

Tutorial *Hive*

Instalação do *Hive* e *Kick-off*

Baseado no trabalho de Nikolay Dimolarov e Romain Rigaux

(<https://towardsdatascience.com/making-big-moves-in-big-data-with-hadoop-hive-parquet-hue-and-docker-320a52ca175>)

Preparar o *Docker Hive*

git clone <https://github.com/tech4242/docker-hadoop-hive-parquet.git>

cd .\docker-hadoop-hive-parquet\

docker-compose up -d

```
PS C:\Users\msi\docker-hadoop-hive-parquet> docker-compose up -d
[+] Running 78/15
  ✓ datanode 8 layers [#####] 0B/0B Pulled 67.4s
  ✓ resourcemanager 3 layers [###] 0B/0B Pulled 67.3s
  ✓ hive-metastore 12 layers [#####] 0B/0B Pulled 70.1s
  ✓ namenode 10 layers [#####] 0B/0B Pulled 67.4s
  ✓ hue 15 layers [#####] 0B/0B Pulled 117.5s
  ✓ hive-metastore-postgresql 14 layers [#####] 0B/0B Pulled 44.0s
  ✓ hive-server Pulled 70.1s
  ✓ huedb 8 layers [#####] 0B/0B Pulled 48.9s

[+] Building 0.0s (0/0)
[+] Running 9/9
  ✓ Network docker-hadoop-hive-parquet_default Created 0.0s
  ✓ Container docker-hadoop-hive-parquet-huedb-1 Started 0.3s
  ✓ Container docker-hadoop-hive-parquet-resource-manager-1 Started 0.3s
  ✓ Container docker-hadoop-hive-parquet-hive-metastore-1 Started 0.3s
  ✓ Container docker-hadoop-hive-parquet-namenode-1 Started... 0.3s
  ✓ Container docker-hadoop-hive-parquet-hive-metastore-postgresql-1 Started 0.3s
  ✓ Container docker-hadoop-hive-parquet-datanode-1 Started... 0.3s
  ✓ Container docker-hadoop-hive-parquet-hive-server-1 St... 0.3s
  ✓ Container docker-hadoop-hive-parquet-hue-1 Started 0.1s
```

Nota: caso obtenha uma mensagem de erro no arranque, executar o comando:

```
net stop winnat
```

Importar dados de ficheiro Parquet

1. Seleccionar um *dataset* com informação interessante (e.g., [Kaggle](https://www.kaggle.com/datasets/samyakb/student-stress-factors/))
 - a. Para efeitos deste tutorial utilizar <https://www.kaggle.com/datasets/samyakb/student-stress-factors/>
2. Descarregar o ficheiro “Student Stress Factors.csv” e renomear para “Student_Stress_Factors.csv”
3. Colocar esse ficheiro junto ao script “parquet_converter.py”
4. Modificar a linha 8 colocando o caminho correcto para o ficheiro
5. Modificar a linha 9 colocando o caminho pretendido para o ficheiro de destino
6. Instalar os seguintes módulos *Python*:

```
pip install pandas
pip install pyarrow
```

7. Executar o script “parquet_converter.py” de modo a obter o ficheiro “Student_Stress_Factors.parquet”

Criar modelo de dados para suportar a integração

1. Instalar os seguintes módulos *Python*:

```
pip install parquet-tools
```

2. Executar o comando na directoria onde está o ficheiro “Student_Stress_Factors.parquet”

```
parquet-tools inspect .\Student_Stress_Factors.parquet
```

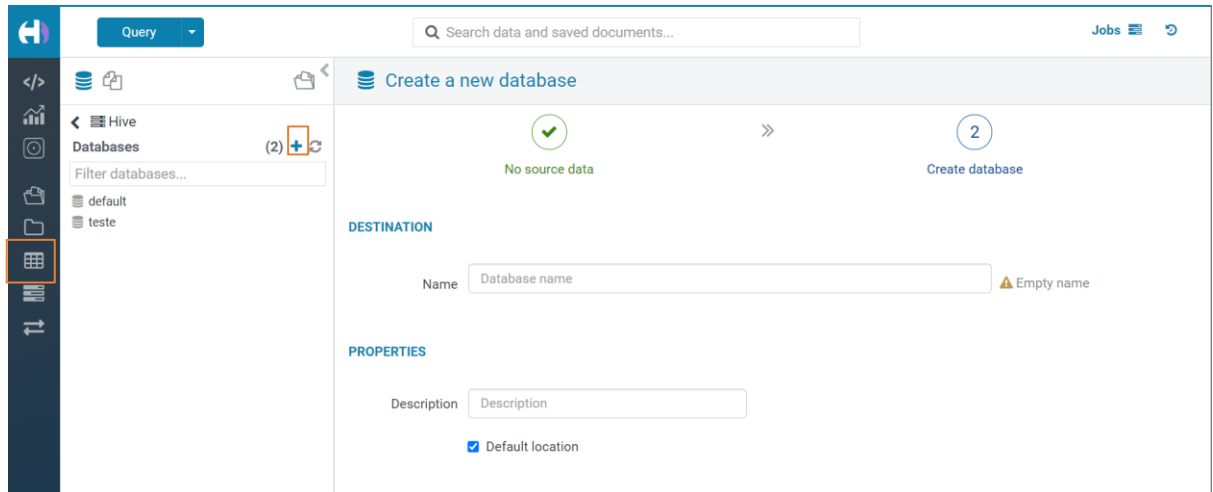
```
PS C:\Users\msi\docker-hadoop-hive-parquet> parquet-tools inspect .\Student_Stress_Factors.parquet

##### file meta data #####
created_by: parquet-cpp-arrow version 14.0.1
num_columns: 7
num_rows: 53
num_row_groups: 1
format_version: 2.6
serialized_size: 6306

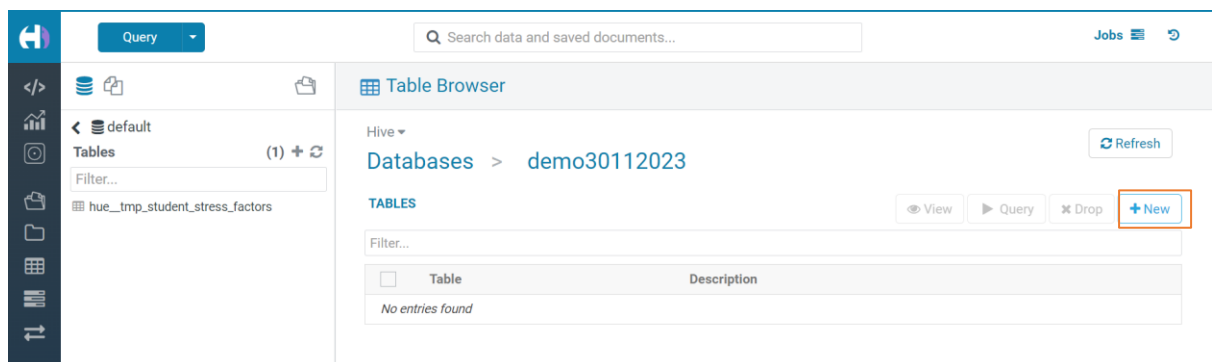
##### Columns #####
Timestamp
Kindly Rate your Sleep Quality 😊
How many times a week do you suffer headaches 🤔?
How would you rate your academic performance 🧑 ?
how would you rate your study load?
How many times a week you practice extracurricular activities 🏀?
How would you rate your stress levels?

##### Column(Timestamp) #####
name: Timestamp
path: Timestamp
max_definition_level: 1
max_repetition_level: 0
physical_type: BYTE_ARRAY
logical_type: String
converted_type (legacy): UTF8
compression: UNCOMPRESSED (space_saved: 0%)
```

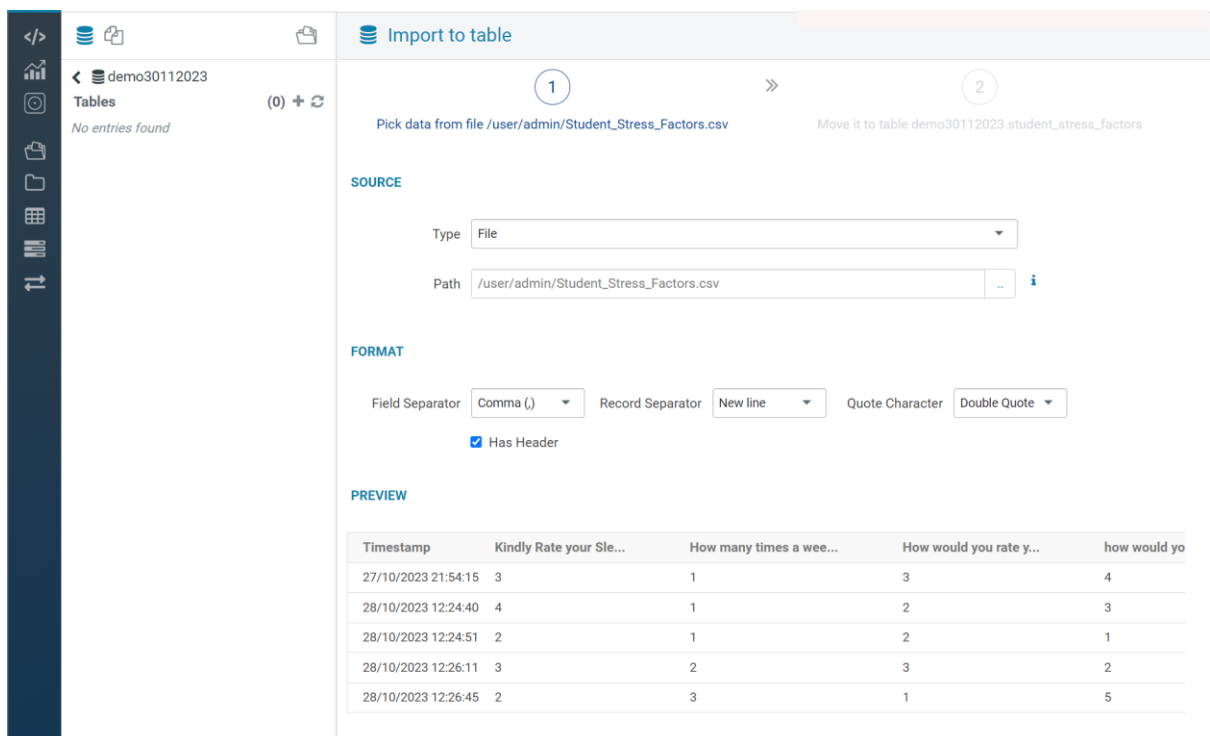
3. Aceder ao hue: <http://localhost:8888>
 - a. Definir a *password* no primeiro acesso
4. Criar uma base de dados



5. Criar uma nova tabela “demo30112023”

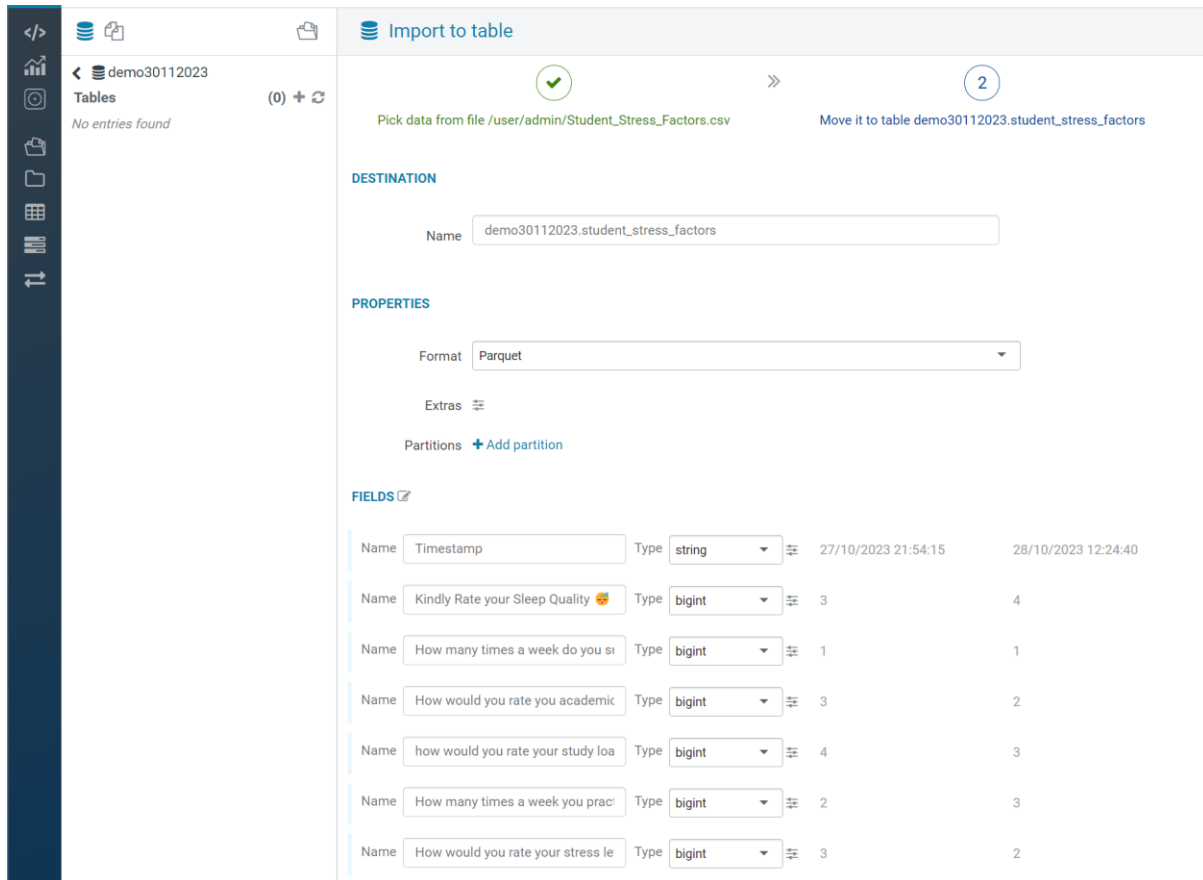


6. Importar os dados do ficheiro “csv” após realizar o *upload*



7. Seleccionar o nome da tabela e o formato “csv” do ficheiro

8. Seleccionar o destino como “Parquet”



Import to table

Pick data from file /user/admin/Student_Stress_Factors.csv Move it to table demo30112023.student_stress_factors

DESTINATION

Name demo30112023.student_stress_factors

PROPERTIES

Format Parquet

Extras

Partitions + Add partition

FIELDS

Name	Type	27/10/2023 21:54:15	28/10/2023 12:24:40
Timestamp	string		
Kindly Rate your Sleep Quality 🤪	bigint	3	4
How many times a week do you si	bigint	1	1
How would you rate you academic	bigint	3	2
how would you rate your study loa	bigint	4	3
How many times a week you prac	bigint	2	3
How would you rate your stress le	bigint	3	2

9. Submit

10. Esta versão utiliza Hive-on-MR (Hive on MapReduce) que foi descontinuado. As alternativas são Spark, que será introduzido na próxima UC.

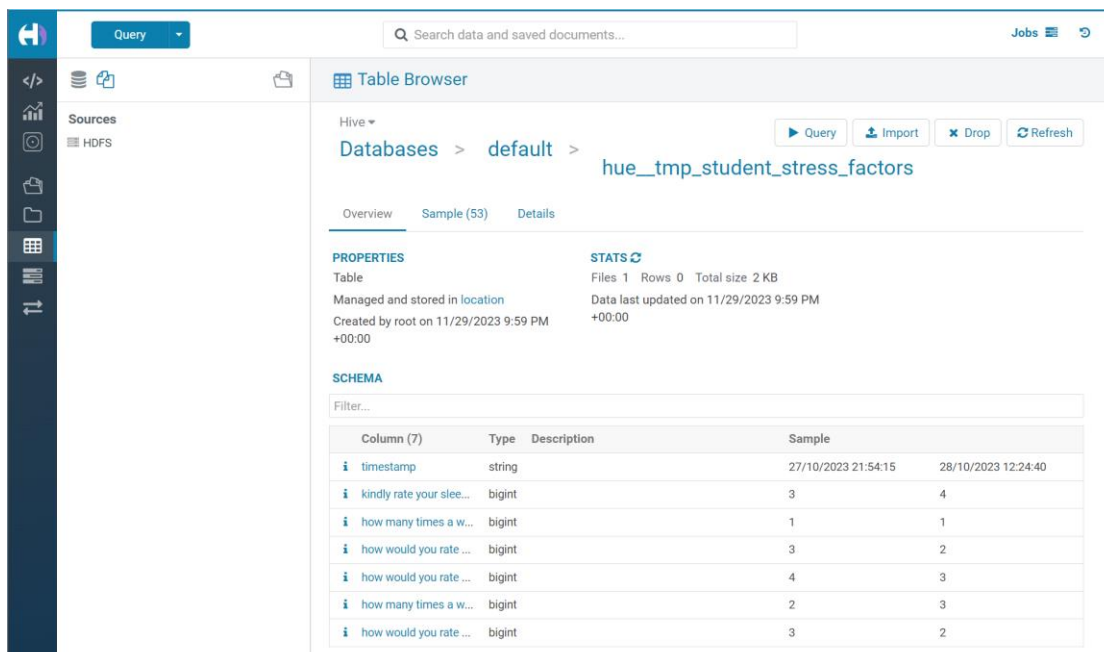


Table Browser

Hive ▾

Databases > default > hue_tmp_student_stress_factors

Query Import Drop Refresh

Overview Sample (53) Details

PROPERTIES

Table

Managed and stored in location

Created by root on 11/29/2023 9:59 PM +00:00

STATS

Files 1 Rows 0 Total size 2 KB

Data last updated on 11/29/2023 9:59 PM +00:00

SCHEMA

Filter...

Column (7)	Type	Description	Sample
i timestamp	string		27/10/2023 21:54:15 28/10/2023 12:24:40
i kindly rate your slee...	bigint		3 4
i how many times a w...	bigint		1 1
i how would you rate ...	bigint		3 2
i how would you rate ...	bigint		4 3
i how many times a w...	bigint		2 3
i how would you rate ...	bigint		3 2

11. Visualizar os dados através do *dashboard* e realizar algumas *queries* sobre os novos dados.

Dica: faça primeiro uma query e depois mude para o gráfico

Hive Execute and watch Add a description... 0.34s Database default Type text ?

```
SELECT * FROM "default"."hue_tmp_student_stress_factors" LIMIT 100;
```

Query History Saved Queries Query Builder Results (100+)

COLUMNS (7) Q

- ☒ hue_tmp_student_stress_factors.timestamp
- ☒ hue_tmp_student_stress_factors.kindly rate your sleep quality
- ☒ hue_tmp_student_stress_factors.how many times a week do you suffer headaches
- ☒ hue_tmp_student_stress_factors.how would you rate your academic performance
- ☒ hue_tmp_student_stress_factors.how would you rate your study load?
- ☒ hue_tmp_student_stress_factors.how many times a week you practice extracurricular activities
- ☒ hue_tmp_student_stress_factors.how would you rate your stress levels?

	hue_tmp_student_stress_factors.timestamp	hue_tmp_student_stress_factors.kindly rate your sleep quality	hue_tmp_student_stress_factors.how many times a week do you suffer headaches
1	27/10/2023 21:54:15	3	1
2	28/10/2023 12:24:40	4	1
3	28/10/2023 12:24:51	2	1
4	28/10/2023 12:26:11	3	2
5	28/10/2023 12:26:45	2	3
6	28/10/2023 12:31:02	3	1
7	28/10/2023 12:34:45	3	5
8	28/10/2023 12:35:43	4	3
9	28/10/2023 12:36:07	2	1
10	28/10/2023 12:36:20	1	2
11	28/10/2023 12:37:22	2	3
12	28/10/2023 12:38:40	3	1
13	28/10/2023 12:39:43	2	3
14	28/10/2023 12:40:50	4	1
15	28/10/2023 12:41:19	4	1

12. Podemos tirar algumas conclusões com os dados?

