Tutorial Big Data
CLEI 2019
Ciudad de Panamá, Panamá

Por: Edwin Montoya
emontoya@eafit.edu.co

Laboratorios

- https://github.com/edwinmontoya/tutorialbigdataclei2019.git
- **Se tendrá acceso a una serie de Recursos en Nube:**
  - **AWS EMR:**
    - http://emr1.emontoya.ml:8888 (hue)
    - http://emr2.emontoya.ml:8888 (si necesita)
    - http://emr3.emontoya.ml:8888 (si necesita)
      - User: admin Password: Clei2019*

    - http://emr1.emontoya.ml:8890 (zeppelin)
    - http://emr2.emontoya.ml:8890 (si necesita)
    - http://emr3.emontoya.ml:8890 (si necesita)
      - Sin autenticación

Laboratorios

- **Base de datos RDBMS – MySQL (Amazon RDS / MySQL**
  - **Host:** database-1.cj1yhistqein.us-east-2.rds.amazonaws.com
  - **Port: 3306**

Base de datos: "cursodb"

Tabla: "employee"

User: curso / curso

$ mysql –u curso –h database-1.cj1yhistqein.us-east-2.rds.amazonaws.com **–p**

Enter password: ****

mysql> use cursodb;

```
mysql> describe employee;
+--------+-------------+------+-----+---------+-------+
| Field  | Type        | Null | Key | Default | Extra |
+--------+-------------+------+-----+---------+-------+
| emp_id | int(11)     | NO   | PRI | NULL    |       |
| name   | varchar(45) | YES  |     | NULL    |       |
| salary | int(11)     | YES  |     | NULL    |       |
+--------+-------------+------+-----+---------+-------+
3 rows in set (0.00 sec)

mysql>
```

Laboratorios

- **Base de datos RDBMS – MySQL (Amazon RDS / MySQL**
  - **Host:** database-1.cj1yhistqein.us-east-2.rds.amazonaws.com
  - **Port: 3306**

Base de datos: "retail_db"

Tabla: <6 tablas>

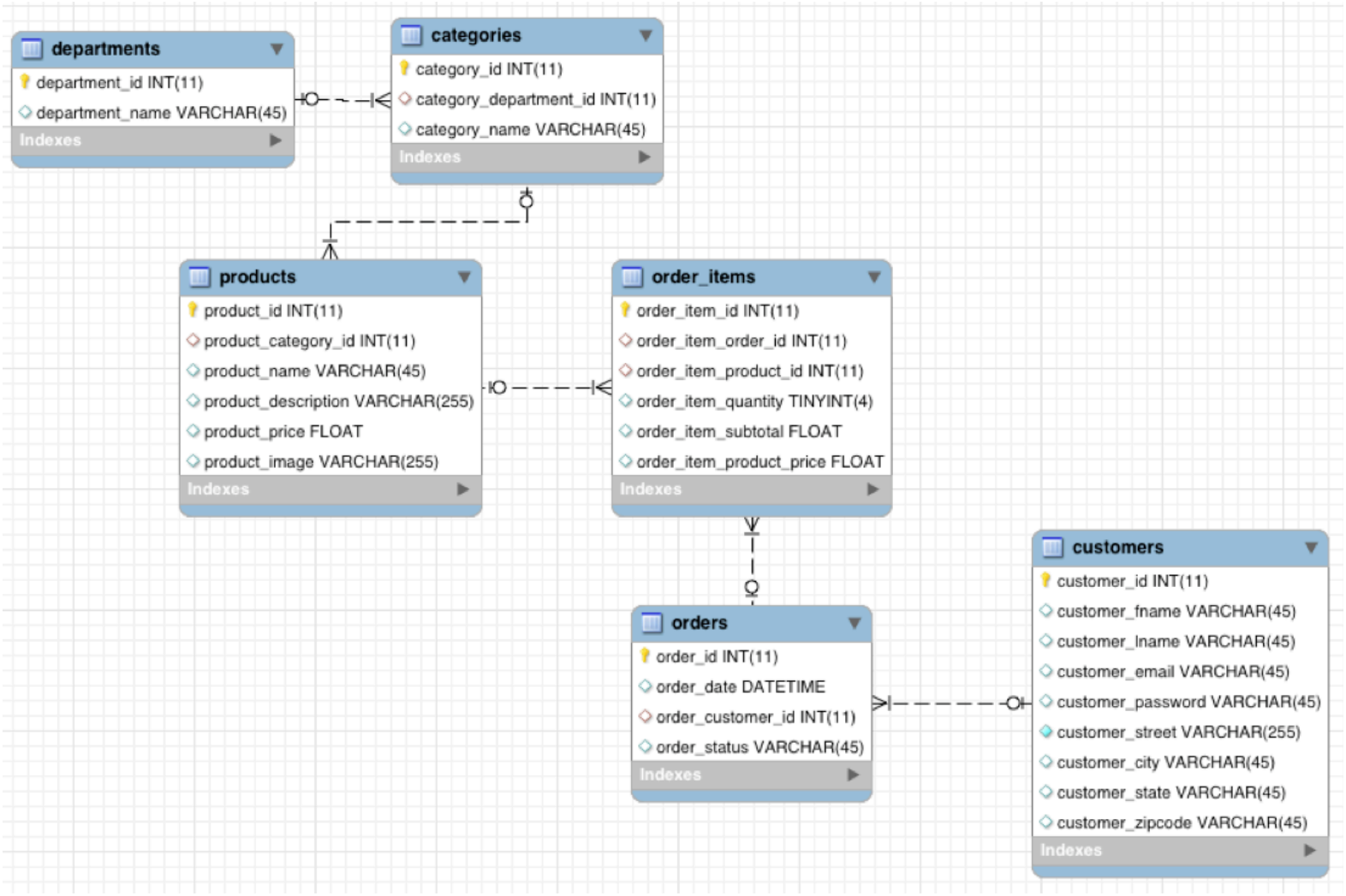
User: retail_dba / retail_dba

$ mysql –u retail_dba –h database-1.cj1yhistqein.us-east-2.rds.amazonaws.com **–p**

Enter password: ****

mysql> use retail_dba;

Retail_DB


departments
department_id INT(11)
department_name VARCHAR(45)
Indexes
categories
category_id INT(11)
category_department_id INT(11)
category_name VARCHAR(45)
Indexes
products
product_id INT(11)
product_category_id INT(11)
product_name VARCHAR(45)
product_description VARCHAR(255)
product_price FLOAT
product_image VARCHAR(255)
Indexes
order_items
order_item_id INT(11)
order_item_order_id INT(11)
order_item_product_id INT(11)
order_item_quantity TINYINT(4)
order_item_subtotal FLOAT
order_item_product_price FLOAT
Indexes
orders
order_id INT(11)
order_date DATETIME
order_customer_id INT(11)
order_status VARCHAR(45)
Indexes
customers
customer_id INT(11)
customer_fname VARCHAR(45)
customer_lname VARCHAR(45)
customer_email VARCHAR(45)
customer_password VARCHAR(45)
customer_street VARCHAR(255)
customer_city VARCHAR(45)
customer_state VARCHAR(45)
customer_zipcode VARCHAR(45)
Indexes

Laboratorios

- **Datasets**
  - Los datos de trabajo para el tutorial se encuentran en varias partes replicados, depende de donde los quiera o requiera acceder:
  - En el github:

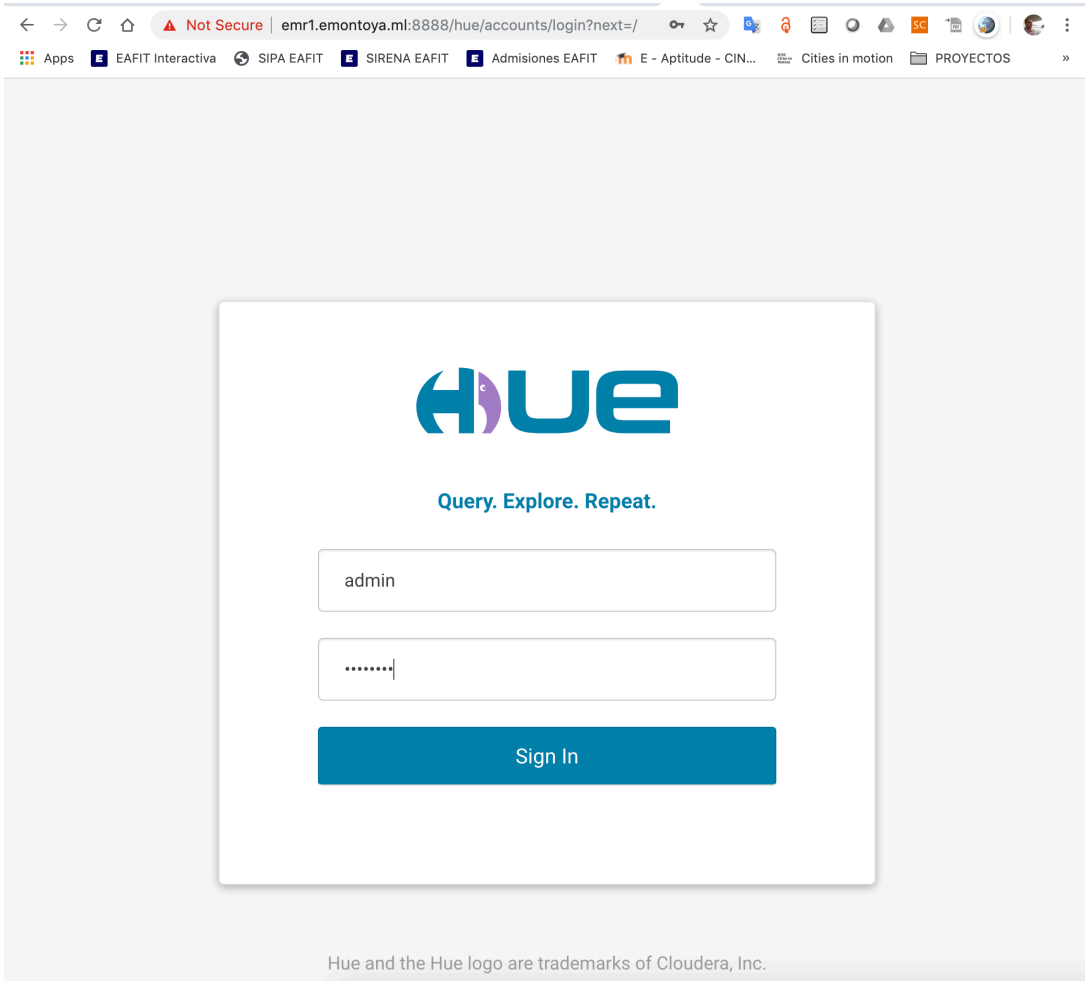
    https://github.com/edwinmontoya/tutorialbigdataclei2019.git
  - En Amazon S3 -> s3://emontoyapublic/datasets
  - DBname: <varias>
  - User: curso o retail_dba
  - Pass: *******

GESTION DE ARCHIVOS DESDE HUE

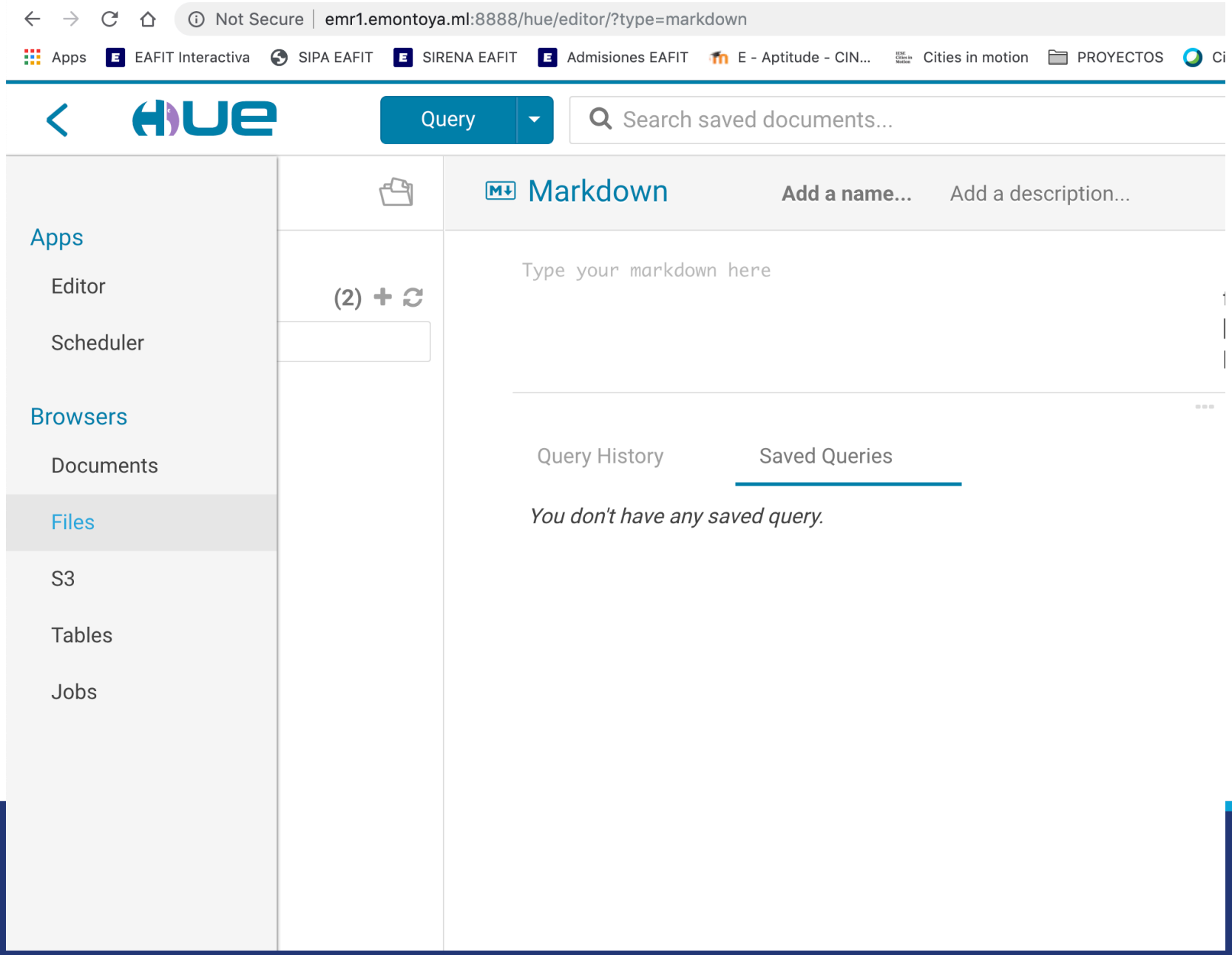
ACCESO WEB A HUE

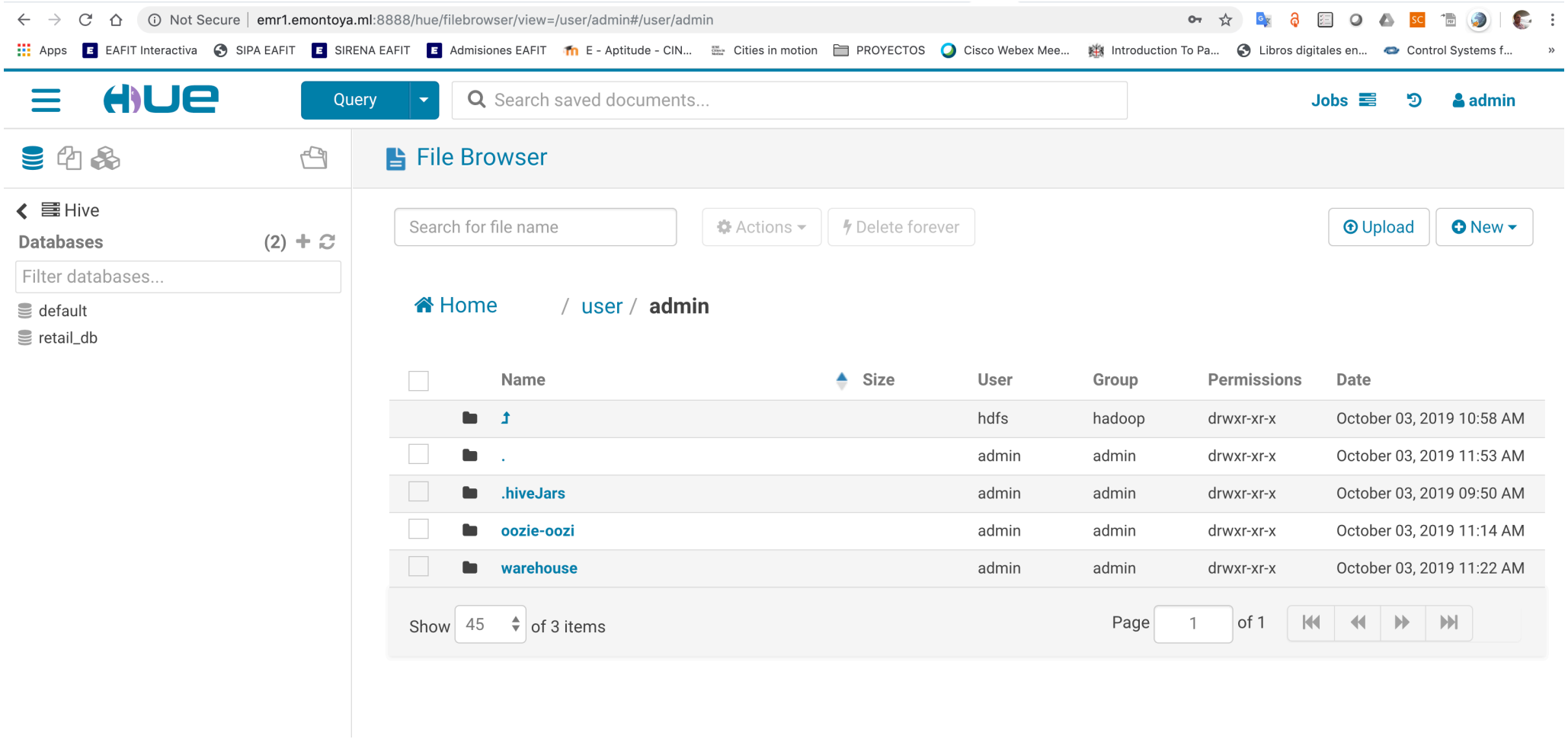
http://emr1.emontoya.ml:8888

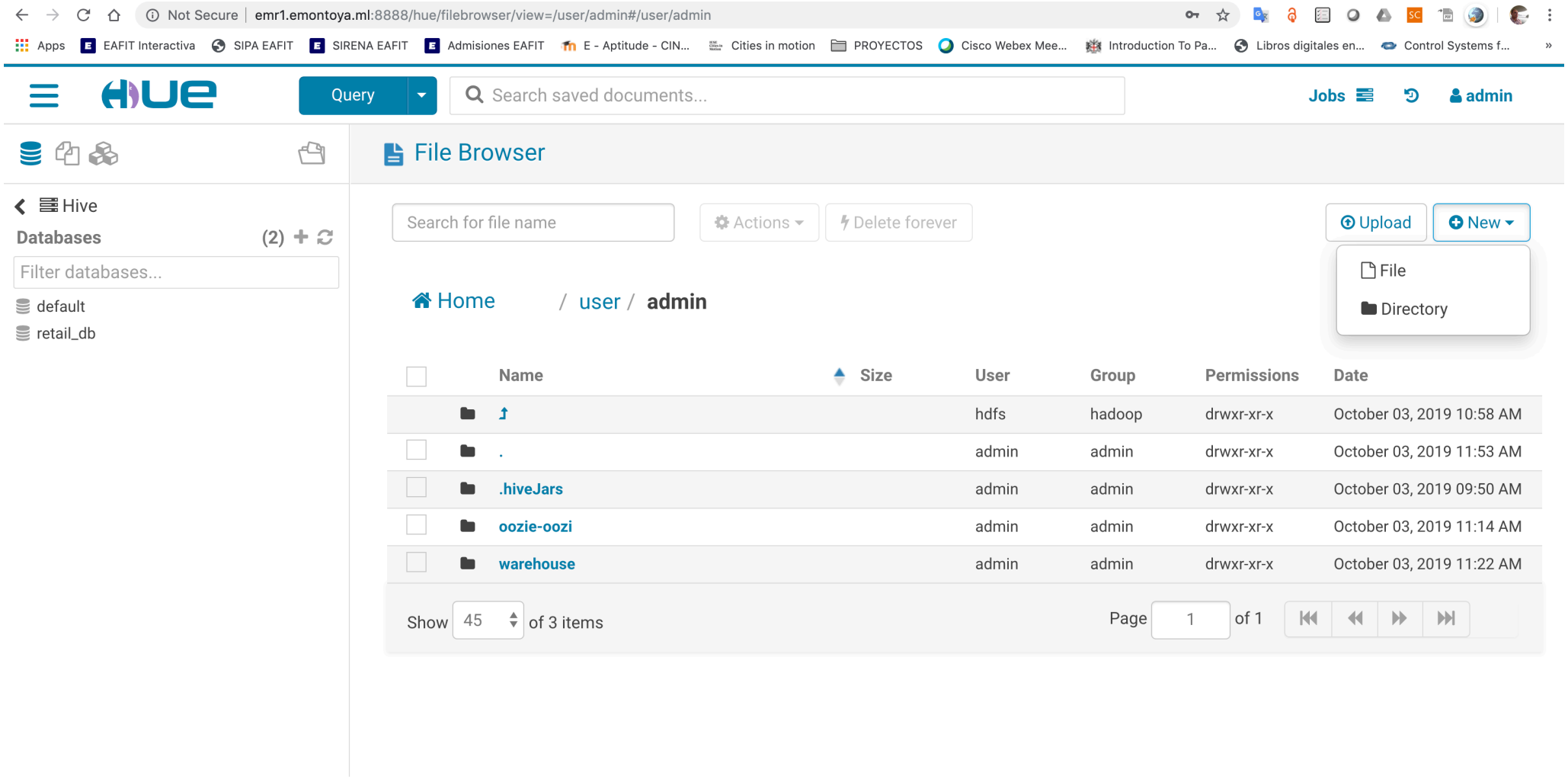
Login

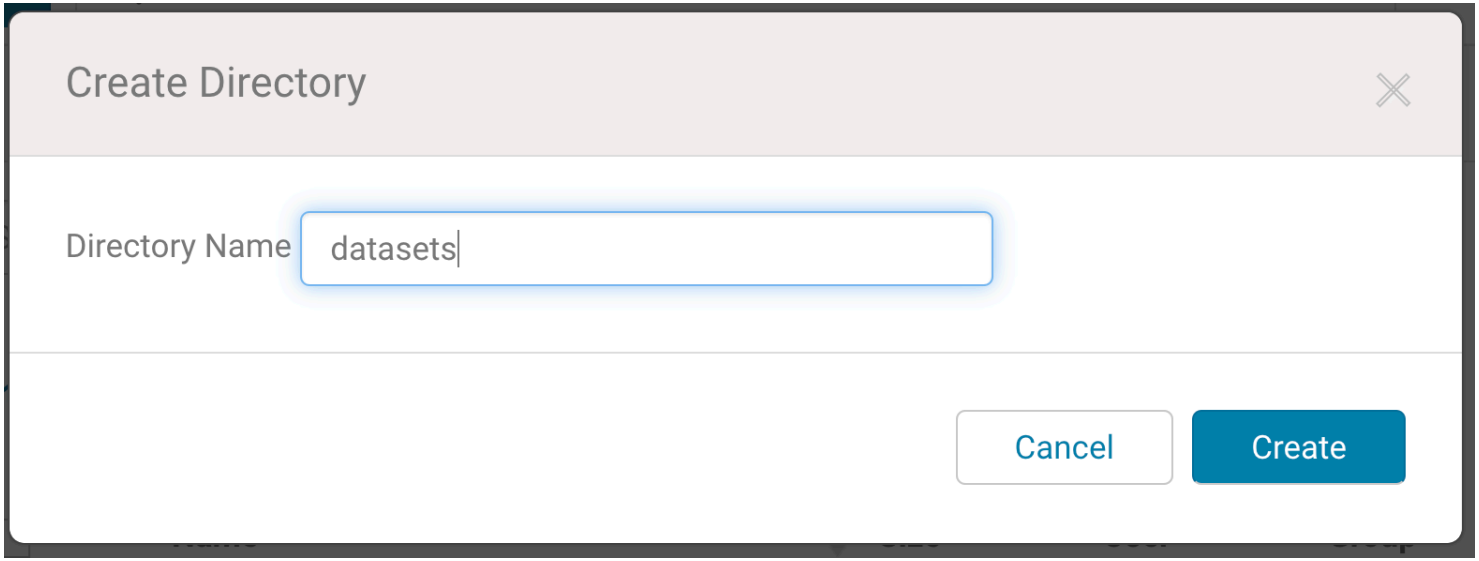

Not Secure | emr1.emontoya.ml:8888/hue/accounts/login?next=/
Apps
EAFIT Interactiva
SIPA EAFIT
SIRENA EAFIT
Admisiones EAFIT
E - Aptitude - CIN...
Cities in motion
PROYECTOS
HUE
Query. Explore. Repeat.
admin
Sign In
Hue and the Hue logo are trademarks of Cloudera, Inc.

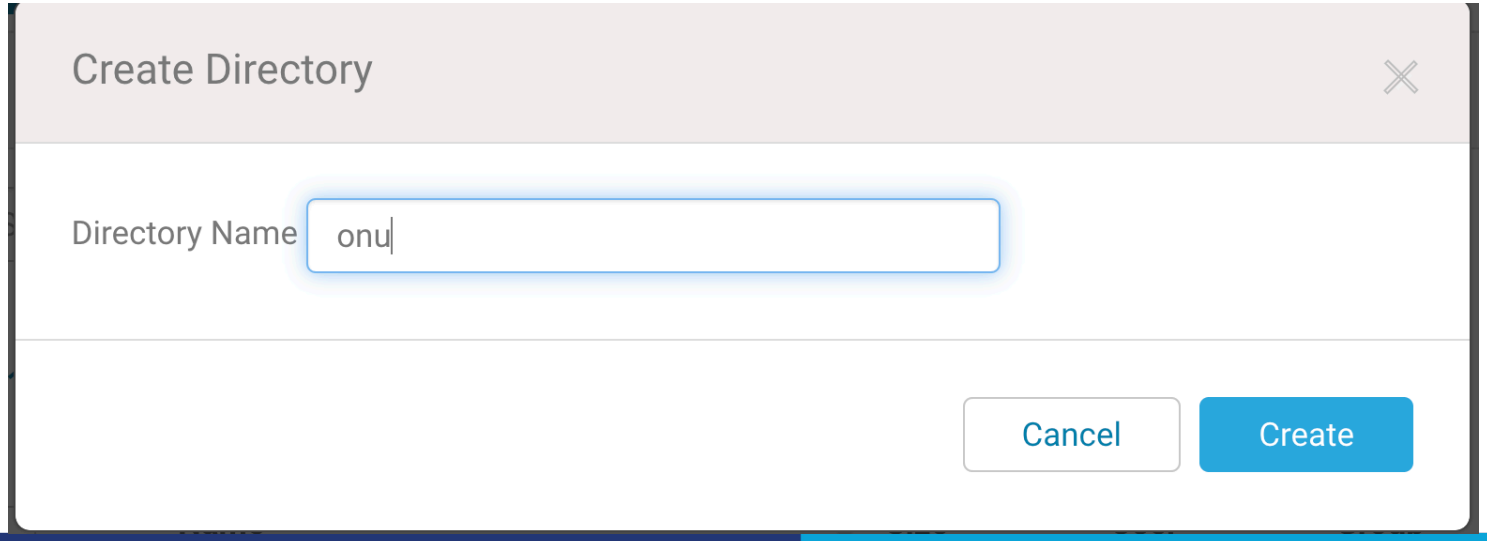
archivos

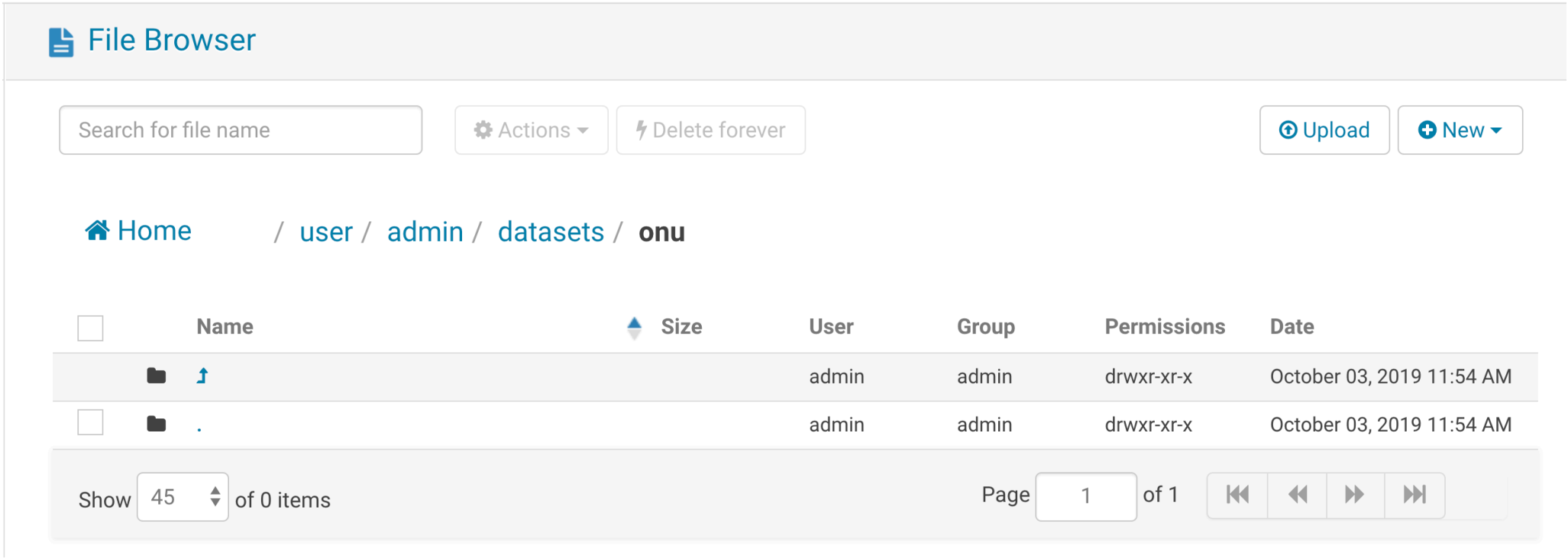

Not Secure | emr1.emontoya.ml:8888/hue/editor/?type=markdown
Apps
EAFIT Interactiva
SIPA EAFIT
SIRENA EAFIT
Admisiones EAFIT
E - Aptitude - CIN...
Cities in motion
PROYECTOS
HUE
Query
Search saved documents...
Markdown
Add a name...
Add a description...
Apps
Editor
Scheduler
Browsers
Documents
Files
S3
Tables
Jobs
(2)
Type your markdown here
Query History
Saved Queries
You don't have any saved query.

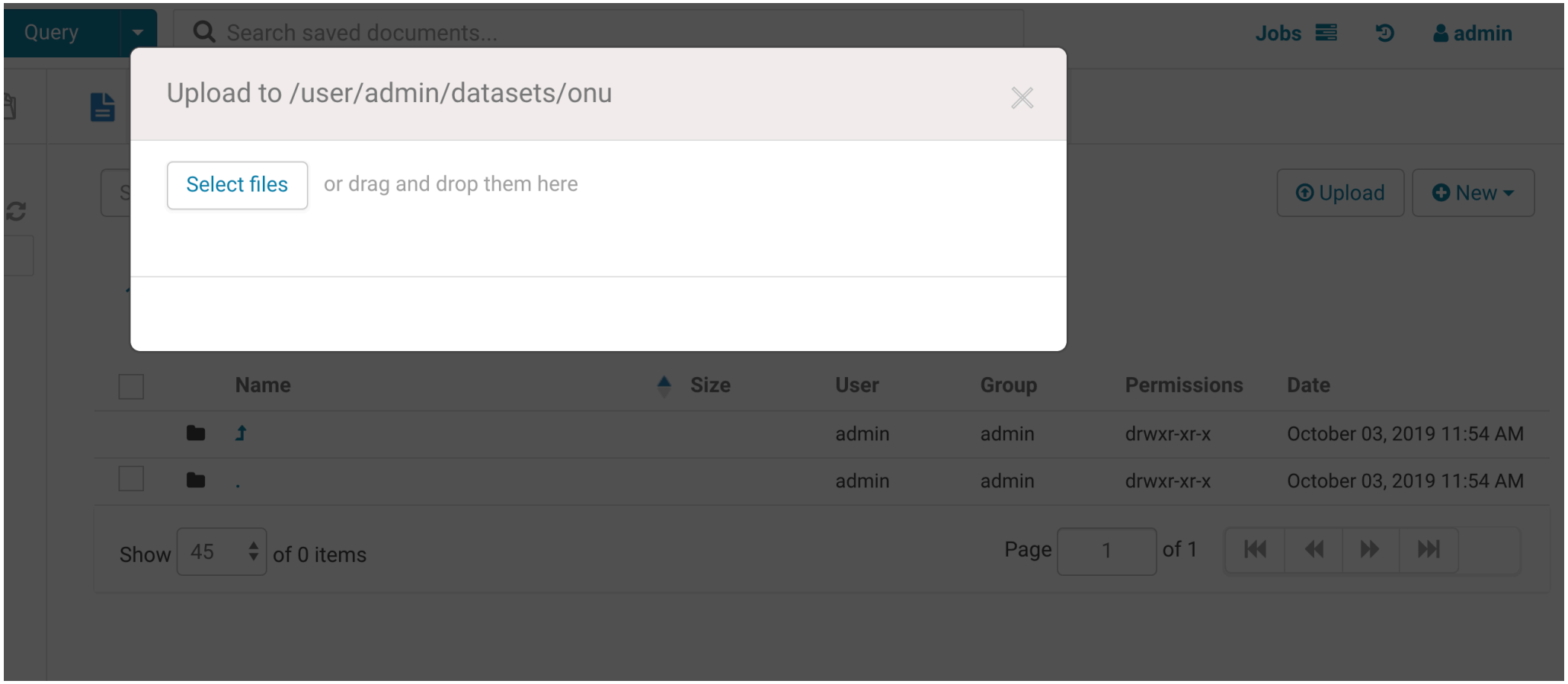

Not Secure
emr1.emontoya.ml:8888/hue/filebrowser/view=/user/admin#/user/admin
Apps
EAFIT Interactiva
SIPA EAFIT
SIRENA EAFIT
Admisiones EAFIT
E - Aptitude - CIN...
Cities in motion
PROYECTOS
Cisco Webex Mee...
Introduction To Pa...
Libros digitales en...
Control Systems f...
HUE
Query
Search saved documents...
Jobs
admin
File Browser
Hive
Databases (2)
Filter databases...
default
retail_db
Search for file name
Actions
Delete forever
Upload
New
Home / user / admin
Name
Size
User
Group
Permissions
Date
hdfs hadoop drwxr-xr-x October 03, 2019 10:58 AM
. admin admin drwxr-xr-x October 03, 2019 11:53 AM
.hiveJars admin admin drwxr-xr-x October 03, 2019 09:50 AM
oozie-oozi admin admin drwxr-xr-x October 03, 2019 11:14 AM
warehouse admin admin drwxr-xr-x October 03, 2019 11:22 AM
Show 45 of 3 items
Page 1 of 1

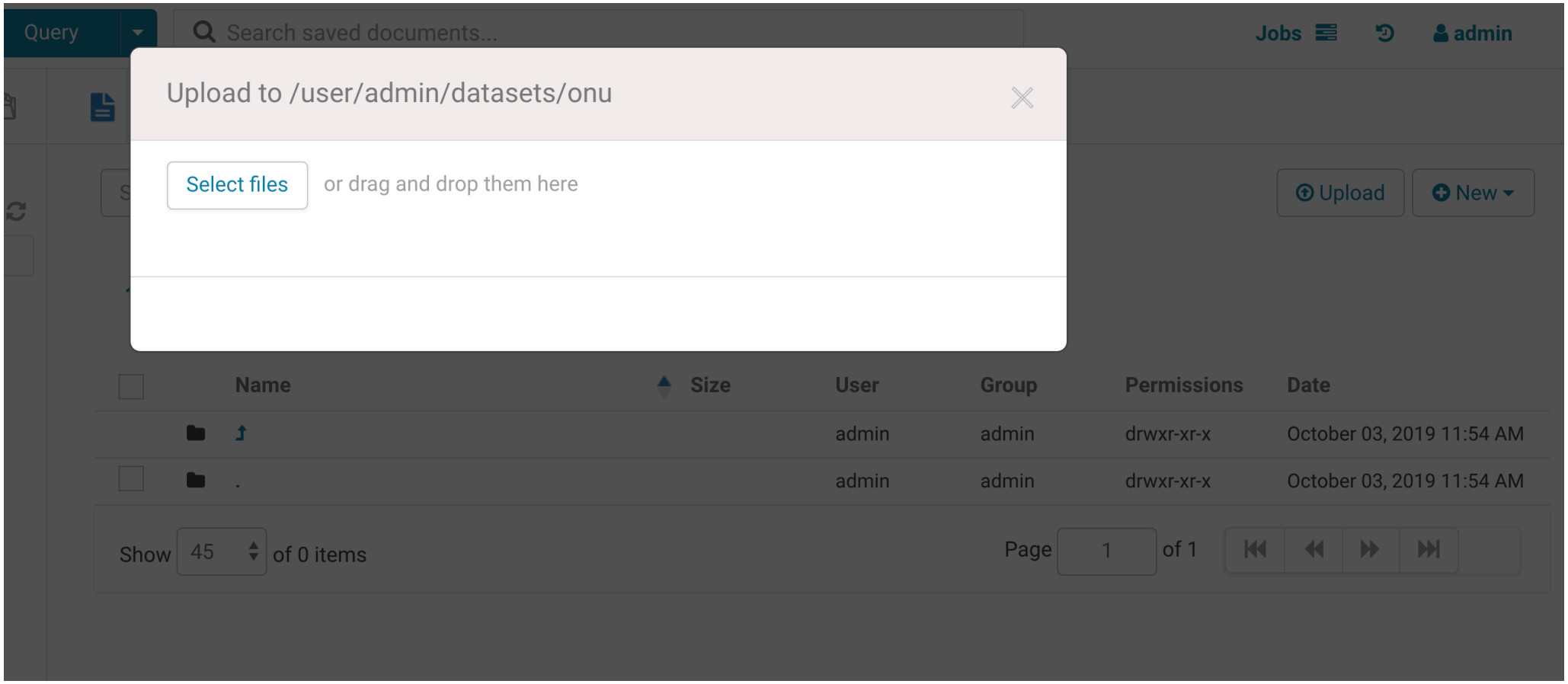

Not Secure
emr1.emontoya.ml:8888/hue/filebrowser/view=/user/admin#/user/admin
Apps
EAFIT Interactiva
SIPA EAFIT
SIRENA EAFIT
Admisiones EAFIT
E - Aptitude - CIN...
Cities in motion
PROYECTOS
Cisco Webex Mee...
Introduction To Pa...
Libros digitales en...
Control Systems f...
HUE
Query
Search saved documents...
Jobs
admin
File Browser
Hive
Databases (2)
Filter databases...
default
retail_db
Search for file name
Actions
Delete forever
Upload
New
File
Directory
Home / user / admin
Name
Size
User
Group
Permissions
Date
hdfs hadoop drwxr-xr-x October 03, 2019 10:58 AM
admin admin drwxr-xr-x October 03, 2019 11:53 AM
.hiveJars admin admin drwxr-xr-x October 03, 2019 09:50 AM
oozie-oozi admin admin drwxr-xr-x October 03, 2019 11:14 AM
warehouse admin admin drwxr-xr-x October 03, 2019 11:22 AM
Show 45 of 3 items
Page 1 of 1

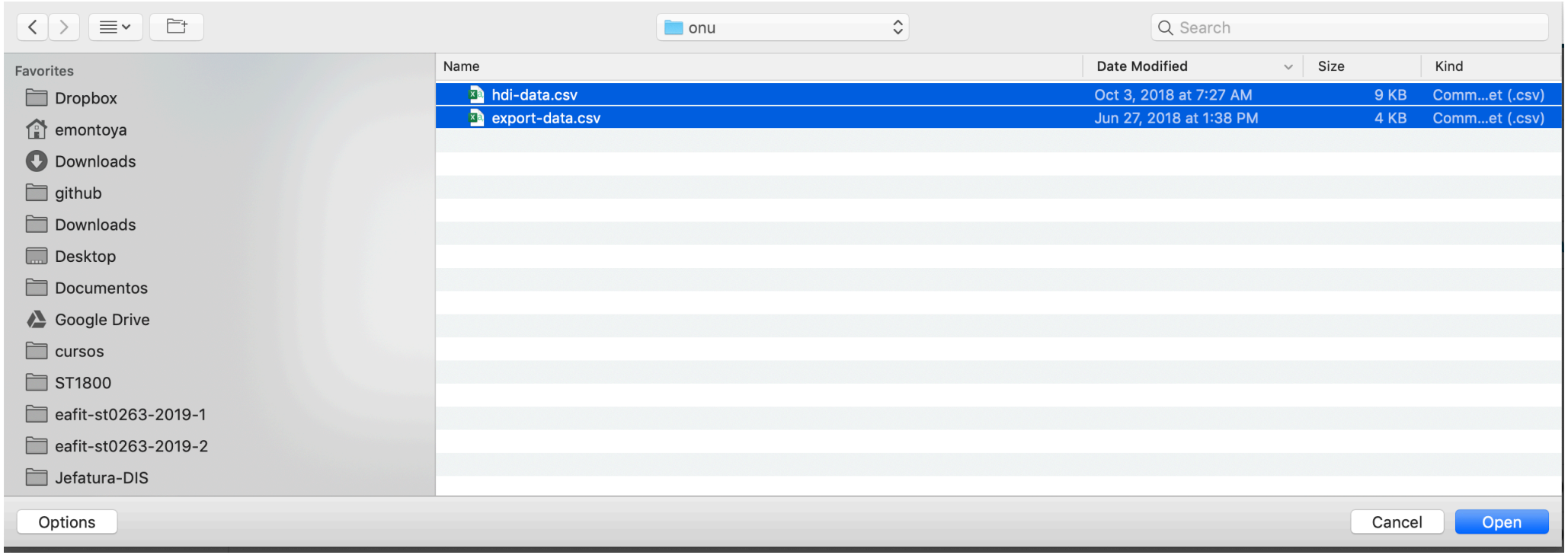

Create Directory
Directory Name
datasets
Cancel
Create

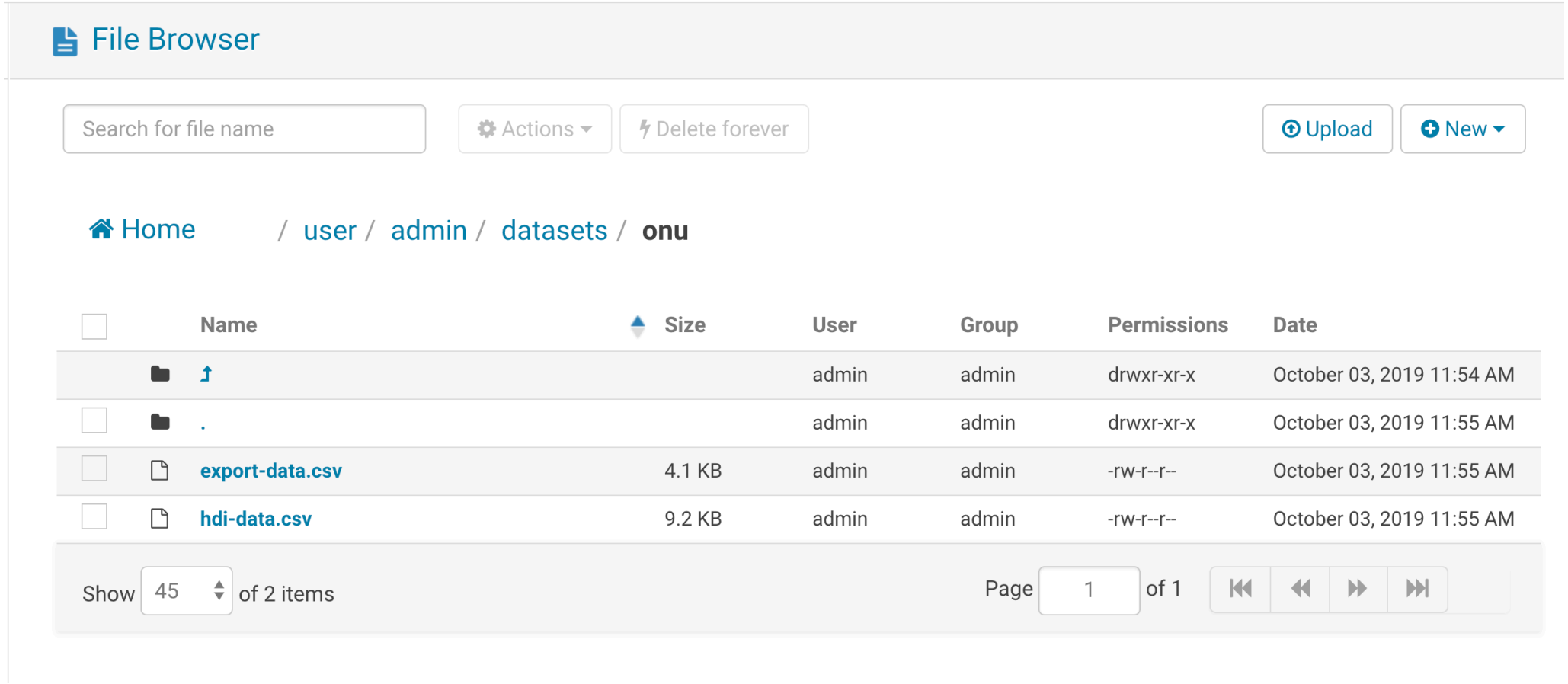

Create Directory
Directory Name
onu
Cancel
Create


File Browser
Search for file name
Actions
Delete forever
Upload
New
Home / user / admin / datasets / onu
Name
Size
User
Group
Permissions
Date
admin admin drwxr-xr-x October 03, 2019 11:54 AM
admin admin drwxr-xr-x October 03, 2019 11:54 AM
Show 45 of 0 items
Page 1 of 1


Query
Search saved documents...
Jobs
admin
Upload to /user/admin/datasets/onu
Select files
or drag and drop them here
Upload
New
Name
Size
User
Group
Permissions
Date
admin
admin
drwxr-xr-x
October 03, 2019 11:54 AM
admin
admin
drwxr-xr-x
October 03, 2019 11:54 AM
Show
45
of 0 items
Page
1
of 1


Query
Search saved documents...
Jobs
admin
Upload to /user/admin/datasets/onu
Select files
or drag and drop them here
Upload
New
Name
Size
User
Group
Permissions
Date
admin
admin
drwxr-xr-x
October 03, 2019 11:54 AM
admin
admin
drwxr-xr-x
October 03, 2019 11:54 AM
Show
45
of 0 items
Page
1
of 1


onu
Search
Favorites
Dropbox
emontoya
Downloads
github
Downloads
Desktop
Documentos
Google Drive
cursos
ST1800
eafit-st0263-2019-1
eafit-st0263-2019-2
Jefatura-DIS
Name
Date Modified
Size
Kind
hdi-data.csv
Oct 3, 2018 at 7:27 AM
9 KB
Comm...et (.csv)
export-data.csv
Jun 27, 2018 at 1:38 PM
4 KB
Comm...et (.csv)
Options
Cancel
Open


File Browser
Search for file name
Actions
Delete forever
Upload
New
Home / user / admin / datasets / onu
Name
Size
User
Group
Permissions
Date
admin admin drwxr-xr-x October 03, 2019 11:54 AM
. admin admin drwxr-xr-x October 03, 2019 11:55 AM
export-data.csv 4.1 KB admin admin -rw-r--r-- October 03, 2019 11:55 AM
hdi-data.csv 9.2 KB admin admin -rw-r--r-- October 03, 2019 11:55 AM
Show 45 of 2 items
Page 1 of 1

File Browser

Back

Edit file

Refresh

View as binary

Download

Last modified
10/03/2019 6:55 PM

User
admin

Group
admin

Size
9.17 KB

Mode
100644

Home

Page 1 to 3 of 3

/ user / admin / datasets / onu / **hdi-data.csv**

```
id,country,Human Development Index (HDI) ,Life expectancy at birth,Mean years of schooling,Expected years o
f schooling,Gross National Income (GNI) per capita,GNI per capita rank minus HDI rank,Nonincome HDI
1,Norway,0.943,81.1,12.6,17.3,47557,6,0.975
2,Australia,0.929,81.9,12,18,34431,16,0.979
3,Netherlands,0.91,80.7,11.6,16.8,36402,9,0.944
4,United States,0.91,78.5,12.4,16,43017,6,0.931
5,New Zealand,0.908,80.7,12.5,18,23737,30,0.978
6,Canada,0.908,81,12.1,16,35166,10,0.944
7,Ireland,0.908,80.6,11.6,18,29322,19,0.959
8,Liechtenstein,0.905,79.6,10.3,14.7,83717,-6,0.877
9,Germany,0.905,80.4,12.2,15.9,34854,8,0.94
10,Sweden,0.904,81.4,11.7,15.7,35837,4,0.936
11,Switzerland,0.903,82.3,11,15.6,39924,0,0.926
12,Japan,0.901,83.4,11.6,15.1,32295,11,0.94
13,Hong Kong China (SAR),0.898,82.8,10,15.7,44805,-4,0.91
14,Iceland,0.898,81.8,10.4,18,29354,11,0.943
15,Korea (Republic of),0.897,80.6,11.6,16.9,28230,12,0.945
16,Denmark,0.895,78.8,11.4,16.9,34347,3,0.926
17,Israel,0.888,81.6,11.9,15.5,25849,14,0.939
18,Belgium,0.886,80,10.9,16.1,33357,2,0.914
19,Austria,0.885,80.9,10.8,15.3,35719,-4,0.908
20,France,0.884,81.5,10.6,16.1,30462,4,0.919
21,Slovenia,0.884,79.3,11.6,16.9,24914,11,0.935
```

Lab Hadoop – HDFS terminal

- **Por Shell o Terminal:**
  - https://hdp1shell.dis.eafit.edu.co
- **Se tienen datos locales al servidor en FS: file:///datasets**
- **Se tienen datos HDFS en: hdfs:///datasets**

Login por Terminal Web

```
ssh -i ~/clei2019.pem hadoop@ec2-13-59-55-130.us-east-2.compute.amazonaws.com
```

Comandos Básicos HDFS

- **Formato:**

$ hdfs dfs -<command-hdfs> <opciones>

Listar archivos y directorios:

$ hdfs dfs -ls /

$ hdfs dfs -ls /datasets

$ hdfs dfs -ls /user

$ hdfs dfs -ls /user/cec##curso

Comandos Básicos HDFS

- **Crear directorio**

$ hdfs dfs -mkdir /tmp/<nombre> (reemplace <nombre> por su propio valor)

- **Copiar archivos al HDFS desde el servidor gateway:**

$ hdfs dfs -mkdir /tmp/<nombre>/data_in

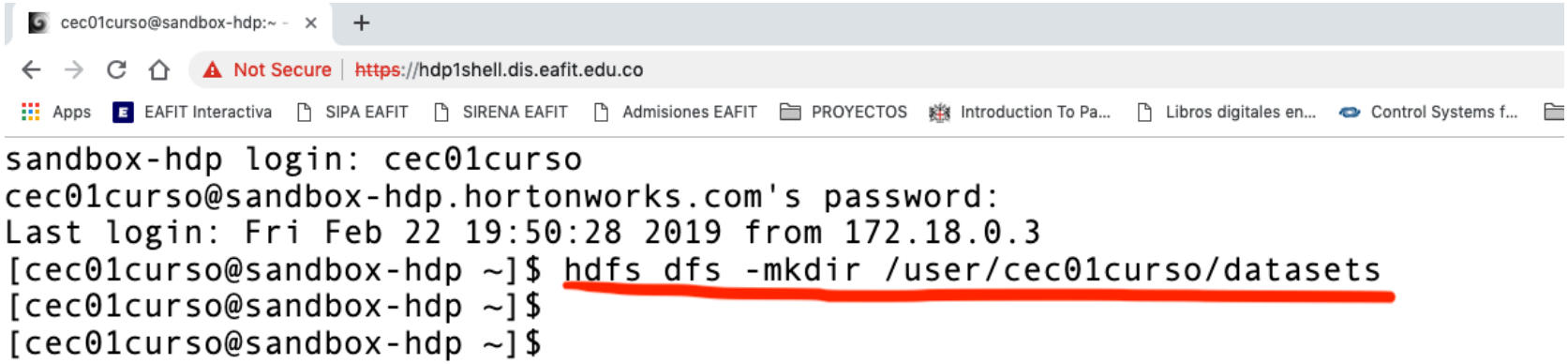
$ hdfs dfs -put *.txt /tmp/<nombre>/data_in

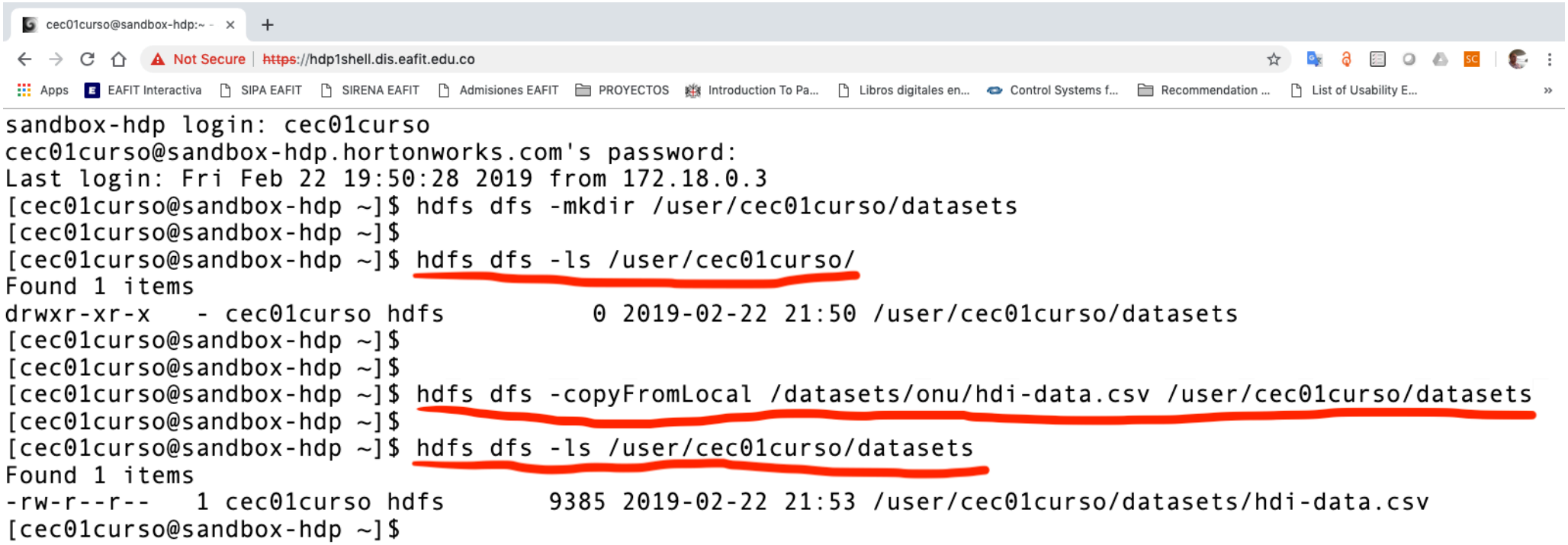
$ hdfs dfs -copyFromLocal *.txt /tmp/<nombre>/data_in (copia recursiva)

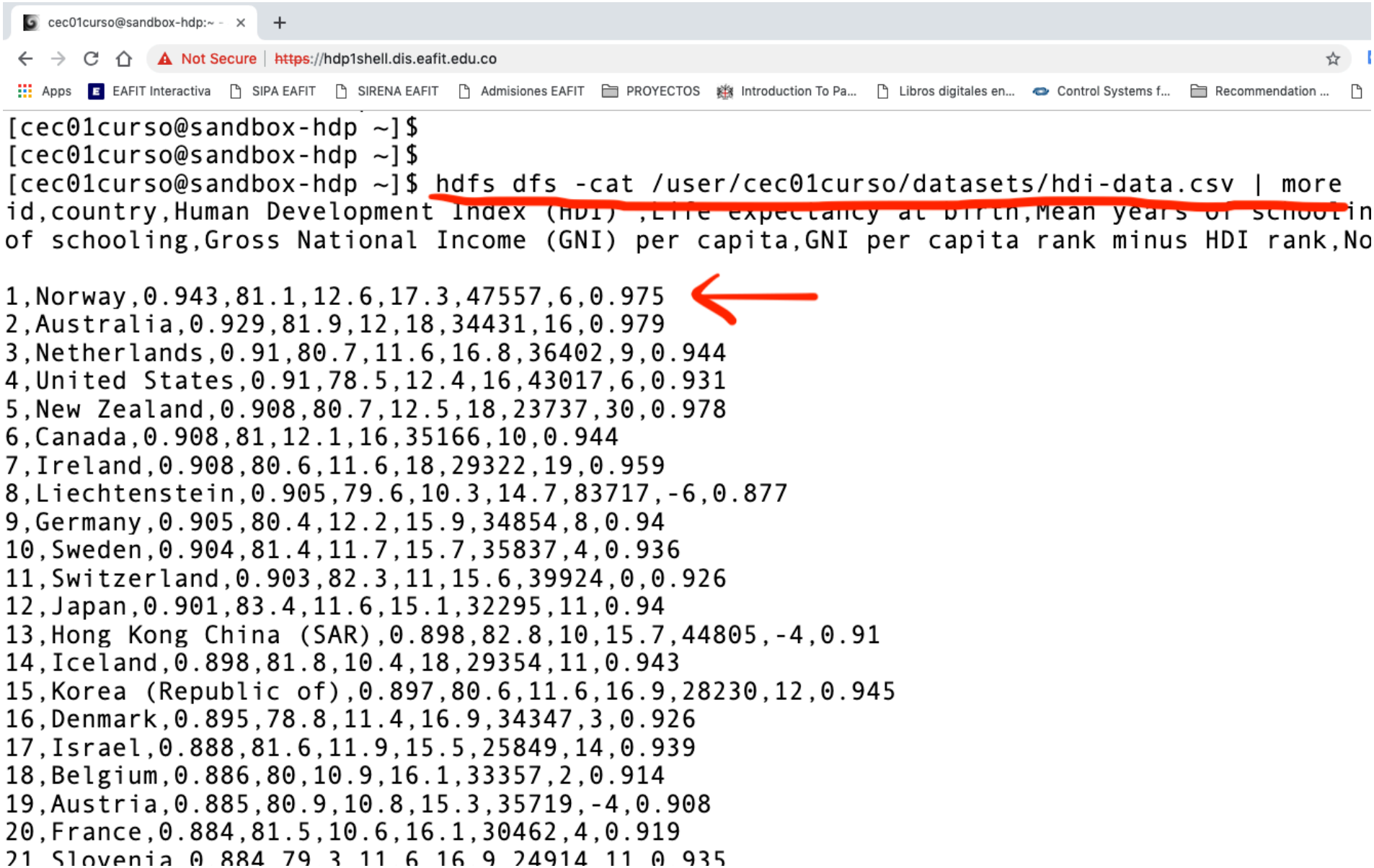
Comandos Básicos HDFS

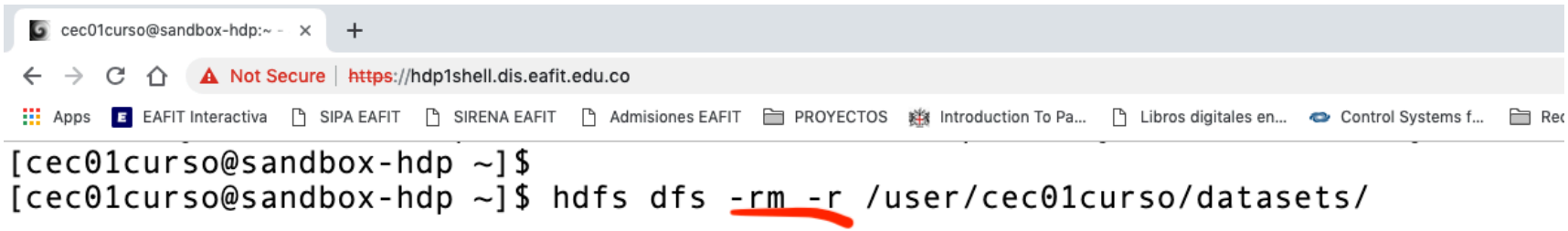
- **Copiar archivos del HDFS al disco servidor gateway:**

$ hdfs dfs -get /tmp/<nombre>/data_in/* /tmp/data_in


cec01curso@sandbox-hdp:~
Not Secure
https://hdp1shell.dis.eafit.edu.co
Apps
EAFIT Interactiva
SIPA EAFIT
SIRENA EAFIT
Admisiones EAFIT
PROYECTOS
Introduction To Pa...
Libros digitales en...
Control Systems f...
sandbox-hdp login: cec01curso
cec01curso@sandbox-hdp.hortonworks.com's password:
Last login: Fri Feb 22 19:50:28 2019 from 172.18.0.3
[cec01curso@sandbox-hdp ~]$ hdfs dfs -mkdir /user/cec01curso/datasets
[cec01curso@sandbox-hdp ~]$
[cec01curso@sandbox-hdp ~]$


cec01curso@sandbox-hdp:~
Not Secure
https://hdp1shell.dis.eafit.edu.co
Apps
EAFIT Interactiva
SIPA EAFIT
SIRENA EAFIT
Admisiones EAFIT
PROYECTOS
Introduction To Pa...
Libros digitales en...
Control Systems f...
Recommendation ...
List of Usability E...
sandbox-hdp login: cec01curso
cec01curso@sandbox-hdp.hortonworks.com's password:
Last login: Fri Feb 22 19:50:28 2019 from 172.18.0.3
[cec01curso@sandbox-hdp ~]$ hdfs dfs -mkdir /user/cec01curso/datasets
[cec01curso@sandbox-hdp ~]$
[cec01curso@sandbox-hdp ~]$ hdfs dfs -ls /user/cec01curso/
Found 1 items
drwxr-xr-x - cec01curso hdfs 0 2019-02-22 21:50 /user/cec01curso/datasets
[cec01curso@sandbox-hdp ~]$
[cec01curso@sandbox-hdp ~]$
[cec01curso@sandbox-hdp ~]$ hdfs dfs -copyFromLocal /datasets/onu/hdi-data.csv /user/cec01curso/datasets
[cec01curso@sandbox-hdp ~]$
[cec01curso@sandbox-hdp ~]$ hdfs dfs -ls /user/cec01curso/datasets
Found 1 items
-rw-r--r-- 1 cec01curso hdfs 9385 2019-02-22 21:53 /user/cec01curso/datasets/hdi-data.csv
[cec01curso@sandbox-hdp ~]$


cec01curso@sandbox-hdp:~
Not Secure
https://hdp1shell.dis.eafit.edu.co
Apps
EAFIT Interactiva
SIPA EAFIT
SIRENA EAFIT
Admisiones EAFIT
PROYECTOS
Introduction To Pa...
Libros digitales en...
Control Systems f...
Recommendation ...
[cec01curso@sandbox-hdp ~]$
[cec01curso@sandbox-hdp ~]$
[cec01curso@sandbox-hdp ~]$ hdfs dfs -cat /user/cec01curso/datasets/hdi-data.csv | more
id,country,Human Development Index (HDI) ,Life expectancy at birth,Mean years of schoolin
of schooling,Gross National Income (GNI) per capita,GNI per capita rank minus HDI rank,No
1,Norway,0.943,81.1,12.6,17.3,47557,6,0.975
2,Australia,0.929,81.9,12,18,34431,16,0.979
3,Netherlands,0.91,80.7,11.6,16.8,36402,9,0.944
4,United States,0.91,78.5,12.4,16,43017,6,0.931
5,New Zealand,0.908,80.7,12.5,18,23737,30,0.978
6,Canada,0.908,81,12.1,16,35166,10,0.944
7,Ireland,0.908,80.6,11.6,18,29322,19,0.959
8,Liechtenstein,0.905,79.6,10.3,14.7,83717,-6,0.877
9,Germany,0.905,80.4,12.2,15.9,34854,8,0.94
10,Sweden,0.904,81.4,11.7,15.7,35837,4,0.936
11,Switzerland,0.903,82.3,11,15.6,39924,0,0.926
12,Japan,0.901,83.4,11.6,15.1,32295,11,0.94
13,Hong Kong China (SAR),0.898,82.8,10,15.7,44805,-4,0.91
14,Iceland,0.898,81.8,10.4,18,29354,11,0.943
15,Korea (Republic of),0.897,80.6,11.6,16.9,28230,12,0.945
16,Denmark,0.895,78.8,11.4,16.9,34347,3,0.926
17,Israel,0.888,81.6,11.9,15.5,25849,14,0.939
18,Belgium,0.886,80,10.9,16.1,33357,2,0.914
19,Austria,0.885,80.9,10.8,15.3,35719,-4,0.908
20,France,0.884,81.5,10.6,16.1,30462,4,0.919
21 Slovenia 0 884 79 3 11 6 16 9 24914 11 0 935



Muy Peligroso -> borra todo el dir 'datasets' con subdirs