diferentes perspectivas, podendo construir e compartilhar dashboards.
3. Você pode ser administrador da plataforma, analista de dados, cientista de dados e engenheiro de dados com diversas funcionalidade.

Sites Importantes

DataBricks

<https://databricks.com/>

Documentação

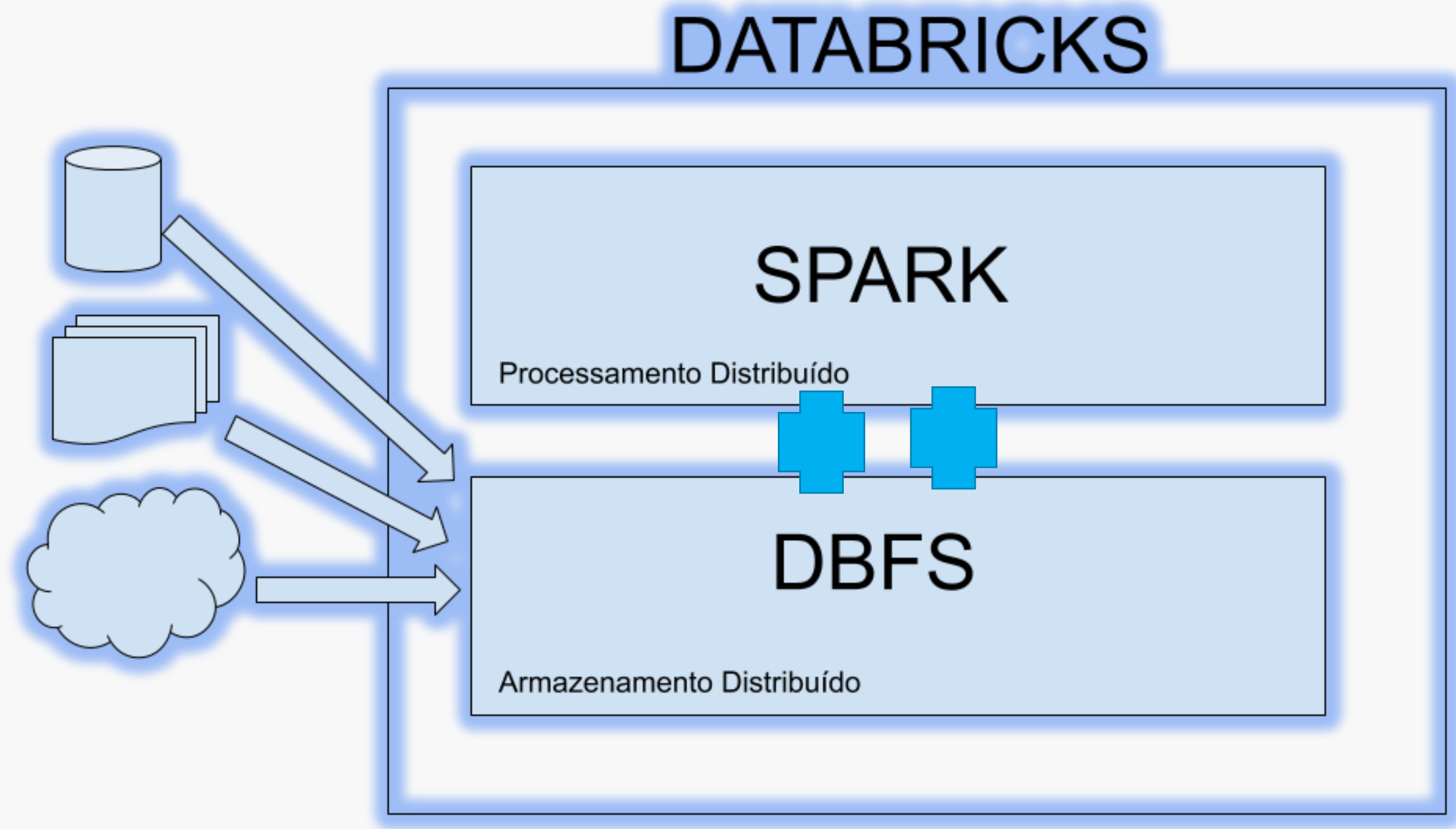
<https://docs.databricks.com/>

**Databricks
Community**

<https://community.cloud.databricks.com/login.html>

Ecossistema

Como funciona a leitura dos dados para serem carregados no Databricks Store e gerenciado pelo SPARK.



Databricks

Uma plataforma aberta para armazenar e gerenciar todos os seus dados para que você possa utilizar em seus projetos com dados.

Databricks

Uma plataforma aberta destinada para armazenar e gerenciar todos os seus dados para todas as suas cargas de dados.

É dividida em :

NOTEBOOKS: Permite que os analistas e cientistas de dados construam seus scripts nas linguagens Python, SQL, R, Scala, fazendo consulta aos dados.

ANALYTICS: Use a poderosa interface SQL Analytics para consultar e visualizar dados e preparar dashboards.

DELTA LAKE: Permite que sejam combinados os diversos tipos de dados inseridos no Databricks, pra que ajustes, tratamentos nos dados possam ser executados.

MACHINE LEARNING: É possível desenvolver modelos matemáticos e estatísticos para a geração de informação para seus negócios.

SPARK

Faz todo o gerenciamento e distribuição do processamento executado no Databricks.

Apache Spark

Apache Spark é o principal **mecanismo** de análise unificado para Big Data e aprendizado de máquina que existe no mundo, sendo utilizado pelas grandes corporações. Explorando nas suas execuções o uso de memória e outras otimizações. Anteriormente as empresas utilizavam o Hadoop.

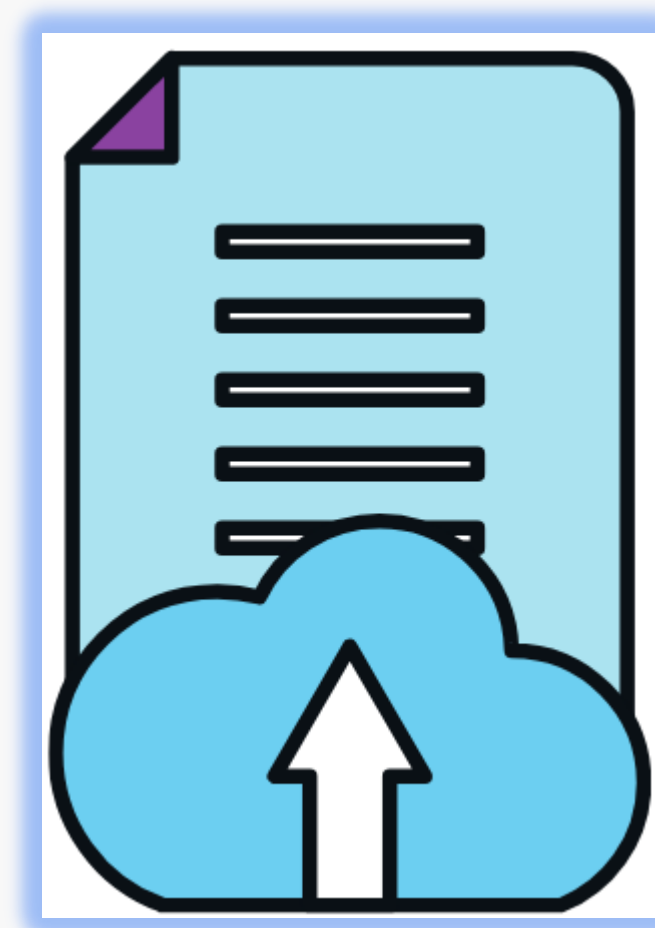


DBFS

Onde os dados são armazenados e gerenciados.

DBFS

Databricks File System (DBFS) é um sistema de arquivos distribuído montado em um espaço de trabalho Databricks e disponível em clusters Databricks.



Cluster no Databricks

É onde os recursos para operacionalização são criados. Precisa criar um Cluster para que o Databricks seja operacional.

Cluster

É um conjunto de recursos e configurações em que você cria os seus projetos, dentro dos chamados notebooks, é possível dentro de um cluster executar cargas de trabalho e analisar seus dados.

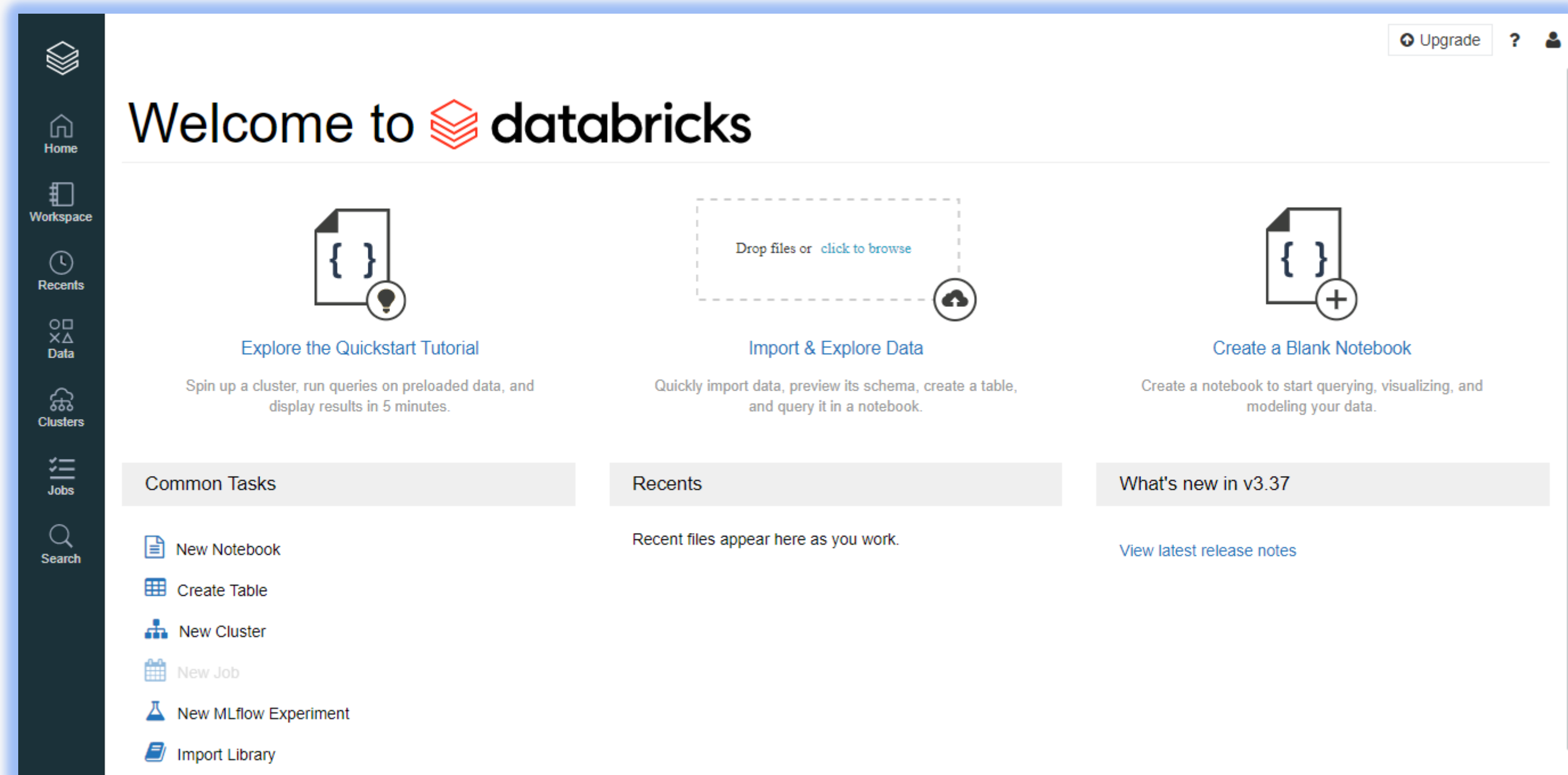


Workspace

É local de trabalho dentro do Databricks.

Workspace

Um espaço de trabalho Databricks é um ambiente para acessar todos os seus ativos Databricks. O espaço de trabalho organiza objetos, e fornece acesso a dados e recursos computacionais, como o clusters.



Workspace

É local de trabalho dentro do Databricks.

Workspace

Um espaço de trabalho Databricks é um ambiente para acessar todos os seus ativos Databricks. O espaço de trabalho organiza objetos, e fornece acesso a dados e recursos computacionais, como o clusters.

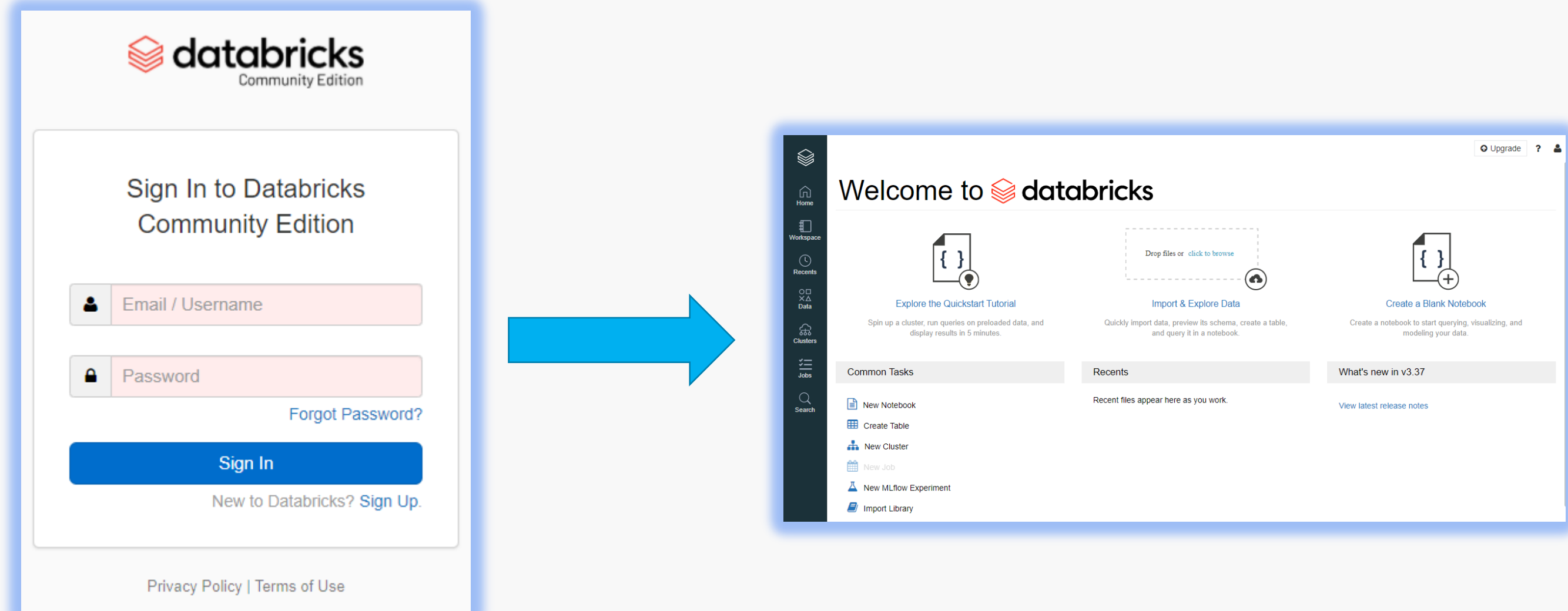


Criar sua conta

Vamos trabalhar na versão
Community gratuita.

Criando sua conta na Community

Você deverá criar um acesso gratuito na Community



- Os usuários terão acesso a clusters de 15GB, um gerenciador de clusters e o ambiente de notebook para protótipos de aplicações simples.
- O cluster com os dados fica disponível por 2 horas.
- Limitado a 3 usuários colaborativos.

Começando a trabalhar

Vamos criar o cluster para iniciarmos
nosso trabalho.

Criando o cluster

Primeiro precisaremos criar uma área para carregar os dados e gerar os
nossos notebooks para análise.

Clusters

All-Purpose ClustersJob Clusters

+ Create Cluster



Create Cluster

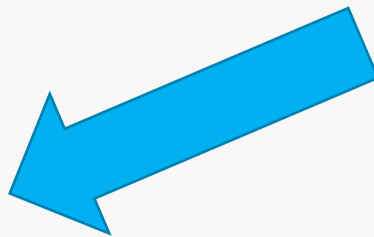
New ClusterCancelCreate Cluster0 Workers: 0.0 GB Memory, 0 Cores, 0 DBU
1 Driver: 15.3 GB Memory, 2 Cores, 1 DBU ⓘ

Cluster Name
Cursodatabricks

Databricks Runtime Version ⓘ
Runtime: 7.5 (Scala 2.12, Spark 3.0.1) | v

New This Runtime version supports only Python 3.

Instance
Free 15GB Memory: As a Community Edition user, your cluster will automatically terminate after an idle period of two hours.
For more configuration options, please upgrade your Databricks subscription.



Clusters

All-Purpose ClustersJob Clusters

+ Create Cluster

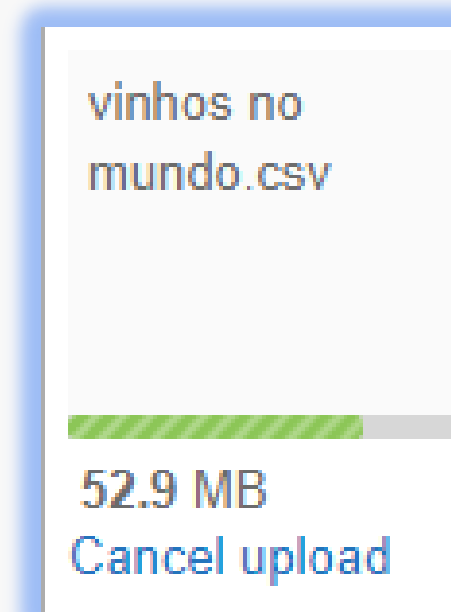
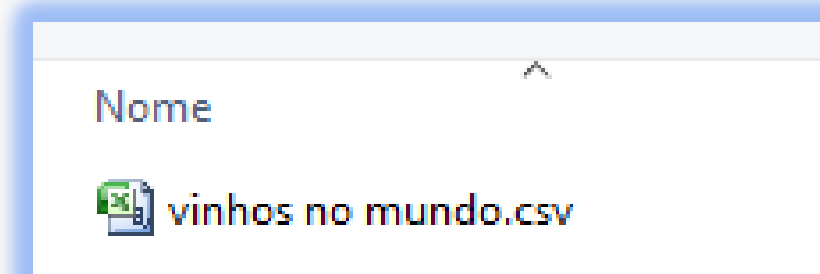
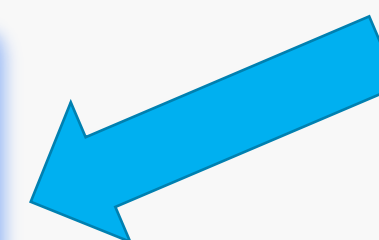
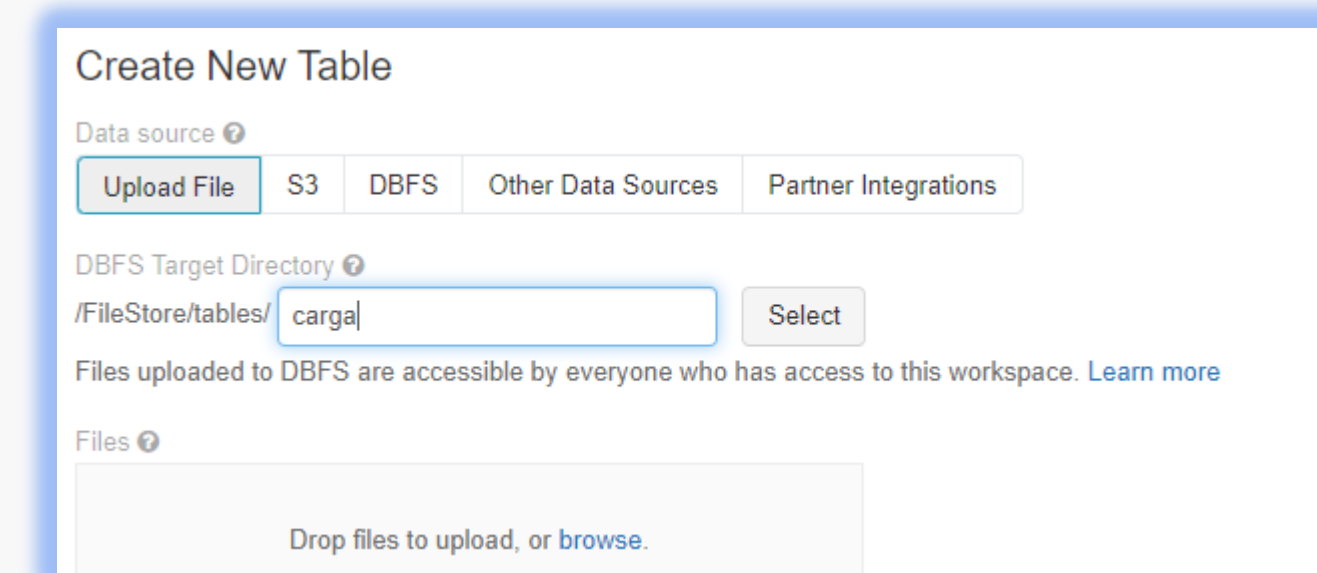
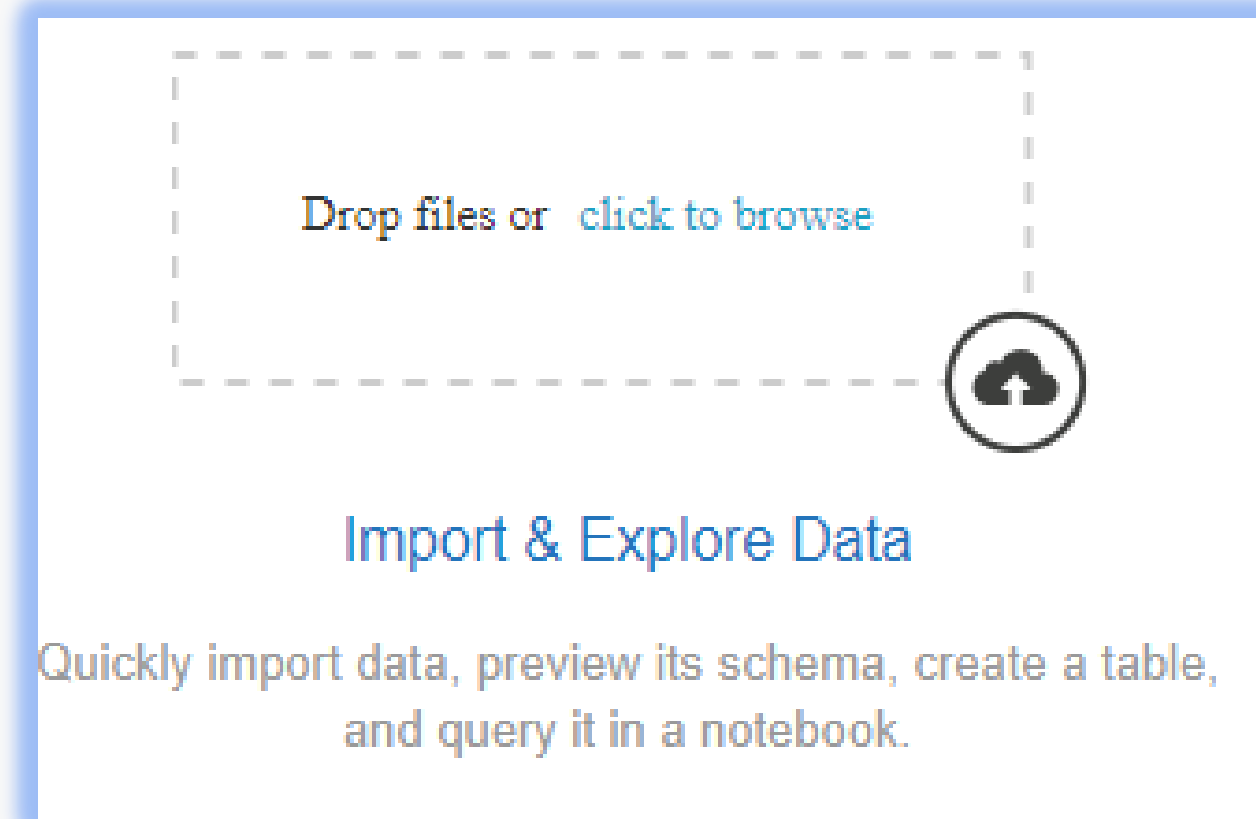
Name ▼	State ▼	Nodes ▼	Runtime ▼
● Cursodatabricks (clone)	Running	1 (0 spot)	7.5 (includes Apache Spark 3.0.1, S...

Começando a trabalhar

Vamos carregar os primeiros dados.

Carregando dados

Vamos carregar 129 mil registros em csv do arquivo **vinhos no mundo.csv**



Começando a trabalhar

Vamos carregar os primeiros dados.

Carregando dados

Vamos carregar 129 mil registros em csv do arquivo **vinhos no mundo.csv**

✓ File uploaded to `/FileStore/tables/carga/vinhos_no_mundo.csv`

Create Table with UI

Select a Cluster to Preview the Table

Choose a cluster with which you will read and preview the data.

Cluster ?

Cursodatabricks (clone)

● Cursodatabricks (clone)

15.25 GB | 2 Cores | DBR 7.5 | Spark 3.0.1 | Scala 2.12

Preview Table

Começando a trabalhar

Vamos carregar os primeiros dados.

Carregando dados

Vamos carregar 129 mil registros em csv do arquivo **vinhos no mundo.csv**. Criando a **tabela vinho**.

Specify Table Attributes

Specify the Table Name, Database and Schema to add this to the data UI for other users to access

Table Name

vinhos

Create in Database

default

File Type

CSV

Column Delimiter

,

☒ First row is header

☐ Infer schema

☐ Multi-line

Create Table

Create Table in Notebook

Table Preview

registro	pais	descricao	designacao	ponto
STRING	STRING	STRING	STRING	STRING
0	Italy	The palate isn't overly expressive, offering unripened apple, citrus	Vulkà Bianco	87
1	Portugal	Firm tannins are filled out with juicy red berry fruits and freshened with acidity. It's already	Avidagos	87
2	US	Some green pineapple pokes through, with crisp acidity	null	87
3	US	aromas. The palate is a bit more opulent, with notes of honey-drizzled guava and mango giving	Reserve Late Harvest	87
4	US	earthy, herbal characteristics. Nonetheless, if you think of it as a	Vintner's Reserve Wild Child Block	87
5	Spain	horseradish. In the mouth, this is fairly full bodied, with tomatoey	Ars In Vitro	87

Table: vinho

vinho

Refresh

Cursodatabricks (clone)

Schema:

	col_name	data_type	comment
1	registro	string	null
2	pais	string	null
3	descricao	string	null
4	designacao	string	null
5	ponto	string	null
6	preco	string	null
7	provincia	string	null
8	regiao	string	null

Showing all 14 rows.

Sample Data:

	registro	pais	descricao	designacao
1	0	Italy	Aromas include tropical fruit, broom, brimstone and dried herb. The palate isn't overly expressive, offering unripened apple, citrus and dried sage alongside brisk acidity.	Vulkà Bianco
2	1	Portugal	This is ripe and fruity, a wine that is smooth while still structured. Firm tannins are filled out with juicy red berry fruits and freshened with acidity. It's already drinkable, although it will certainly be better from 2016.	Avidagos

Começando criar um notebook de trabalho

Local onde analisaremos os dados.

Criando um notebook

Local que leremos os dados e faremos análises nos dados, podem serem gerados gráficos. Os notebook podem ser em R, Python, Scala e SQL.

Create Notebook

Name

vinho

Default Language

SQL

| v

Cluster

Cursodatabricks

| v

Cancel

Create

vinho (SQL)

Cursodatabricks

Cmd 1

Lista todos os vinhos

```
1 select
2 *
3 from
4 vinho
```

(1) Spark Jobs

	registro	pais	descricao	designacao
1	0	Italy	Aromas include tropical fruit, broom, brimstone and dried herb. The palate isn't overly expressive, offering unripened apple, citrus and dried sage alongside brisk acidity.	Vulkà Bianco
2	1	Portugal	This is ripe and fruity, a wine that is smooth while still structured. Firm tannins are filled out with juicy red berry fruits and freshened with acidity. It's already drinkable, although it will certainly be better from 2016.	Avidagos
3	2	US	Tart and snappy, the flavors of lime flesh and rind dominate. Some green pineapple pokes through, with crisp acidity underscoring the flavors. The wine was all stainless-steel fermented.	null
4	3	US	Pineapple rind, lemon pith and orange blossom start off the aromas. The palate is a bit more opulent, with notes of honey-drizzled guava and mango giving way to a slightly astringent, semidry finish.	Reserve Late
	4	US	Much like the regular bottling from 2012 this comes across as rather rough and tannic, with rustic, earthy, herbal characteristics.	Vintner's Res

Showing the first 1000 rows.

Table

Bar

Download

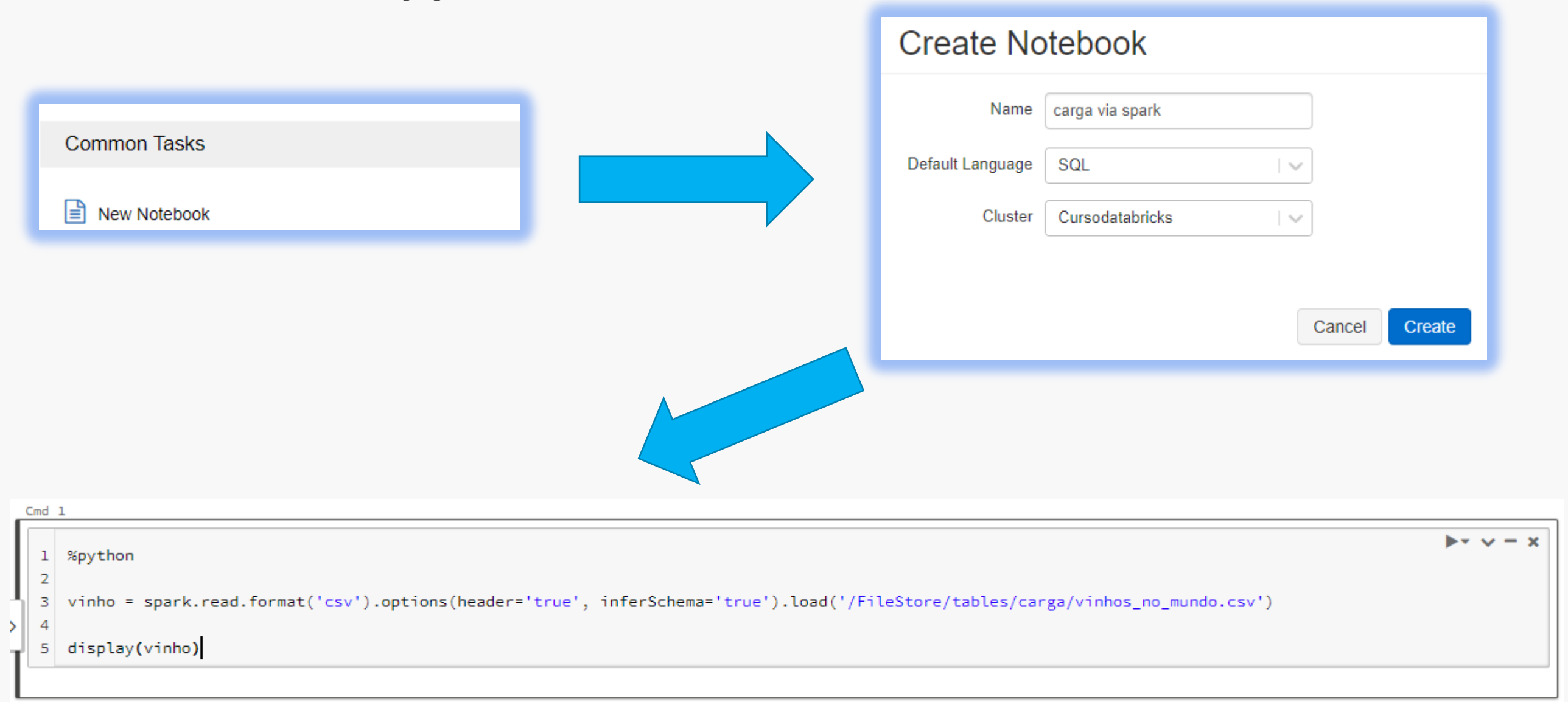
Command took 5.67 seconds -- by grimaldo_lopes@hotmail.com at 08/02/2021 14:42:48 on Cursodatabricks

Começando a trabalhar

Vamos carregar os dados de uma forma diferente.

Carregando dados

Vamos carregar + 10 mil registros em csv do arquivo **clientes cartao.csv**. Só que guardando no dataframe do spark. **Diretamente nos notebooks, trabalhando com python, Scala e SQL.**



%python

```
clientes = spark.read.format('csv').options(header='true',  
inferSchema='true',  
delimiter=';').load('/FileStore/tables/carga/clientes_cartao.csv'  
)
```

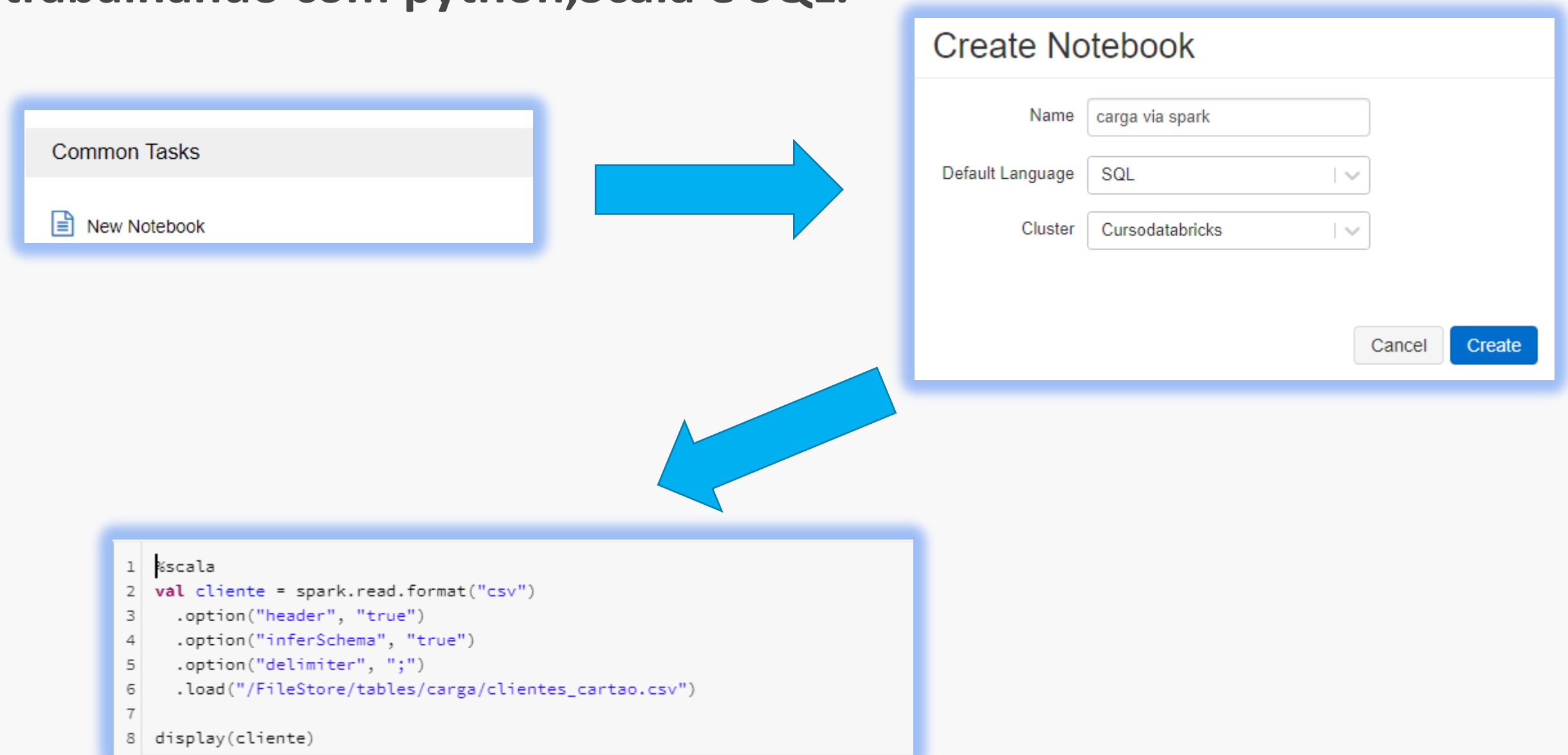
```
display(clientes)
```

Começando a trabalhar

Vamos carregar os dados de uma forma diferente.

Carregando dados

Vamos carregar + 10 mil registros em csv do arquivo **clientes cartao.csv**. Só que carregando direto do spark. **Diretamente nos notebooks, trabalhando com python, Scala e SQL.**



%scala

```
val cliente = spark.read.format("csv")
  .option("header", "true")
  .option("inferSchema", "true")
  .option("delimiter", ";")
  .load("/FileStore/tables/carga/clientes_cartao.csv")
```

```
display(cliente)
```


Começando a trabalhar

Vamos carregar os dados de uma forma diferente.

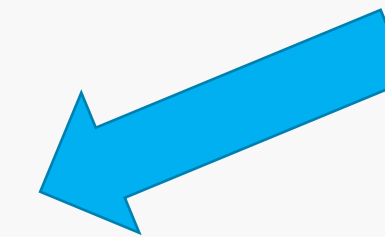
Carregando dados

Vamos carregar + 10 mil registros em csv do arquivo **clientes cartao.csv**. Só que carregando direto do spark. **Diretamente nos notebooks, trabalhando com python, Scala e SQL.**

```
Cmd 3
1 %scala
2 cliente.createOrReplaceTempView("dados_cliente")
3
4 |
```



```
1 %sql
2 select * from dados_cliente
```



```
%scala
cliente.createOrReplaceTempView("dados_cliente")
```



```
%sql
select * from dados_cliente
```

Começando a trabalhar

Preparando o primeiro gráfico.

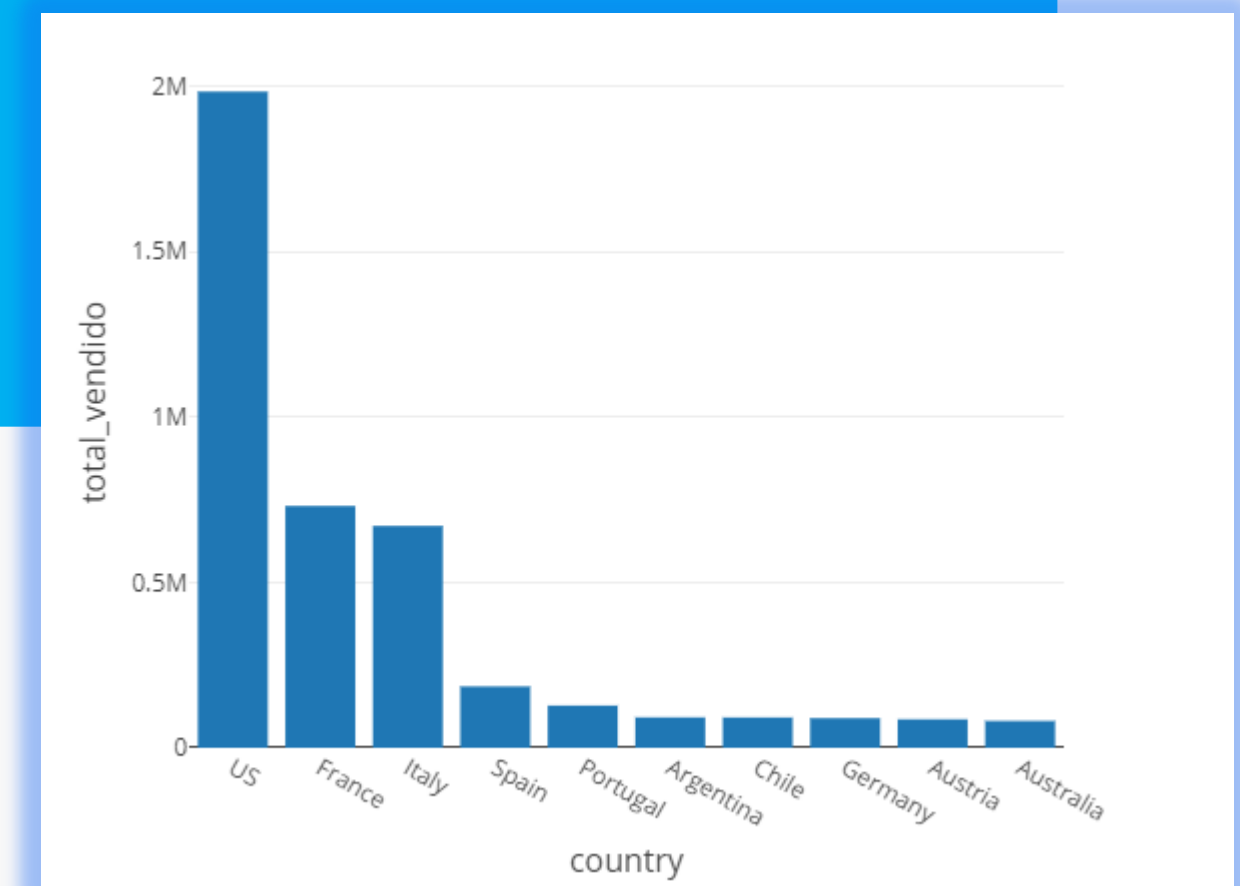
Primeiro gráfico

Vamos executar uma query cujo os resultados vamos representar graficamente.

```
1 %sql
2 select country, sum(price) as total_vendido from dados_vinho
3 where price > 0
4 group by country
5 order by total_vendido desc
6 limit 10
7
```

	country ▲	total_vendido ▲
1	US	1984106
2	France	731224
3	Italy	670882
4	Spain	185520
5	Portugal	127814
6	Argentina	92060
7	Chile	91793
8	Germany	89586

```
%sql
select pais, sum(preco) as total_vendido from vinho
where preco > 0
group by pais
order by total_vendido desc
limit 10
```



Começando a trabalhar

Preparando o segundo gráfico.

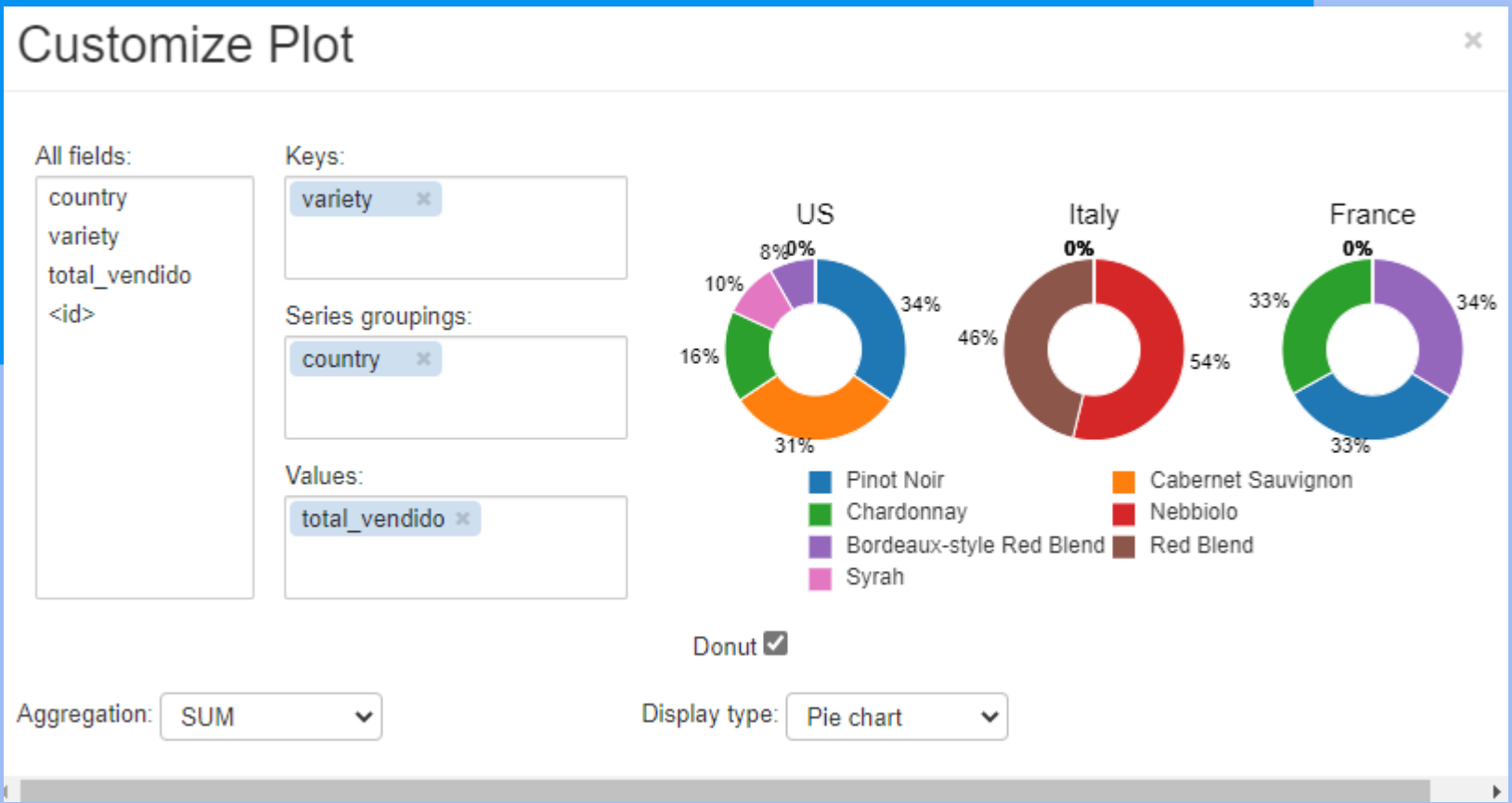
Segundo gráfico

Vamos executar uma query cujo os resultados vamos representar graficamente.

```
1 %sql
2 select country, sum(price) as total_vendido from dados_vinho
3 where price > 0
4 group by country
5 order by total_vendido desc
6 limit 10
7
```

	country	variety	total_vendido
1	US	Pinot Noir	439530
2	US	Cabernet Sauvignon	398345
3	US	Chardonnay	207482
4	Italy	Nebbiolo	150432
5	France	Bordeaux-style Red Blend	129764
6	Italy	Red Blend	129494
7	France	Pinot Noir	128934
8	France	Chardonnay	127724
9	US	Syrah	125775
10	US	Bordeaux-style Red Blend	105443

```
%sql
select pais, variante, sum(preco) as total_vendido from vinho
where preco > 0
group by pais, variante
order by total_vendido desc
limit 10
```



Começando a trabalhar

Preparando o terceiro gráfico.

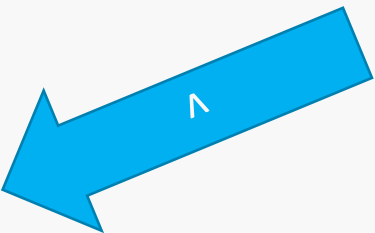
Terceiro gráfico

Vamos executar uma query cujo os resultados vamos representar graficamente.

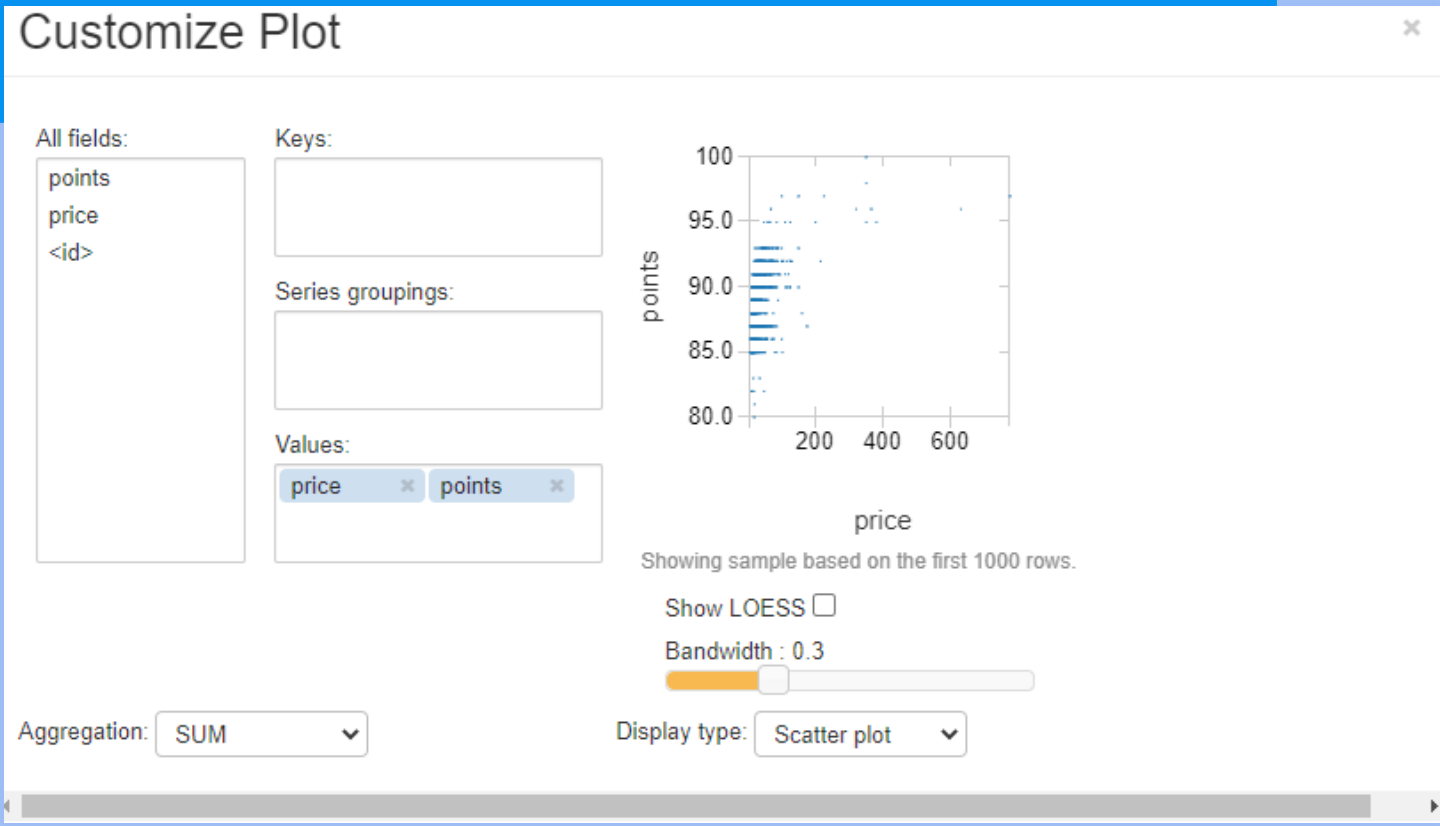
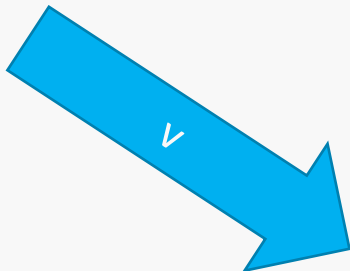
```
1 %sql
2 select points, price from dados_vinho
3
```



	points	price
1	87	null
2	87	15.0
3	87	14.0
4	87	13.0
5	87	65.0
6	87	15.0
7	87	16.0
8	87	24.0
9	87	12.0
10	87	27.0
11	87	19.0



```
%sql
select pontos, preco from vinho
```



Veja como criar um stream dos arquivos json.

Veja como criar um stream dos arquivos json.

Carregando JSON

Executa Script

Lista de arquivos Json que estão armazenado no DBFS

```
1 %fs ls /FileStore/tables/financeiro/
```

Showing all 1 rows.

```
1 %fs head /FileStore/tables/financeiro/2016q1.json
```

[Truncated to first 65536 bytes]

```
{ "num.txt": "adsh\ttag\tversion\tcoreg\tddate\tqttrs\ttuom\tvvalue\tfootnote\n0000003545-16-0  
4407000.0000\t\n0000003545-16-000130\tAccountsPayableCurrent\ttus-gAAP/2014\t\tt20151231\t0  
\t\ttus-gAAP/2014\t\tt20150930\t0\t\tUSD\t291100000.0000\t\n00000004127-16-000044\tAccountsPayab  
00004187-16-000034\tAccountsPayableCurrent\ttus-gAAP/2014\t\tt20141231\t0\t\tUSD\t2622000.0000  
\t20151231\t0\t\tUSD\t2152000.0000\t\n00000006955-16-000051\tAccountsPayableCurrent\ttus-gAAP,  
\tAccountsPayableCurrent\ttus-gAAP/2014\t\tt20151130\t0\t\tUSD\t124383000.0000\t\n00000006955-1  
150831\t0\t\tUSD\t14700000.0000\t\n00000006955-16-000051\tAccountsPayableCurrent\ttus-gAAP/20  
6-000051\tAccountsPayableCurrent\ttus-gAAP/2014\tNonGuarantorSubsidiaries\t20150831\t0\t\tUS  
us-gAAP/2014\tNonGuarantorSubsidiaries\t20151130\t0\t\tUSD\t92317000.0000\t\n00000006955-16-0  
s\t20150831\t0\t\tUSD\t19213000.0000\t\n00000006955-16-000051\tAccountsPayableCurrent\ttus-gAP  
\n00000006955-16-000051\tAccountsPayableCurrent\ttus-gAAP/2014\tConsolidationEliminations\t2  
urrent\ttus-gAAP/2014\tConsolidationEliminations\t20151130\t0\t\tUSD\t0.0000\t\n00000008947-16
```

VISUALIZAÇÃO

1. É um formato de arquivo de coluna que fornece otimizações para acelerar consultas
2. Tradicionalmente o armazenamento de dados em banco de dados é em linhas, no Parquet armazena os dados de forma colunar
3. Parquet foi criado para suportar compressão
4. O arquivo é dividido em dados e metadados
5. Está presente no Spark, Hive, Impala no ecossistema Hadoop em geral

Pontos Principais Arquivo Parquet

Formato	Espaço utilizado	Tempo de execução	Escaneado
CSV	2 TB	472 seg	2.3 TB
Parquet	260 GB	13,56 seg	5.02 GB

Criando arquivos parquet

Veja como criar um arquivo parquet e realizar sua leitura.

Veremos agora demonstrar a potencialidade do spark na execução de scripts para gravação e leitura de arquivos parquet.

DATAFRAME

```
#criando um dataframe com dados fixos
dados = [("Grimaldo ", "Oliveira", "Brasileira", "Professor", "M", 3000),
        ("Ana ", "Santos", "Portuguesa", "Atriz", "F", 4000),
        ("Roberto", "Carlos", "Brasileira", "Analista", "M", 4000),
        ("Maria ", "Santanna", "Italiana", "Dentista", "F", 6000),
        ("Jeane", "Andrade", "Portuguesa", "Medica", "F", 7000)]
colunas=["Primeiro_Nome", "Ultimo_nome", "Nacionalidade", "Trabalho", "Genero", "Salario"]
datafparquet=spark.createDataFrame(data,colunas)
datafparquet.show()
```

Executa Script

Gravando o arquivo Parquet

```
1 #criando o arquivo parquet
2 datafparquet.write.parquet("/FileStore/tables/parquet/pessoal.parquet")
```

Arquivo gerado

pessoal.parquet	<div><div><div></div><div>_SUCCESS</div></div><div><div></div><div>_committed_1372109968542957...</div></div><div><div></div><div>_committed_2705869040626374...</div></div><div><div></div><div>_committed_573820509918871798</div></div><div><div></div><div>_started_2705869040626374083</div></div><div><div></div><div>_started_573820509918871798</div></div><div><div></div><div>Nacionalidade=Brasileira</div></div><div><div></div><div>Nacionalidade=Italiana</div></div><div><div></div><div>Nacionalidade=Portuguesa</div></div></div>
-----------------	--

Criando arquivos parquet

Veja como criar um arquivo parquet lendo um arquivo CSV.

Vamos ler um arquivo CSV e vamos guardar no arquivo parquet.

DATAFRAME lendo de um arquivo CSV

```
1 %python
2 #Leitura de arquivo CSV
3 dataframesp= spark.read.format("csv").option("header", "true").load("/FileStore/tables/arquivo/Datafiniti_Hotel_Reviews_Jun19.csv")
4 dataframesp.show()
```

Executa Script de gravação em parquet

```
1 #criando o arquivo parquet
2 dataframesp.write.parquet("/FileStore/tables/parquet/csvparquet3.parquet")
```

Arquivo gerado

📁 csvparquet3.parquet

Sistema de Arquivos

Comandos de manipulação.

Comandos

Acessar o sistema de arquivos usando comandos `%fs` (sistema de **arquivos**) ou `%sh` (shell de comando). Existe a biblioteca **dbutils.fs** Python. São formas diferentes de obter os mesmos resultados.

DBFS (Sistema de arquivos Databricks)

`%fs ls temporario`

`Dbutils.fs.ls("/temporario/")`

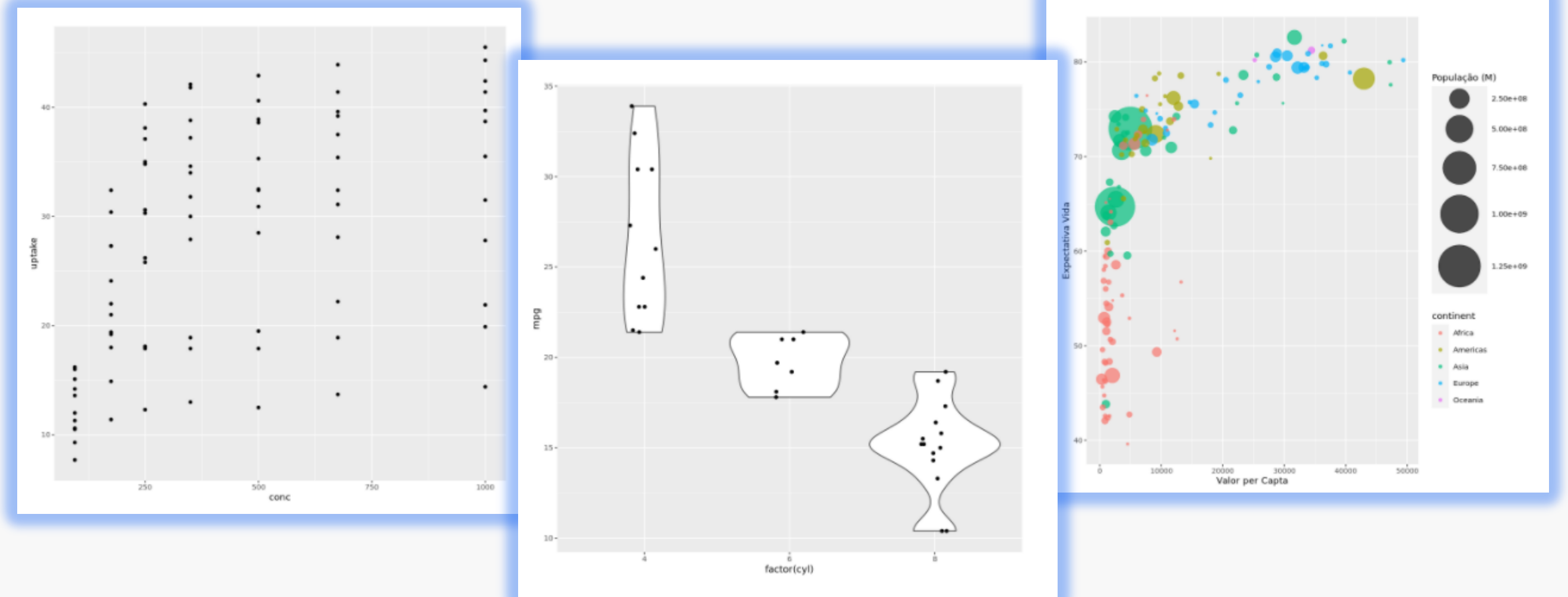
`%sh ls /dbfs/temporario`

Construção de dashboard

Visão dos gráficos gerados.

Dashboard

É possível construir um dashboard com gráficos que são gerados. Vamos gerar gráficos em R para aprender sobre esta construção.



Leia os dados do Databricks no Power BI com uma simples configuração.

Vamos conectar o Databricks diretamente no Power BI, ferramenta de visualização de dados.

The diagram illustrates the process of obtaining data from Databricks. It starts with a configuration page on the left, which has tabs for 'Instances', 'Spark', 'JDBC/ODBC' (selected), and 'Per'. The configuration fields are: 'Server Hostname' with value 'community.cloud.databricks.com', 'Port' with value '443', 'Protocol' with value 'HTTPS', and 'HTTP Path' with value 'sql/protocolv1/o/1215337267419013/'. A blue arrow points from this page to a box titled 'Obter Dados'. This box contains a search bar with 'databricks' and a dropdown menu with 'Tudo' and 'Azure'. To the right of the search bar is a section labeled 'Tudo' with a Databricks logo and the text 'Azure Databricks'. A second blue arrow points down from the 'Obter Dados' box to a box titled 'Azure Databricks'. This box contains two input fields: 'Server Hostname' with a help icon and an empty field, and 'HTTP Path' with a help icon and an example value 'sql/protocolv1/o/1814582234607533/7508-187377-agent704'.

```
graph LR; A["Instances | Spark | JDBC/ODBC | Per"] --> B["Obter Dados"]; B --> C["Azure Databricks"];
```

Instances | **Spark** | **JDBC/ODBC** | Per

Server Hostname

community.cloud.databricks.com

Port

443

Protocol

HTTPS

HTTP Path

sql/protocolv1/o/1215337267419013/

Obter Dados

databricks X

Tudo

Azure

Tudo

Azure Databricks

Azure Databricks

Server Hostname ⓘ

HTTP Path ⓘ

Exemplo: sql/protocolv1/o/1814582234607533/7508-187377-agent704

PRÁTICA

ENVIE AO PROFESSOR

PREPARE UM ESTUDO

Carregando qualquer arquivo do site **kaggle.com** e crie um cluster, scripts, visualizações, dashboards e analise seus dados.

Muito boa sorte e conte comigo!

```
1 %sql
2 select country, sum(price) as total_vendido from dados_vinho
3 where price > 0
4 group by country
5 order by total_vendido desc
6 limit 10
7
```

	country	variety	total_vendido
1	US	Pinot Noir	439530
2	US	Cabernet Sauvignon	398345
3	US	Chardonnay	207482
4	Italy	Nebbiolo	150432
5	France	Bordeaux-style Red Blend	129764
6	Italy	Red Blend	129494
7	France	Pinot Noir	128934
8	France	Chardonnay	127724
9	US	Syrah	125775
10	US	Bordeaux-style Red Blend	105443

Customize Plot

