

M015: Crèdits de Iliure elecció

UF2-Big Data

APACHE HADOOP I APACHE SPARK (*i ECOSISTEMA*)



Curs: 2019-20

CFGS: DAM2

Alumne: Arnau Subirós Puigarnau

Data:

Nom i Cognoms	Data
Arnau Subirós Puigarnau	21-04-2020

APACHE HADOOP I APACHE SPARK (i ECOSISTEMA)

Descripció

Es vol implementar una solució de Big Data a partir de diversos orígens de dades obtinguts de <https://www.kaggle.com/datasets> Cada grup triarà 3 orígens de dades diferents. Un s'importarà directament des de fitxers, l'altre des d'un SGBDR i el tercer han de ser dades en streaming importades amb Apache Flume.

Objectius

- Crear una infraestructura d'emmagatzematge distribuït amb Hadoop.
- Utilizar diversos orígens de dades i importar-hos en HDFS.
- Utilitzar consultes Hive/Impala sobre els fitxers emmagatzemats.
- Fer ús de dades en streaming
- Realitzar processament distribuït amb Spark.
- Consultes de dades estructurades amb Spark SQL
- Visualitzar les dades emmagatzemades amb Plot.ly i Apache Zeppelin.
- Utilització d'algorismes de ML amb Spark

Nom i Cognoms	Data
Arnau Subirós Puigarnau	21-04-2020

INDEX

pàgina

1. Importació de dades al HDFS

- [Introducció](#)
- [Formats:](#)
 - [Textfile](#)
 - [Avro](#)

2. Recollida de dades en temps real

3. Transformació i persistència d'un RDD amb spark shell

4. Creació d'una aplicació spark

5. Explotació de dades amb spark.

6. Utilització de Spark SQL

7. Visualització de dades

8. Bibliografia

- https://docs.cloudera.com/documentation/enterprise/6/6.3/topics/impala_create_table.html
- <https://www.kaggle.com/datasets>
- <https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset>
- https://docs.cloudera.com/documentation/enterprise/6/6.3/topics/impala_avro.html#avro_create_table
- https://docs.cloudera.com/documentation/other/tutorial/CDH5/topics/ht_flume_to_hdfs.html
- <https://stackoverflow.com/questions/18657548/how-to-setup-a-http-source-for-testing-flume-setup>
- <http://www.bigdatareflections.net/blog/?p=35>
- <https://www.cloudsigma.com/realtime-twitter-data-ingestion-using-flume/>
- <http://flume.apache.org/FlumeUserGuide.html>

Nom i Cognoms	Data
Arnau Subirós Puigarnau	21-04-2020

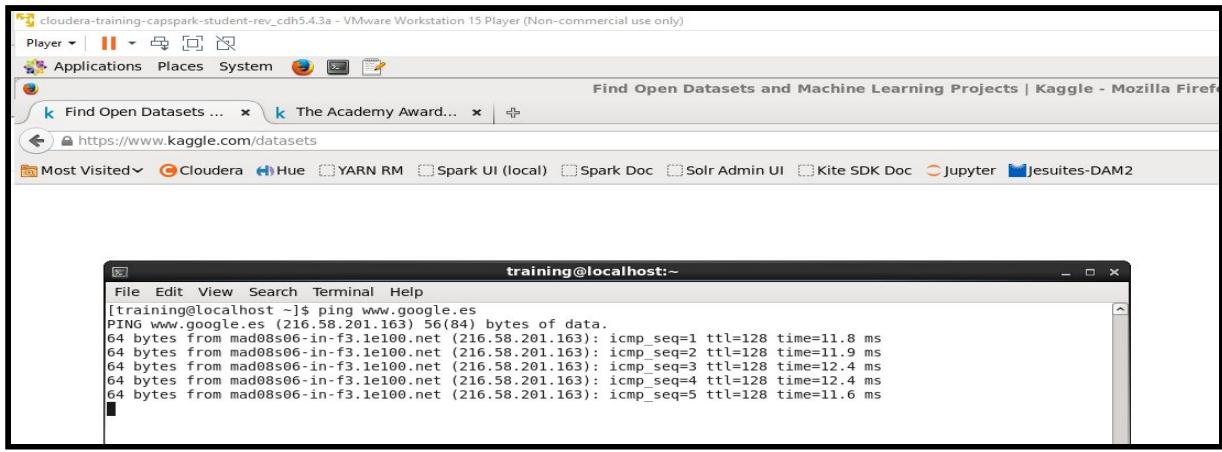
1. Importació de dades al HDFS

- **Introducció**

- Primer de tot obrim la màquina virtual



- Des de el navegador Firefox accedim a <https://www.kaggle.com/datasets> a la recerca de dades. En la màquina virtual, confirmo que tinc connexió a la xarxa, però la versió del navegador deu ser obsolet i no es pot visualitzar tot correctament.



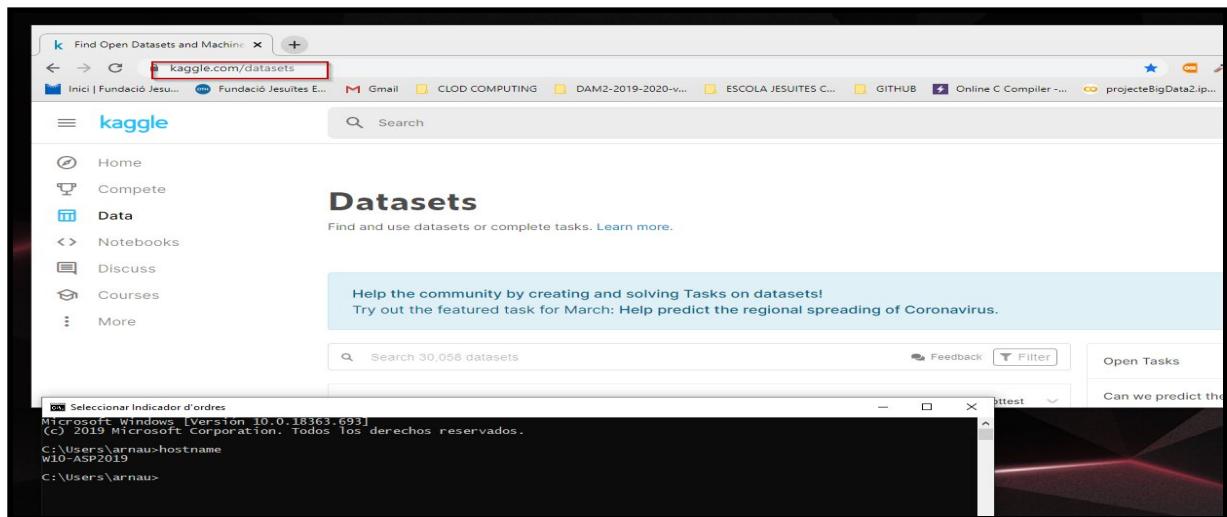
- Per tan desde el Windows 10 faré la recerca i guardarà l'arxiu csv en una carpeta compartida.

Nom i Cognoms

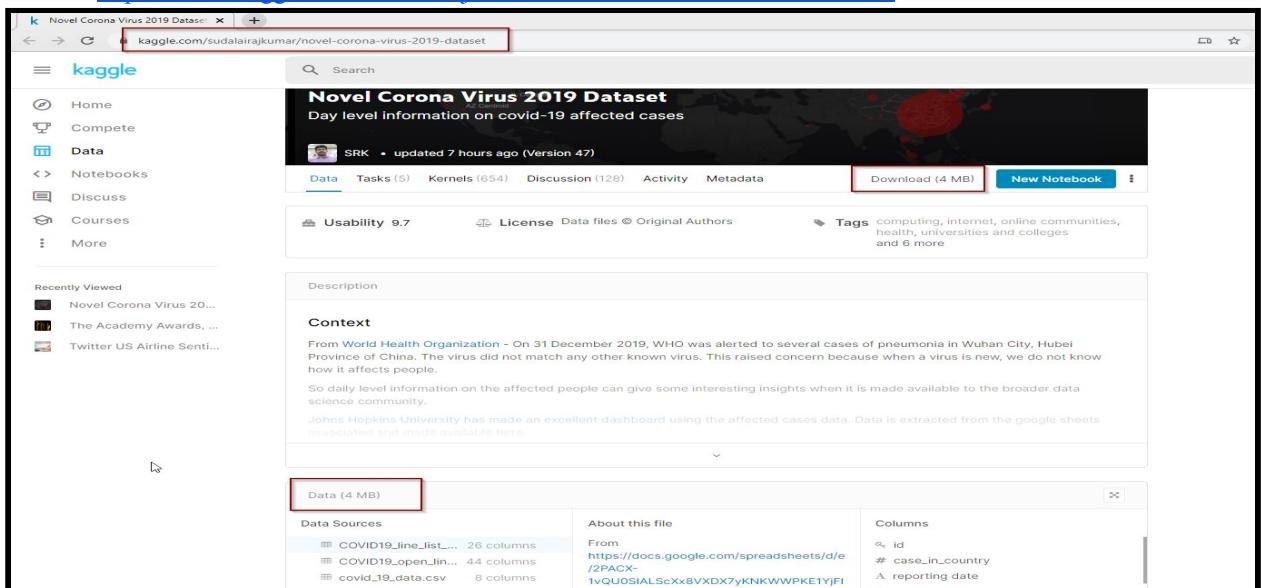
Arnau Subirós Puigarnau

Data

21-04-2020



- Seleccione els datasets amb els quals voldrem treballar
<https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset>



Data Sources	About this file	Columns
<ul style="list-style-type: none"> COVID19_line_list... COVID19_open_jin... covid_19_data.csv 	<p>From: https://docs.google.com/spreadsheets/d/e/2PACX-1vQUOSIALScxx8VxDX7yKNKWWPK1yjFI</p>	<ul style="list-style-type: none"> id case_in_country reporting date

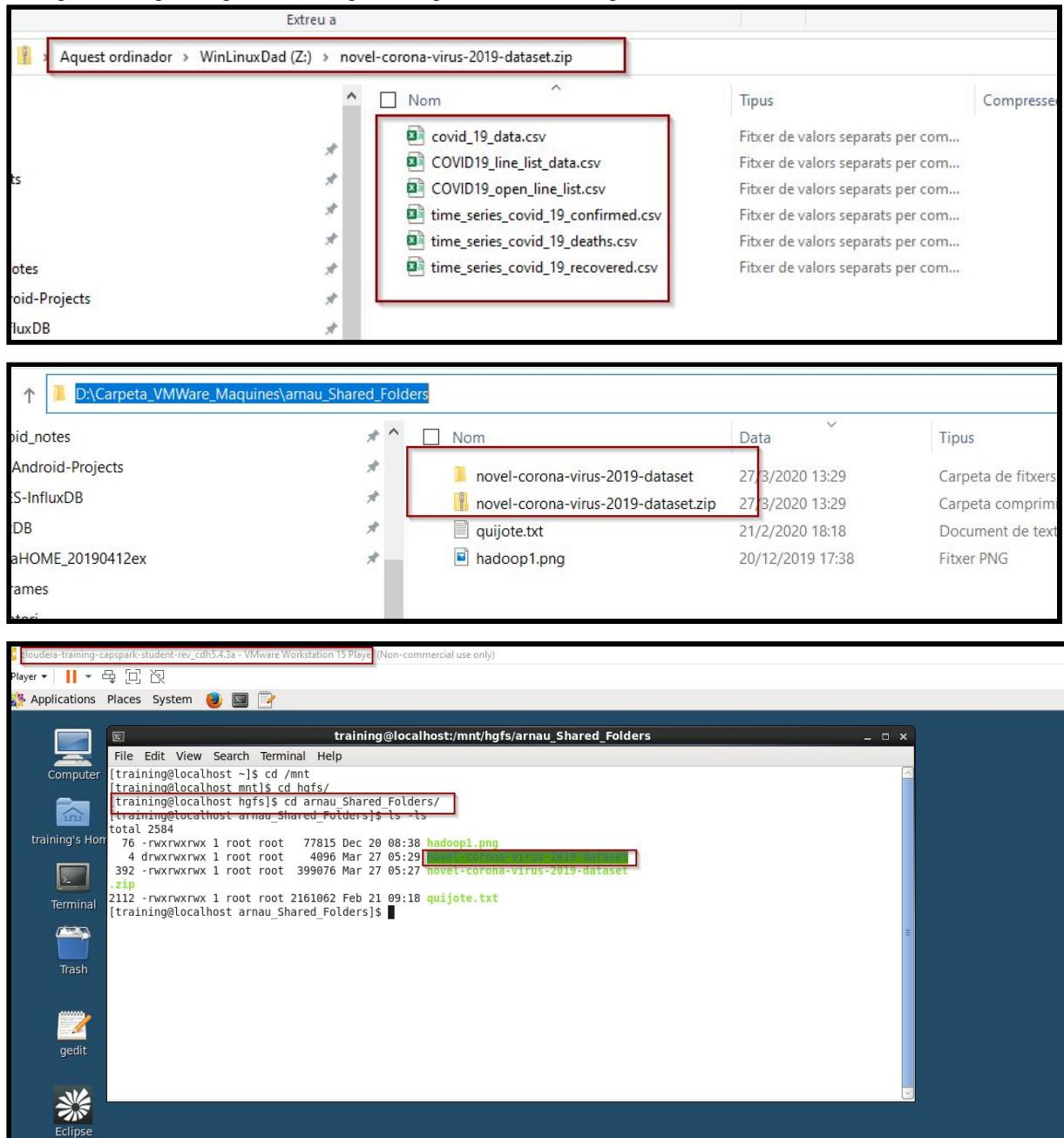
Nom i Cognoms

Arnau Subirós Puigarnau

Data

21-04-2020

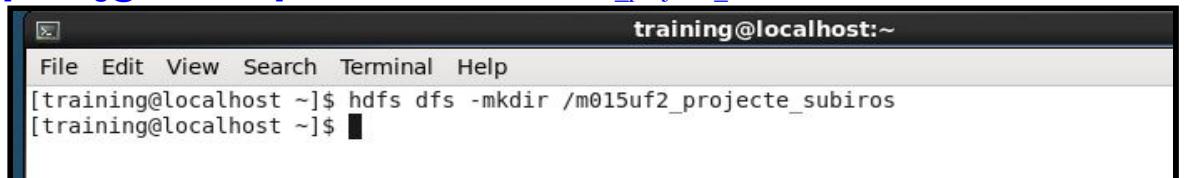
- Un cop tinc el zip, el copio a una carpeta compartida i el descomprimeixo



Nom i Cognoms	Data
Arnau Subirós Puigarnau	21-04-2020

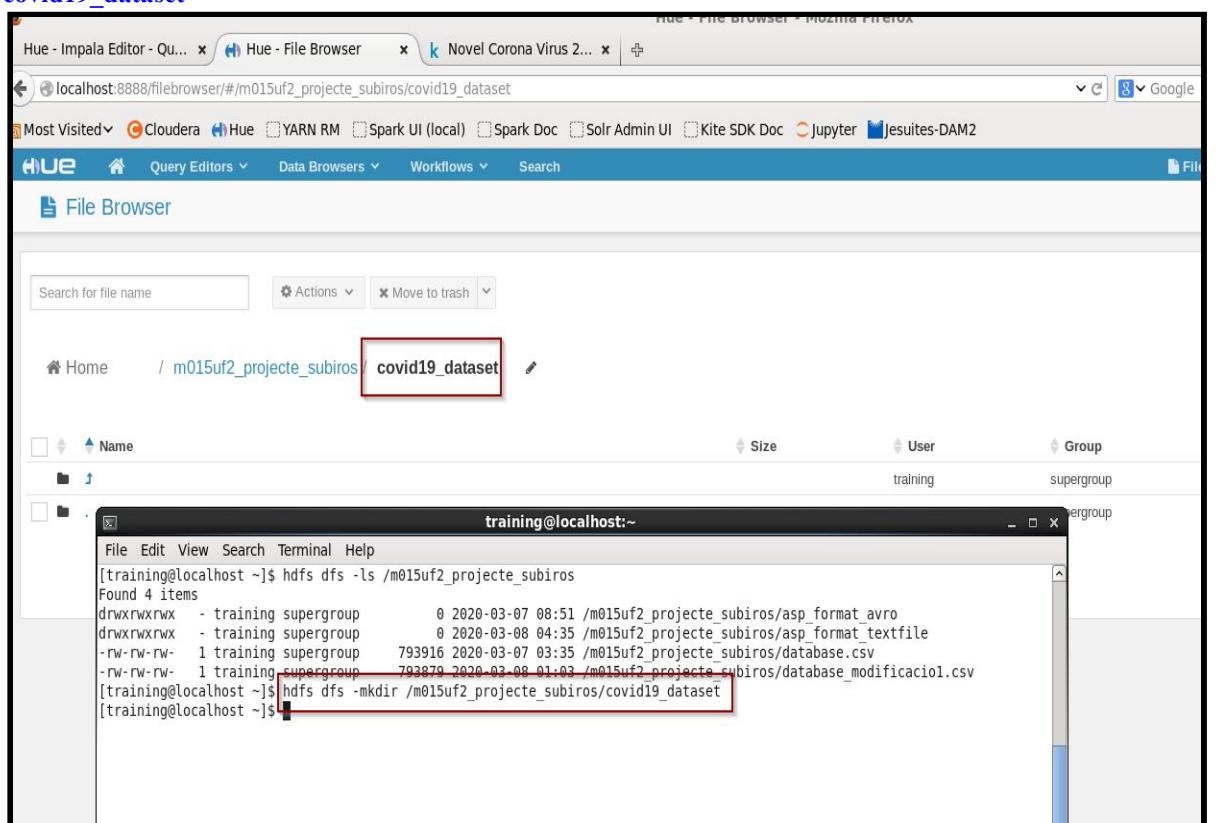
- Ara que tenim l'arxiu descarregat ens interessa guardar-ho a HDFS, però abans d'això crearem un carpeta anomenada : [m015uf2_projecte_subiros](#)

[training@localhost ~]\$ hdfs dfs -mkdir /m015uf2_projecte_subiros



```
File Edit View Search Terminal Help
[training@localhost ~]$ hdfs dfs -mkdir /m015uf2_projecte_subiros
[training@localhost ~]$
```

- Com que encara no vull eliminar les proves que he fet amb l'anterior dataset, creare una directori anomenat [covid19_dataset](#)



Hue - Impala Editor - Qu... x Hue - File Browser x Novel Corona Virus 2... x +

localhost:8888/filebrowser/#/m015uf2_projecte_subiros/covid19_dataset

Most Visited ▾ Cloudera Hue YARN RM Spark UI (local) Spark Doc Solr Admin UI Kite SDK Doc Jupyter Jesuites-DAM2

HUE Home Query Editors Data Browsers Workflows Search

File Browser

Search for file name Actions Move to trash

Home / m015uf2_projecte_subiros covid19_dataset

Name	Size	User	Group
training		training	supergroup

training@localhost:~

```
[training@localhost ~]$ hdfs dfs -ls /m015uf2_projecte_subiros
Found 4 items
drwxrwxrwx - training supergroup 0 2020-03-07 08:51 /m015uf2_projecte_subiros/asp_format_avro
drwxrwxrwx - training supergroup 0 2020-03-08 04:35 /m015uf2_projecte_subiros/asp_format_textfile
-rw-rw-rw- 1 training supergroup 793916 2020-03-07 03:35 /m015uf2_projecte_subiros/database.csv
-rw-rw-rw- 1 training supergroup 793879 2020-03-08 01:03 /m015uf2_projecte_subiros/database_modificacio1.csv
[training@localhost ~]$ hdfs dfs -mkdir /m015uf2_projecte_subiros/covid19_dataset
[training@localhost ~]$
```

Nom i Cognoms

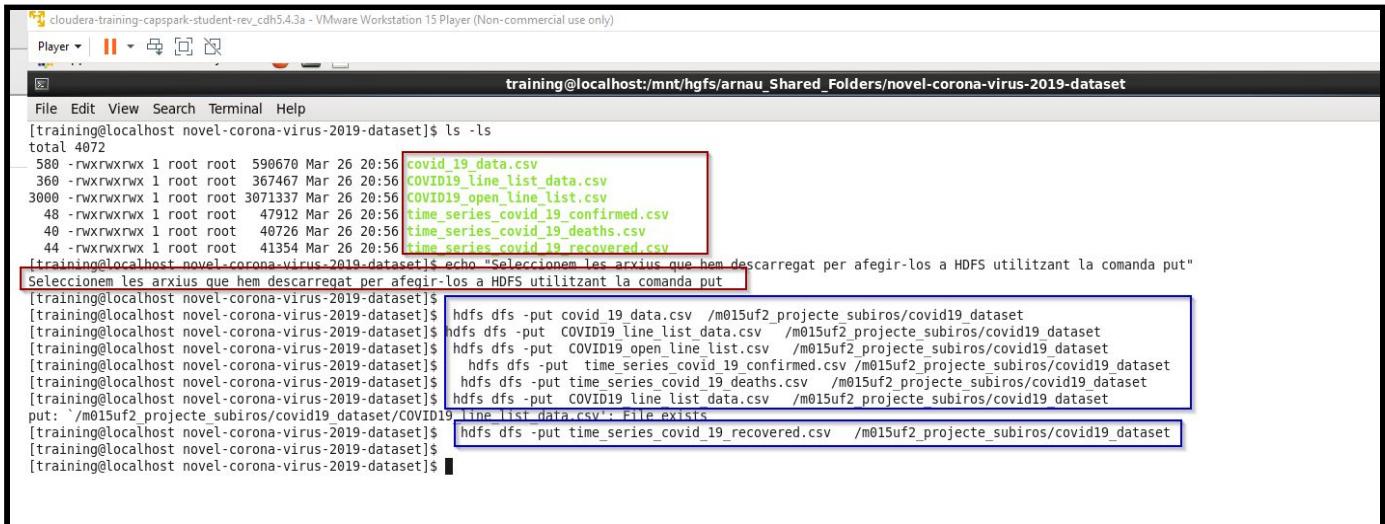
Arnau Subirós Puigarnau

Data

21-04-2020

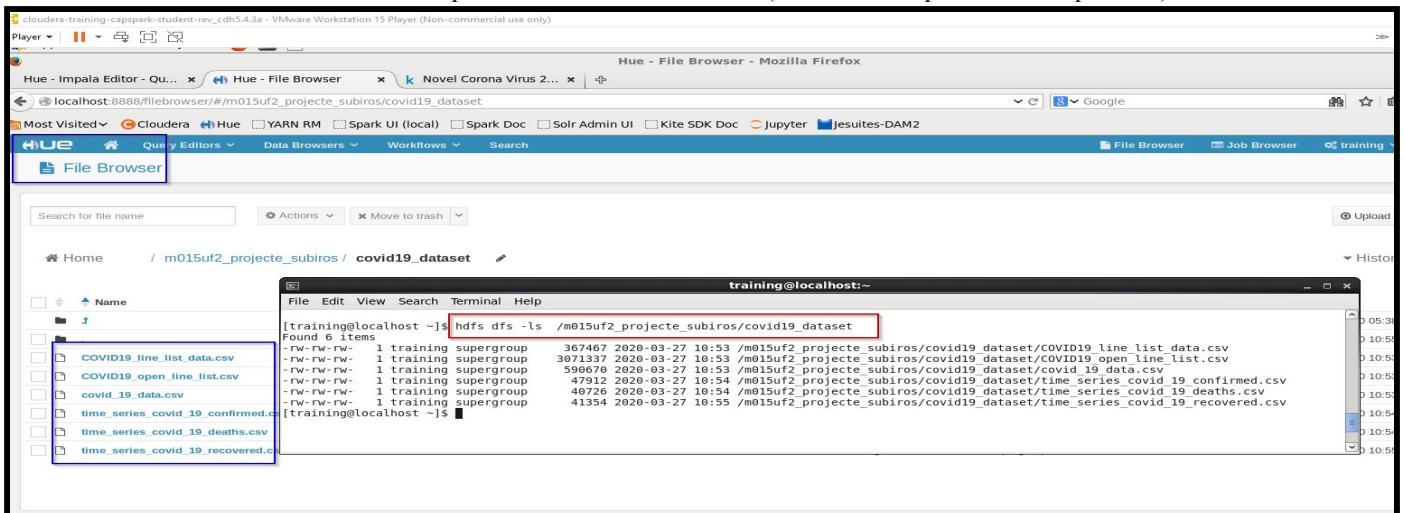
- Inserció de dades a HDFS

```
[training@localhost ~]$ hdfs dfs -put covid_19_data.csv /m015uf2_proyecte_subiros/covid19_dataset
[training@localhost ~]$ hdfs dfs -put COVID19_line_list_data.csv /m015uf2_proyecte_subiros/covid19_dataset
[training@localhost ~]$ hdfs dfs -put COVID19_open_line_list.csv /m015uf2_proyecte_subiros/covid19_dataset
[training@localhost ~]$ hdfs dfs -put time_series_covid_19_confirmed.csv /m015uf2_proyecte_subiros/covid19_dataset
[training@localhost ~]$ hdfs dfs -put time_series_covid_19_deaths.csv /m015uf2_proyecte_subiros/covid19_dataset
[training@localhost ~]$ hdfs dfs -put time_series_covid_19_recovered.csv /m015uf2_proyecte_subiros/covid19_dataset
```



```
File Edit View Search Terminal Help
[training@localhost novel-corona-virus-2019-dataset]$ ls -ls
total 4072
580 -rwxrwxrwx 1 root root 590670 Mar 26 20:56 covid_19_data.csv
360 -rwxrwxrwx 1 root root 367467 Mar 26 20:56 COVID19_line_list_data.csv
3000 -rwxrwxrwx 1 root root 3071337 Mar 26 20:56 COVID19_open_line_list.csv
48 -rwxrwxrwx 1 root root 47912 Mar 26 20:56 time_series_covid_19_confirmed.csv
40 -rwxrwxrwx 1 root root 40726 Mar 26 20:56 time_series_covid_19_deaths.csv
44 -rwxrwxrwx 1 root root 41354 Mar 26 20:56 time_series_covid_19_recovered.csv
[training@localhost novel-corona-virus-2019-dataset]$ echo "Seleccionem les arxius que hem descarregat per afegir-los a HDFS utilitzant la comanda put"
[Selectem els arxius que hem descarregat per afegir-los a HDFS utilitzant la comanda put]
[training@localhost novel-corona-virus-2019-dataset]$ hdfs dfs -put covid_19_data.csv /m015uf2_proyecte_subiros/covid19_dataset
[training@localhost novel-corona-virus-2019-dataset]$ hdfs dfs -put COVID19_line_list_data.csv /m015uf2_proyecte_subiros/covid19_dataset
[training@localhost novel-corona-virus-2019-dataset]$ hdfs dfs -put COVID19_open_line_list.csv /m015uf2_proyecte_subiros/covid19_dataset
[training@localhost novel-corona-virus-2019-dataset]$ hdfs dfs -put time_series_covid_19_confirmed.csv /m015uf2_proyecte_subiros/covid19_dataset
[training@localhost novel-corona-virus-2019-dataset]$ hdfs dfs -put time_series_covid_19_deaths.csv /m015uf2_proyecte_subiros/covid19_dataset
[training@localhost novel-corona-virus-2019-dataset]$ hdfs dfs -put COVID19_line_list_data.csv /m015uf2_proyecte_subiros/covid19_dataset
put: '/m015uf2_proyecte_subiros/covid19_dataset/COVID19_line_list_data.csv': File exists
[training@localhost novel-corona-virus-2019-dataset]$ hdfs dfs -put time_series_covid_19_recovered.csv /m015uf2_proyecte_subiros/covid19_dataset
[training@localhost novel-corona-virus-2019-dataset]$
```

- Desde HDFS confirmo que s'han creat correctament. (ho visualitzo per terminal i per Hue)



```
Hue - Impala Editor - Qu... □ Hue - File Browser □ Novel Corona Virus 2... □
localhost:8888/filebrowser/#/m015uf2_proyecte_subiros/covid19_dataset
File Browser □ Job Browser □ training □
File Browser

Search for file name Actions Move to trash

Name
COVID19_line_list_data.csv
COVID19_open_line_list.csv
covid_19_data.csv
time_series_covid_19_confirmed.csv
time_series_covid_19_deaths.csv
time_series_covid_19_recovered.csv

File Edit View Search Terminal Help
[training@localhost ~]$ hdfs dfs -ls /m015uf2_proyecte_subiros/covid19_dataset
Found 6 items
-rw-rw-rw- 1 training supergroup 367467 2020-03-27 10:53 /m015uf2_proyecte_subiros/covid19_dataset/COVID19_line_list_data.csv
-rw-rw-rw- 1 training supergroup 3071337 2020-03-27 10:53 /m015uf2_proyecte_subiros/covid19_dataset/COVID19_open_line_list.csv
-rw-rw-rw- 1 training supergroup 580 2020-03-27 10:53 /m015uf2_proyecte_subiros/covid19_dataset/covid_19_data.csv
-rw-rw-rw- 1 training supergroup 47912 2020-03-27 10:54 /m015uf2_proyecte_subiros/covid19_dataset/time_series_covid_19_confirmed.csv
-rw-rw-rw- 1 training supergroup 40726 2020-03-27 10:54 /m015uf2_proyecte_subiros/covid19_dataset/time_series_covid_19_deaths.csv
-rw-rw-rw- 1 training supergroup 41354 2020-03-27 10:55 /m015uf2_proyecte_subiros/covid19_dataset/time_series_covid_19_recovered.csv
```

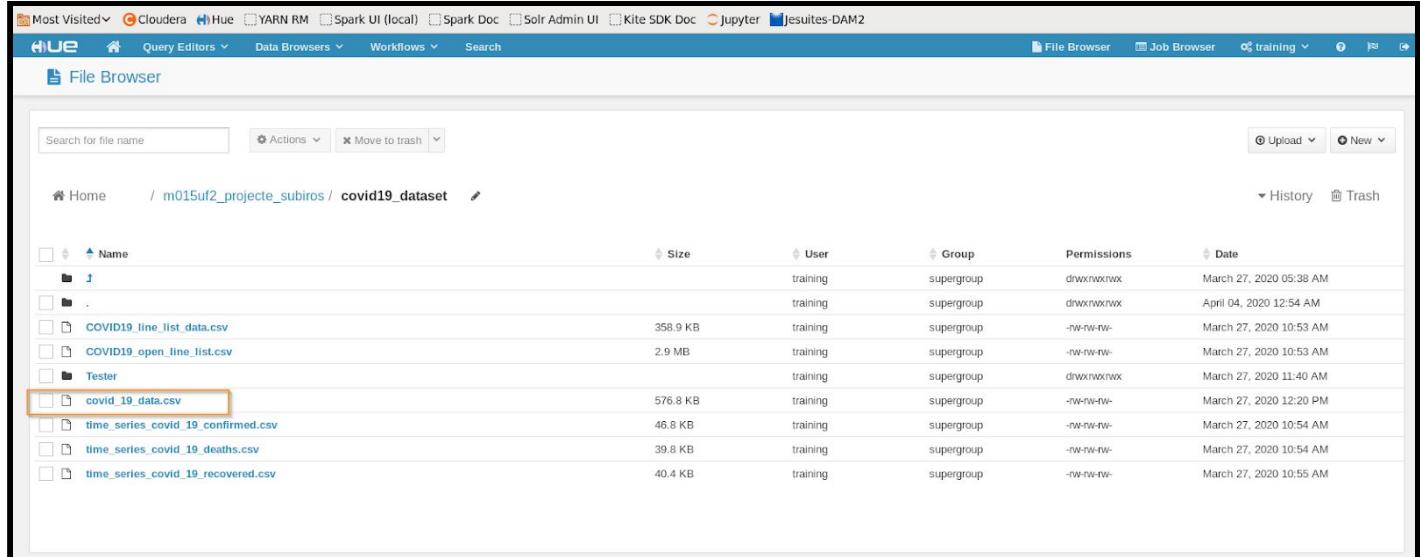
Nom i Cognoms	Data
Arnau Subirós Puigarnau	21-04-2020

● FORMAT : TEXTFILE

Ens basarem en l' arxiu : [covid_19_data.csv](#)

```
CREATE EXTERNAL TABLE COVID19TESTER3 (
Serial int,
ObservationDate String,
Province String ,
State String,
Country String,
Region String,
LastUpdate String,
Confirmed float,
Deaths float,
Recovered float
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
```

STORED AS TEXTFILE
LOCATION '/m015uf2_proyecto_subiros/covid19_dataset';



The screenshot shows the Apache Hue File Browser interface. The top navigation bar includes links for Most Visited, Cloudera, Hue, YARN RM, Spark UI (local), Spark Doc, Solr Admin UI, Kite SDK Doc, Jupyter, and jesuites-DAM2. The main area is titled 'File Browser' and shows a list of files under the path '/m015uf2_proyecto_subiros / covid19_dataset'. The files listed are:

Name	Size	User	Group	Permissions	Date
.		training	supergroup	drwxrwxrwx	March 27, 2020 05:38 AM
COVID19_line_list_data.csv	358.9 KB	training	supergroup	drwxrwxrwx	April 04, 2020 12:54 AM
COVID19_open_line_list.csv	2.9 MB	training	supergroup	-rw-rw-rw-	March 27, 2020 10:53 AM
Tester		training	supergroup	drwxrwxrwx	March 27, 2020 11:40 AM
covid_19_data.csv	576.8 KB	training	supergroup	-rw-rw-rw-	March 27, 2020 12:20 PM
time_series_covid_19_confirmed.csv	46.6 KB	training	supergroup	-rw-rw-rw-	March 27, 2020 10:54 AM
time_series_covid_19_deaths.csv	39.8 KB	training	supergroup	-rw-rw-rw-	March 27, 2020 10:54 AM
time_series_covid_19_recovered.csv	40.4 KB	training	supergroup	-rw-rw-rw-	March 27, 2020 10:55 AM

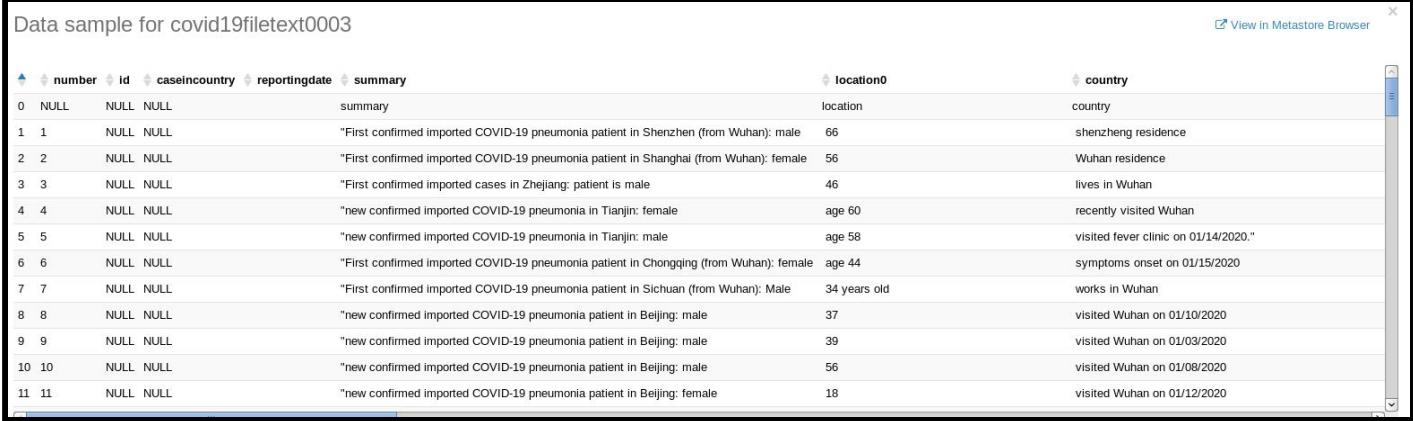
Nom i Cognoms
Data

Arnaud Subirós Puigarnau

21-04-2020

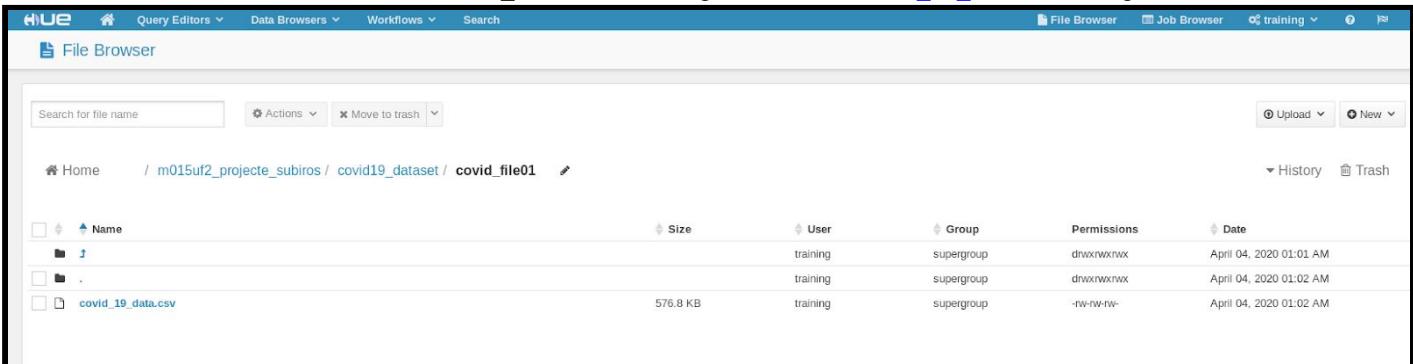
- Com que hi ha diversos arxius la consulta produeix errors

Data sample for covid19filetext00003



	number	id	caseincountry	reportingdate	summary	location	country
0	NULL	NULL	NULL		summary	location	country
1	1	NULL	NULL		"First confirmed imported COVID-19 pneumonia patient in Shenzhen (from Wuhan): male	66	shenzheng residence
2	2	NULL	NULL		"First confirmed imported COVID-19 pneumonia patient in Shanghai (from Wuhan): female	56	Wuhan residence
3	3	NULL	NULL		"First confirmed imported cases in Zhejiang: patient is male	46	lives in Wuhan
4	4	NULL	NULL		"new confirmed imported COVID-19 pneumonia in Tianjin: female	age 60	recently visited Wuhan
5	5	NULL	NULL		"new confirmed imported COVID-19 pneumonia in Tianjin: male	age 58	visited fever clinic on 01/14/2020."
6	6	NULL	NULL		"First confirmed imported COVID-19 pneumonia patient in Chongqing (from Wuhan): female	age 44	symptoms onset on 01/15/2020
7	7	NULL	NULL		"First confirmed imported COVID-19 pneumonia patient in Sichuan (from Wuhan): Male	34 years old	works in Wuhan
8	8	NULL	NULL		"new confirmed imported COVID-19 pneumonia patient in Beijing: male	37	visited Wuhan on 01/10/2020
9	9	NULL	NULL		"new confirmed imported COVID-19 pneumonia patient in Beijing: male	39	visited Wuhan on 01/03/2020
10	10	NULL	NULL		"new confirmed imported COVID-19 pneumonia patient in Beijing: male	56	visited Wuhan on 01/08/2020
11	11	NULL	NULL		"new confirmed imported COVID-19 pneumonia patient in Beijing: female	18	visited Wuhan on 01/12/2020

- Creo una subdirector anomenat covid_file01 fem una copia de l'arxiu [covid_19_data.csv](#) i el dipositem dins



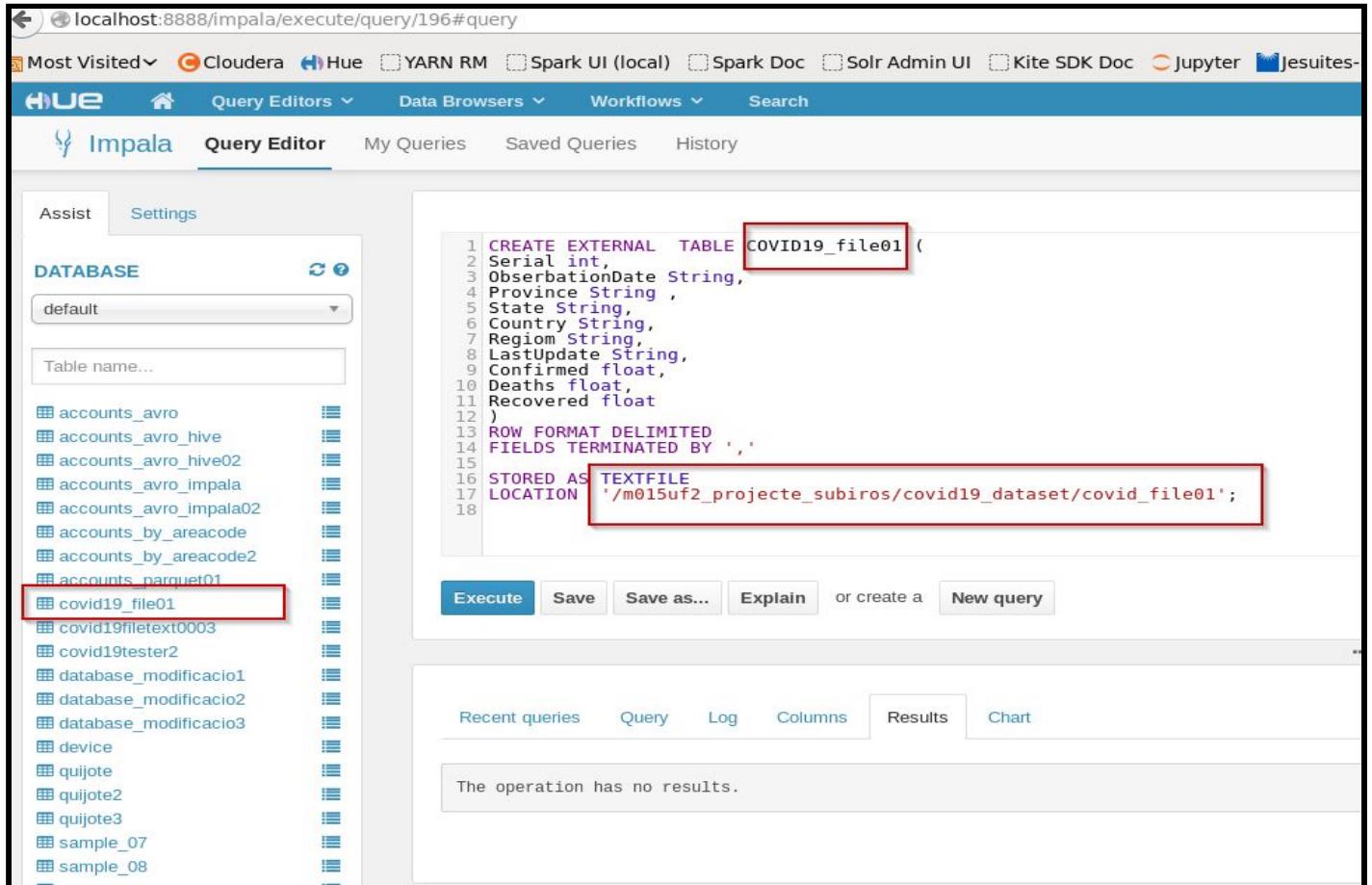
Name	Size	User	Group	Permissions	Date
f		training	supergroup	drwxrwxrwx	April 04, 2020 01:01 AM
.		training	supergroup	drwxrwxrwx	April 04, 2020 01:02 AM
covid_19_data.csv	576.8 KB	training	supergroup	-rw-rw-rw-	April 04, 2020 01:02 AM

Nom i Cognoms	Data
Arnau Subirós Puigarnau	21-04-2020

```

CREATE EXTERNAL TABLE COVID19_File01 (
    Serial int,
    ObservationDate String,
    Province String ,
    State String,
    Country String,
    Region String,
    LastUpdate String,
    Confirmed float,
    Deaths float,
    Recovered float
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
    
```

STORED AS TEXTFILE
 LOCATION '/m015uf2_projecte_subiros/covid19_dataset/covid_file01';



The screenshot shows the Hue Query Editor interface. On the left, the 'DATABASE' dropdown is set to 'default'. A list of tables is shown, with 'covid19_file01' highlighted by a red box. In the central 'Query Editor' pane, a code editor contains the SQL command for creating the external table, also enclosed in a red box. The code is as follows:

```

1 CREATE EXTERNAL TABLE COVID19_file01 (
2     Serial int,
3     ObservationDate String,
4     Province String ,
5     State String,
6     Country String,
7     Region String,
8     LastUpdate String,
9     Confirmed float,
10    Deaths float,
11    Recovered float
12 )
13 ROW FORMAT DELIMITED
14 FIELDS TERMINATED BY ','
15
16 STORED AS TEXTFILE
17 LOCATION '/m015uf2_projecte_subiros/covid19_dataset/covid_file01';
18
    
```

Below the code editor are several buttons: 'Execute' (highlighted in blue), 'Save', 'Save as...', 'Explain', and 'New query'. At the bottom of the editor, it says 'The operation has no results.'

Nom i Cognoms

Arnaud Subirós Puigarnau

Data

21-04-2020

Hive Editor Query Editor My Queries Saved Queries History

Data sample for covid19_file01

	serial	observationdate	province	state	country	region	lastupdate	confirmed	deaths	recovered
0	NULL	ObservationDate	Province/State	Country/Region	Last Update	Confirmed	Deaths	NULL	NULL	NULL
1	1	01/22/2020	Anhui	Mainland China	1/22/2020 17:00	1.0	0.0	0.0	NULL	NULL
2	2	01/22/2020	Beijing	Mainland China	1/22/2020 17:00	14.0	0.0	0.0	NULL	NULL
3	3	01/22/2020	Chongqing	Mainland China	1/22/2020 17:00	6.0	0.0	0.0	NULL	NULL
4	4	01/22/2020	Fujian	Mainland China	1/22/2020 17:00	1.0	0.0	0.0	NULL	NULL
5	5	01/22/2020	Gansu	Mainland China	1/22/2020 17:00	0.0	0.0	0.0	NULL	NULL
6	6	01/22/2020	Guangdong	Mainland China	1/22/2020 17:00	26.0	0.0	0.0	NULL	NULL
7	7	01/22/2020	Guangxi	Mainland China	1/22/2020 17:00	2.0	0.0	0.0	NULL	NULL
8	8	01/22/2020	Guizhou	Mainland China	1/22/2020 17:00	1.0	0.0	0.0	NULL	NULL
9	9	01/22/2020	Hainan	Mainland China	1/22/2020 17:00	4.0	0.0	0.0	NULL	NULL
10	10	01/22/2020	Hebei	Mainland China	1/22/2020 17:00	1.0	0.0	0.0	NULL	NULL
11	11	01/22/2020	Heilongjiang	Mainland China	1/22/2020 17:00	0.0	0.0	0.0	NULL	NULL
12	12	01/22/2020	Henan	Mainland China	1/22/2020 17:00	5.0	0.0	0.0	NULL	NULL

View in Metastore Browser

Ok

- Reviso la taula , al veure errors , faig unes modificacions i creo la versio 2

localhost:8888/impala/execute/query/201#query/results

Most Visited ▾ Cloudera Hue YARN RM Spark UI (local) Spark Doc Solr Admin UI Kite SDK Doc Jupyter Jesuites-DAM2

HUE Home Query Editors Data Browsers Workflows Search

Impala Query Editor My Queries Saved Queries History

Assist Settings

DATABASE

default

Table name...

accounts_avro accounts_avro_hive accounts_avro_hive02 accounts_avro_impala accounts_avro_impala02 accounts_by_areacode accounts_by_areacode2 accounts_parquet01 covid19_file01 covid19filetext0003 database_modificacio1 database_modificacio2 database_modificacio3 device quiete

```

1 CREATE EXTERNAL TABLE COVID19_file02 (
2     Serial int,
3     ObservationDate String,
4     Province State String,
5     Country Region String,
6     LastUpdate float,
7     Confirmed float,
8     Deaths float,
9     Recovered float
10 )
11 ROW FORMAT DELIMITED
12 FIELDS TERMINATED BY ','
13
14 STORED AS TEXTFILE
15 LOCATION '/m015uf2_projecte_subiros/covid19_dataset/covid_file01';
16

```

Execute Save Save as... Explain or create a New query

Recent queries Query Log Columns Results Chart

Nom i Cognoms

Arnaud Subirós Puigarnau

Data

21-04-2020

Data sample for covid19_file01

[View in Metastore Browser](#)

serial	observationdate	province_state	country_region	lastupdate	confirmed	deaths	recovered
0	NULL	ObservationDate	Province/State	Country/Region	NULL	NULL	NULL
1	1	01/22/2020	Anhui	Mainland China	NULL	1.0	0.0
2	2	01/22/2020	Beijing	Mainland China	NULL	14.0	0.0
3	3	01/22/2020	Chongqing	Mainland China	NULL	6.0	0.0
4	4	01/22/2020	Fujian	Mainland China	NULL	1.0	0.0
5	5	01/22/2020	Gansu	Mainland China	NULL	0.0	0.0
6	6	01/22/2020	Guangdong	Mainland China	NULL	26.0	0.0
7	7	01/22/2020	Guangxi	Mainland China	NULL	2.0	0.0
8	8	01/22/2020	Guizhou	Mainland China	NULL	1.0	0.0
9	9	01/22/2020	Hainan	Mainland China	NULL	4.0	0.0
10	10	01/22/2020	Hebei	Mainland China	NULL	1.0	0.0
11	11	01/22/2020	Heilongjiang	Mainland China	NULL	0.0	0.0
12	12	01/22/2020	Henan	Mainland China	NULL	5.0	0.0

- Ara ens interessa fer el mateix pero amb un arxiu de més tamany (2.9MB anomenat COVID19_open_line_list.csv) . Per evitar problemes, creare un subdirectorí anomenat **covid_list**(com he fet abans amb aquest arxiu)

localhost:8888/filebrowser/view/m015uf2_projecte_subiros/covid19_dataset

File Browser

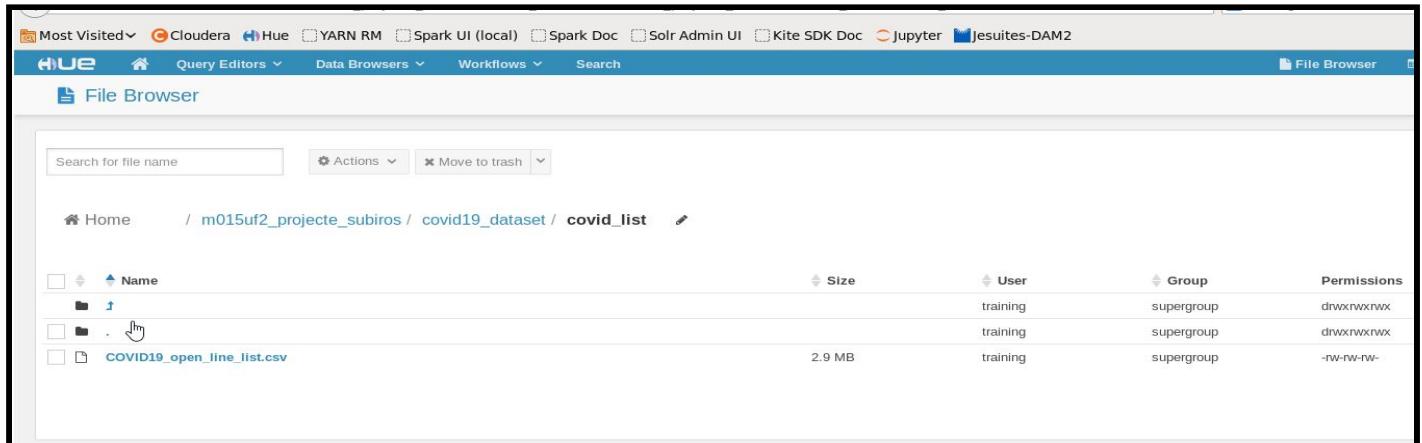
Search for file name Actions Move to trash Upload New

Home / m015uf2_projecte_subiros / covid19_dataset

History Trash

Name	Size	User	Group	Permissions	Date
COVID19_line_list_data.csv	358.9 KB	training	supergroup	drwxrwxrwx	March 27, 2020 10:53 AM
COVID19_open_line_list.csv	2.9 MB	training	supergroup	-rw-rw-rw-	March 27, 2020 10:53 AM
Tester		training	supergroup	drwxrwxrwx	March 27, 2020 11:40 AM
covid_19_data.csv	576.8 KB	training	supergroup	-rw-rw-rw-	March 27, 2020 12:20 PM
covid_file01		training	supergroup	drwxrwxrwx	April 04, 2020 01:02 AM
covid_list		training	supergroup	drwxrwxrwx	April 04, 2020 01:23 AM
time_series_covid_19_confirmed.csv	46.8 KB	training	supergroup	-rw-rw-rw-	March 27, 2020 10:54 AM
time_series_covid_19_deaths.csv	39.8 KB	training	supergroup	-rw-rw-rw-	March 27, 2020 10:54 AM
time_series_covid_19_recovered.csv	40.4 KB	training	supergroup	-rw-rw-rw-	March 27, 2020 10:55 AM

Nom i Cognoms	Data
Arnau Subirós Puigarnau	21-04-2020



Name	Size	User	Group	Permissions
..		training	supergroup	drwxrwxrwx
.		training	supergroup	drwxrwxrwx
COVID19_open_line_list.csv	2.9 MB	training	supergroup	-rw-rw-rw-

Ens basarem en l' arxiu : [COVID19_open_line_list.csv](#)

```

CREATE EXTERNAL TABLE COVID_list01 (
ID int,
Age int,
Sex string,
City String,
Province String,
Country String,
Wuhan_not_Wuhan Boolean,
Latitude float,
Longitude float,
geo_resolution String,
date_on_set_syphoms timestamp,
date_on_set_admission_hospital timestamp,
date_confirmation timestamp,
syphoms String,
lives_in_Wuhan int,
travel_history_dates timestamp,
travel_history_location String,
reported_market_exposure String,
additional_information,
chronic_disease_binary,
chronic_disease String,
source String,
sequence_available String,
outcome String,
date_death_or_discharge String,
notes_for_discussion String,
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','

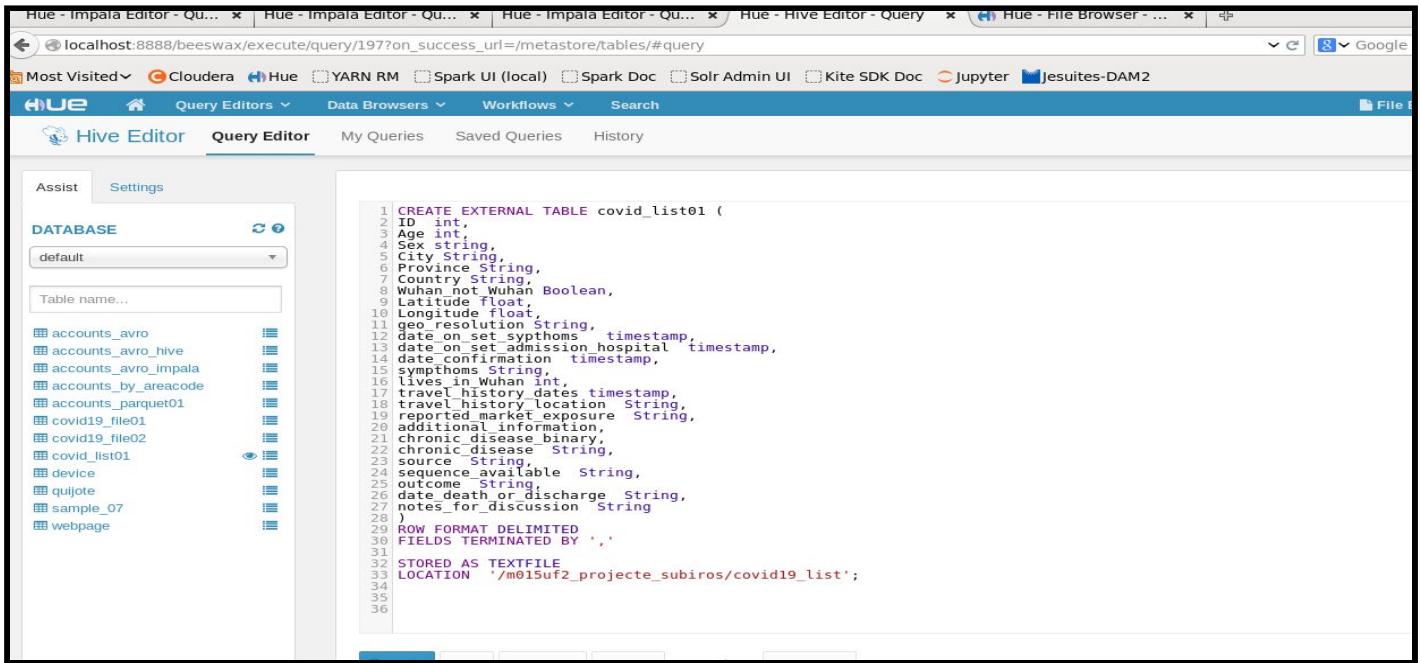
STORED AS TEXTFILE
LOCATION '/m015uf2_projecte_subiros/covid19_dataset/covid19_list';

```

Nom i Cognoms
Data

Arnau Subirós Puigarnau

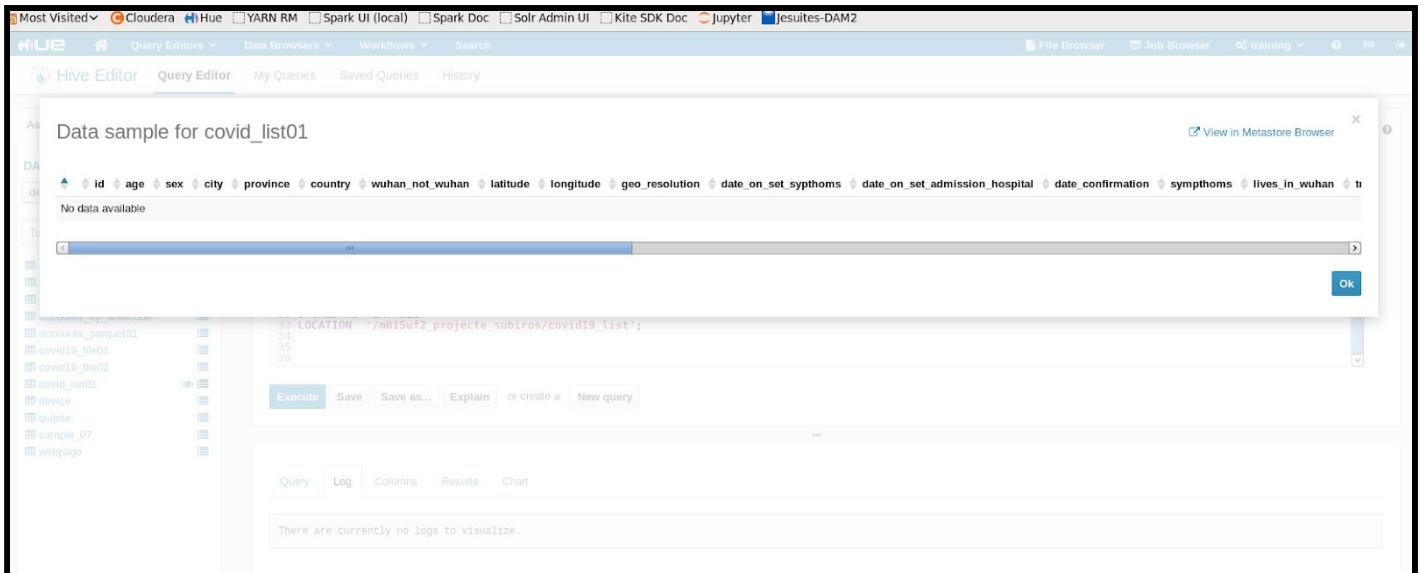
21-04-2020



```

1 CREATE EXTERNAL TABLE covid_list01 (
2   ID int,
3   Age int,
4   Sex string,
5   City String,
6   Province String,
7   Country String,
8   Wuhan_not_Wuhan Boolean,
9   Latitude float,
10  Longitude float,
11  geo_resolution String,
12  date_on_set_syphoms timestamp,
13  date_on_set_admission_hospital timestamp,
14  date_confirmation timestamp,
15  syphoms String,
16  lives_in_Wuhan int,
17  travel_history_dates timestamp,
18  travel_history_location String,
19  reported_market_exposure String,
20  additional_information,
21  chronic_disease_binary,
22  chronic_disease_string,
23  source_String,
24  sequence_available String,
25  outcome String,
26  date_of_death_discharge String,
27  notes_for_discussion String
28 )
29 ROW FORMAT DELIMITED
30 FIELDS TERMINATED BY ','
31
32 STORED AS TEXTFILE
33 LOCATION '/m015uf2_proyecto_subiros/covid19_list';
34
35
36

```

Pendent de solucionar el problema


id	age	sex	city	province	country	wuhan_not_wuhan	latitude	longitude	geo_resolution	date_on_set_syphoms	date_on_set_admission_hospital	date_confirmation	symptoms	lives_in_wuhan	outcome

No data available

LOCATION '/m015uf2_proyecto_subiros/covid19_list';

Execute Save Save as... Explain or create a New query

Query Log Columns Results Chart

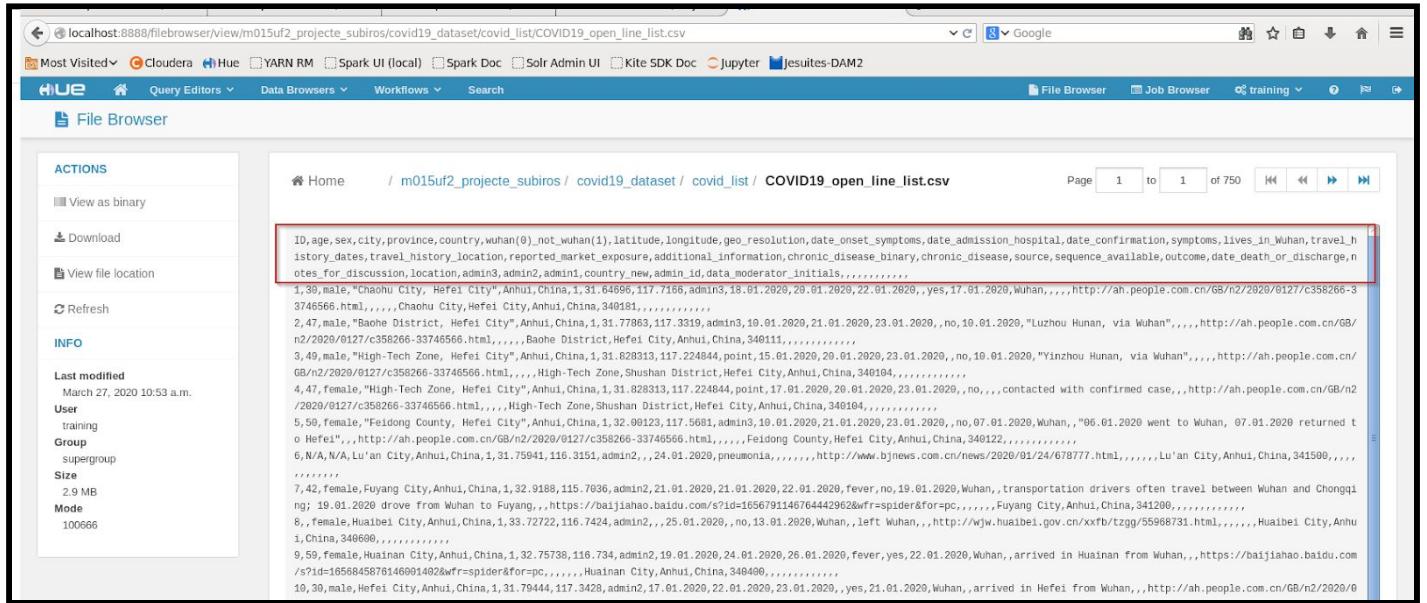
There are currently no logs to visualize.

Nom i Cognoms

Arnaud Subirós Puigarnau

Data

21-04-2020

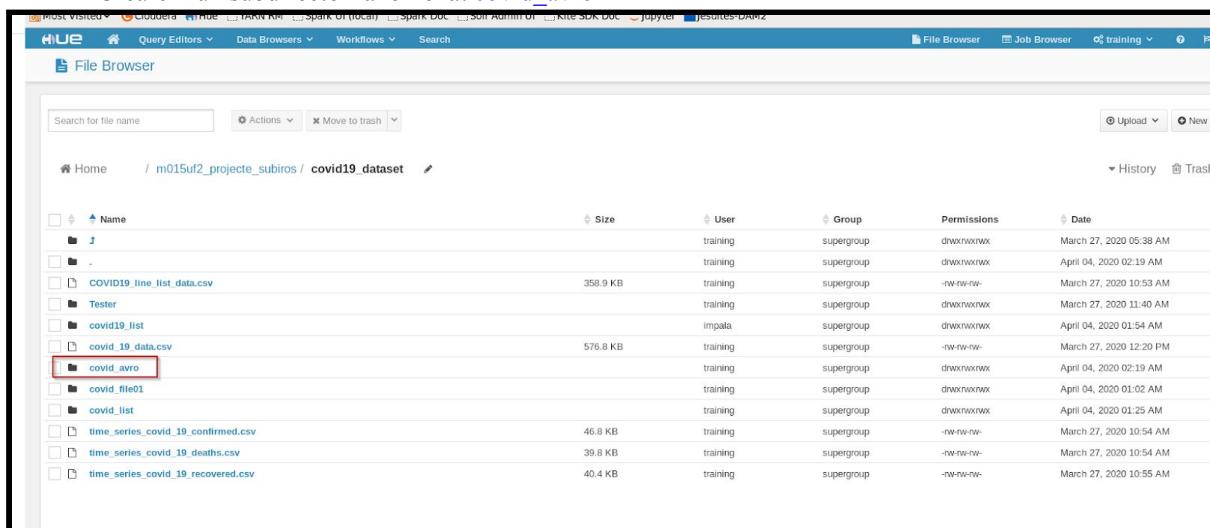


The screenshot shows a web browser window with the URL `localhost:8888/filebrowser/view/m015uf2_projecte_subiros/covid19_dataset/covid_list/COVID19_open_line_list.csv`. The browser's address bar and various tabs are visible at the top. Below the address bar, the Hue interface has a navigation bar with links like 'File Browser', 'Job Browser', and 'training'. The main content area is titled 'COVID19_open_line_list.csv' and displays a large block of CSV data. The data includes columns such as ID, age, sex, city, province, country, wuhan(0)_not_wuhan(1), latitude, longitude, geo_resolution, date_onset_symptoms, date_admission_hospital, date_confirmation, symptoms, lives_in_Wuhan, travel_history_dates, travel_history_location, reported_market_exposure, additional_information, chronic_disease_binary, chronic_disease, source, sequence_available, outcome, date_death_or_discharge, notes_for_discussion, location, admin3, admin2, admin1, country_new, admin_id, data_moderator_initials, etc. The data is presented as a table with many rows, some of which are highlighted with red boxes. The bottom of the page shows pagination controls.

Nom i Cognoms	Data
Arnau Subirós Puigarnau	21-04-2020

● FORMAT : AVRO

- Ens basarem en l' arxiu : [covid_19_data.csv](#)
- Crearem un subdirectorí anomenat **covid_avro**



Name	Size	User	Group	Permissions	Date
COVID19_line_list_data.csv	358.9 KB	training	supergroup	drwxrwxrwx	March 27, 2020 05:38 AM
Tester		training	supergroup	drwxrwxrwx	April 04, 2020 02:19 AM
covid19_list		training	supergroup	drwxrwxrwx	March 27, 2020 10:53 AM
covid_19_data.csv	576.8 KB	training	supergroup	drwxrwxrwx	March 27, 2020 11:40 AM
covid_avro		training	supergroup	drwxrwxrwx	April 04, 2020 01:54 AM
covid_file01		training	supergroup	drwxrwxrwx	April 04, 2020 02:19 AM
covid_llist		training	supergroup	drwxrwxrwx	April 04, 2020 01:02 AM
time_series_covid_19_confirmed.csv	46.8 KB	training	supergroup	drwxrwxrwx	April 04, 2020 01:25 AM
time_series_covid_19_deaths.csv	39.8 KB	training	supergroup	drwxrwxrwx	March 27, 2020 10:54 AM
time_series_covid_19_recovered.csv	40.4 KB	training	supergroup	drwxrwxrwx	March 27, 2020 10:55 AM

- Intento crear una taula en format **avro** em dona error . Intento fer-ho amb 2 opcions:

1. Opció 1

```
CREATE EXTERNAL TABLE database_covid9_avro1 (
    Serial int, ObservationDate String, Province String, State String, Country, Region String, LastUpdate String, Confirmed float, Deaths float, Recovered float)
STORED AS AVRO
LOCATION '/m015uf2_proyecte_subiros/covid19_dataset/covid_avro'
TBLPROPERTIES ('avro.schema.literal' =
```

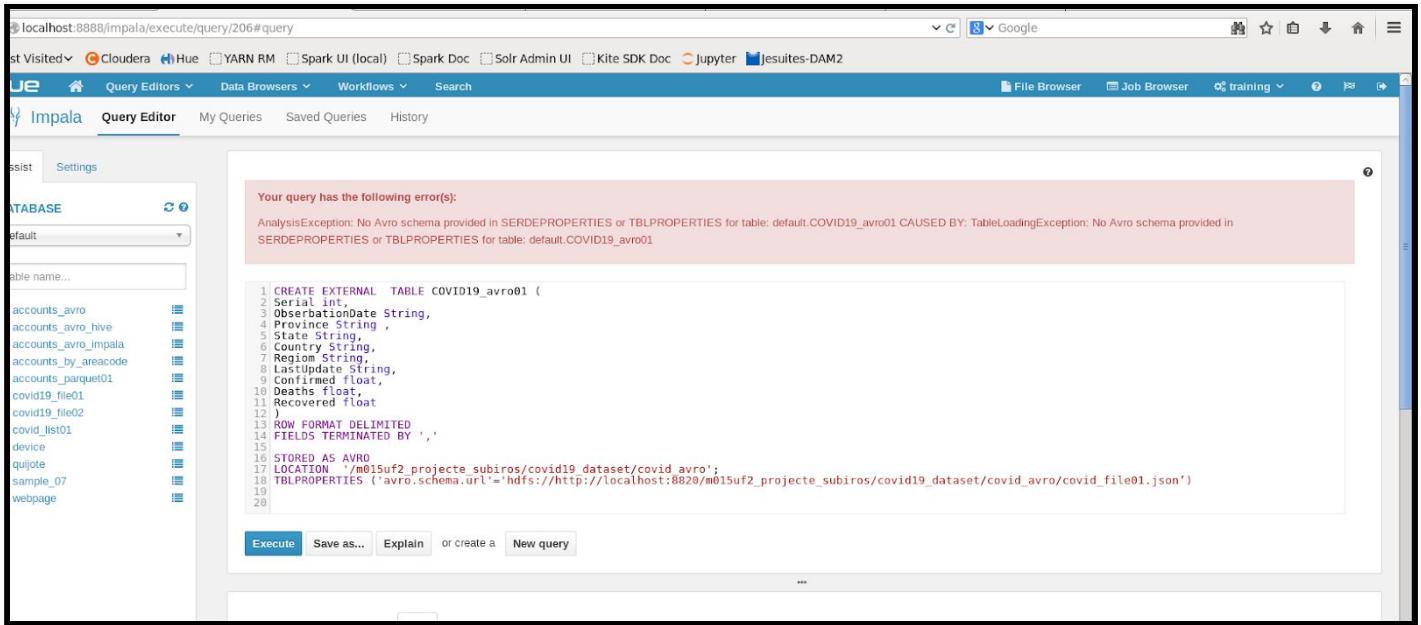
```
'{"name": "database_covid9_avro1",
"type": "record",
"fields": [
    {"name": "Serial", "type": "int"},
    {"name": "ObservationDate", "type": "String"},
    {"name": "State", "type": "string"}
    {"name": "Country", "type": "string"}
    {"name": "Region", "type": "string"}
    {"name": "LastUpdate", "type": "string"}
    {"name": "Confirmed", "type": "float"}
    {"name": "Deaths", "type": "float"}
    {"name": "Recovered", "type": "float"}]
```

Nom i Cognoms	Data
Arnau Subirós Puigarnau	21-04-2020

2. Opció 2

```
CREATE EXTERNAL TABLE COVID19_Avro011 (
Serial int,
ObservationDate String,
Province String ,
State String,
Country String,
Region String,
LastUpdate String,
Confirmed float,
Deaths float,
Recovered float
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS AVRO
LOCATION '/m015uf2_projecte_subiros/covid19_dataset/covid_avro';
TBLPROPERTIES
('avro.schema.url'='hdfs://http://localhost:8820/m015uf2_projecte_subiros/covid19_dataset/covid_avro/covid_file01.json');
```

- Intento provar una nova forma sense èxit
- L'arxiu covid_19_data.csv el converteixo a jason utilitzant aquest enllaç <https://www.convertcsv.com/csv-to-json.htm> després el pujo a HDFS al directori /m015uf2_projecte_subiros/covid19_dataset/covid_avro



The screenshot shows the Impala Query Editor interface. In the main query editor area, there is an error message: "Your query has the following error(s): AnalysisException: No Avro schema provided in SERDEPROPERTIES or TBLPROPERTIES for table: default.COVID19_avro01 CAUSED BY: TableLoadingException: No Avro schema provided in SERDEPROPERTIES or TBLPROPERTIES for table: default.COVID19_avro01". Below the message, the full SQL code is displayed:

```
1 CREATE EXTERNAL TABLE COVID19_avro01 (
2 Serial int,
3 ObservationDate String,
4 Province String ,
5 State String,
6 Country String,
7 Region String,
8 LastUpdate String,
9 Confirmed float,
10 Deaths float,
11 Recovered float
12 )
13 ROW FORMAT DELIMITED
14 FIELDS TERMINATED BY ','
15
16 STORED AS AVRO
17 LOCATION '/m015uf2_projecte_subiros/covid19_dataset/covid_avro';
18 TBLPROPERTIES ('avro.schema.url'='hdfs://http://localhost:8820/m015uf2_projecte_subiros/covid19_dataset/covid_avro/covid_file01.json')
19
20
```

At the bottom of the query editor, there are buttons for "Execute", "Save as...", "Explain", "or create a", and "New query".

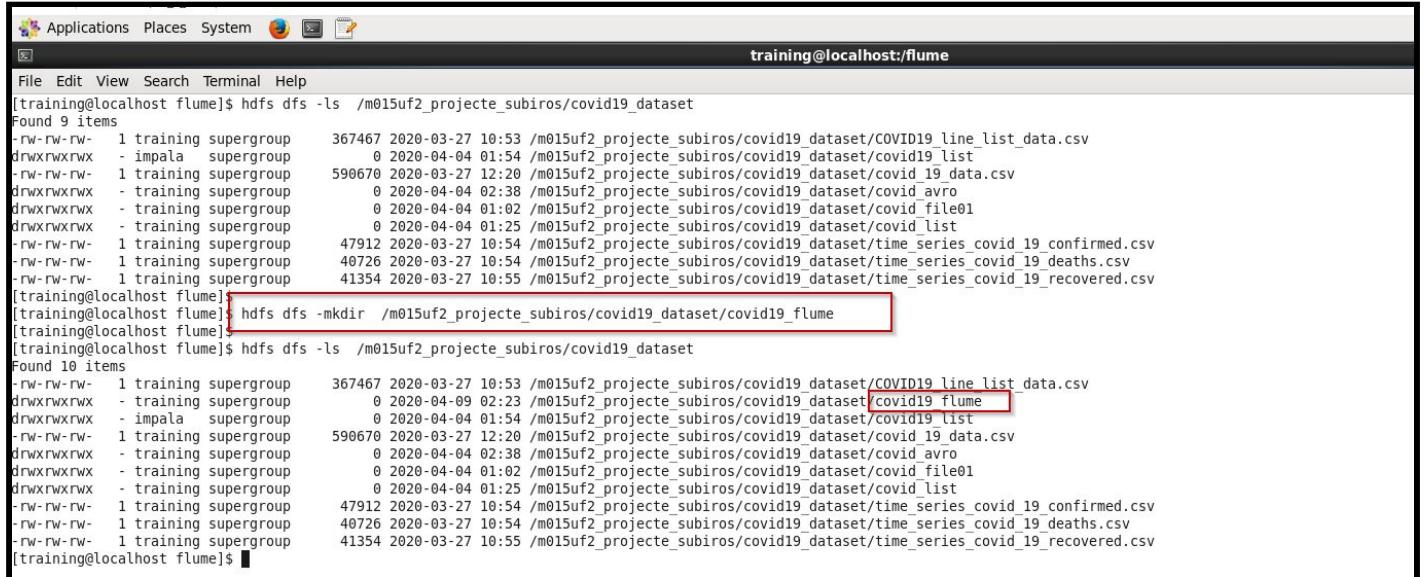
Nom i Cognoms	Data
Arnau Subirós Puigarnau	21-04-2020

2. Recollida de dades en temps real

- **Flume Sources :Spooldir**

- Crearem un directori per la ingestió de dades a HDFS

```
hdfs dfs -mkdir /m015uf2_proyecto_subiros/covid19_dataset/covid19_flume
```

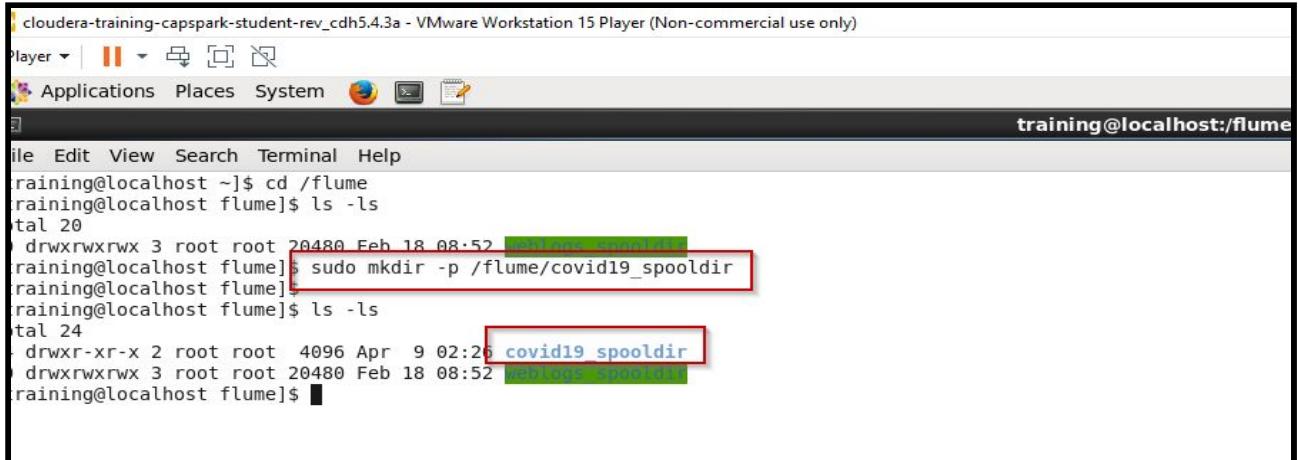


```
File Edit View Search Terminal Help
[training@localhost flume]$ hdfs dfs -ls /m015uf2_proyecto_subiros/covid19_dataset
Found 9 items
-rw-rw-rw- 1 training supergroup 367467 2020-03-27 10:53 /m015uf2_proyecto_subiros/covid19_dataset/COVID19_line_list_data.csv
drwxrwxrwx - impala supergroup 0 2020-04-04 01:54 /m015uf2_proyecto_subiros/covid19_dataset/covid19_list
-rw-rw-rw- 1 training supergroup 590670 2020-03-27 12:20 /m015uf2_proyecto_subiros/covid19_dataset/covid_19_data.csv
drwxrwxrwx - training supergroup 0 2020-04-04 02:38 /m015uf2_proyecto_subiros/covid19_dataset/covid_avro
drwxrwxrwx - training supergroup 0 2020-04-04 01:02 /m015uf2_proyecto_subiros/covid19_dataset/covid_file01
drwxrwxrwx - training supergroup 0 2020-04-04 01:25 /m015uf2_proyecto_subiros/covid19_dataset/covid_list
-rw-rw-rw- 1 training supergroup 47912 2020-03-27 10:54 /m015uf2_proyecto_subiros/covid19_dataset/time_series_covid_19_confirmed.csv
-rw-rw-rw- 1 training supergroup 40726 2020-03-27 10:54 /m015uf2_proyecto_subiros/covid19_dataset/time_series_covid_19_deaths.csv
-rw-rw-rw- 1 training supergroup 41354 2020-03-27 10:55 /m015uf2_proyecto_subiros/covid19_dataset/time_series_covid_19_recovered.csv
[training@localhost flume]$
[training@localhost flume]$ hdfs dfs -mkdir /m015uf2_proyecto_subiros/covid19_dataset/covid19_flume
[training@localhost flume]$
[training@localhost flume]$ hdfs dfs -ls /m015uf2_proyecto_subiros/covid19_dataset
Found 10 items
-rw-rw-rw- 1 training supergroup 367467 2020-03-27 10:53 /m015uf2_proyecto_subiros/covid19_dataset/COVID19_line_list_data.csv
drwxrwxrwx - training supergroup 0 2020-04-09 02:23 /m015uf2_proyecto_subiros/covid19_dataset/covid19_flume
drwxrwxrwx - impala supergroup 0 2020-04-04 01:54 /m015uf2_proyecto_subiros/covid19_dataset/covid19_list
-rw-rw-rw- 1 training supergroup 590670 2020-03-27 12:20 /m015uf2_proyecto_subiros/covid19_dataset/covid_19_data.csv
drwxrwxrwx - training supergroup 0 2020-04-04 02:38 /m015uf2_proyecto_subiros/covid19_dataset/covid_avro
drwxrwxrwx - training supergroup 0 2020-04-04 01:02 /m015uf2_proyecto_subiros/covid19_dataset/covid_file01
drwxrwxrwx - training supergroup 0 2020-04-04 01:25 /m015uf2_proyecto_subiros/covid19_dataset/covid_list
-rw-rw-rw- 1 training supergroup 47912 2020-03-27 10:54 /m015uf2_proyecto_subiros/covid19_dataset/time_series_covid_19_confirmed.csv
-rw-rw-rw- 1 training supergroup 40726 2020-03-27 10:54 /m015uf2_proyecto_subiros/covid19_dataset/time_series_covid_19_deaths.csv
-rw-rw-rw- 1 training supergroup 41354 2020-03-27 10:55 /m015uf2_proyecto_subiros/covid19_dataset/time_series_covid_19_recovered.csv
[training@localhost flume] $
```

- Crearem un directori local(no al HDFS) per el log de sortid del web server

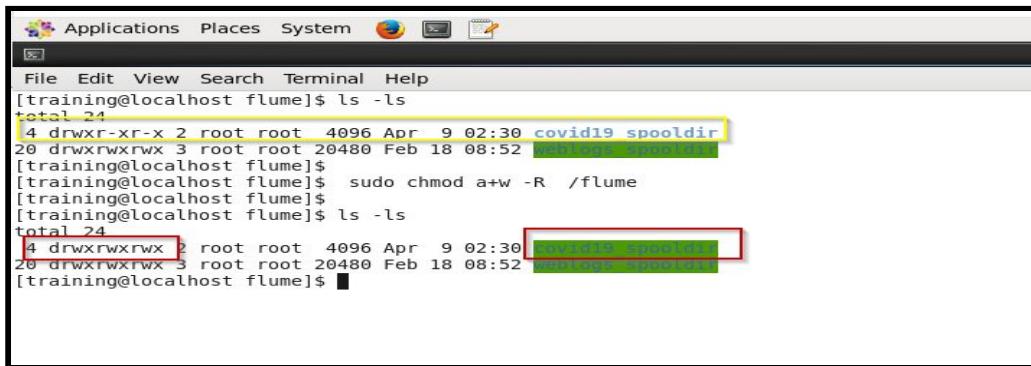
```
/training@localhost flume$ sudo mkdir -p /flume/covid19_spooldir
```

Nom i Cognoms	Data
Arnau Subirós Puigarnau	21-04-2020



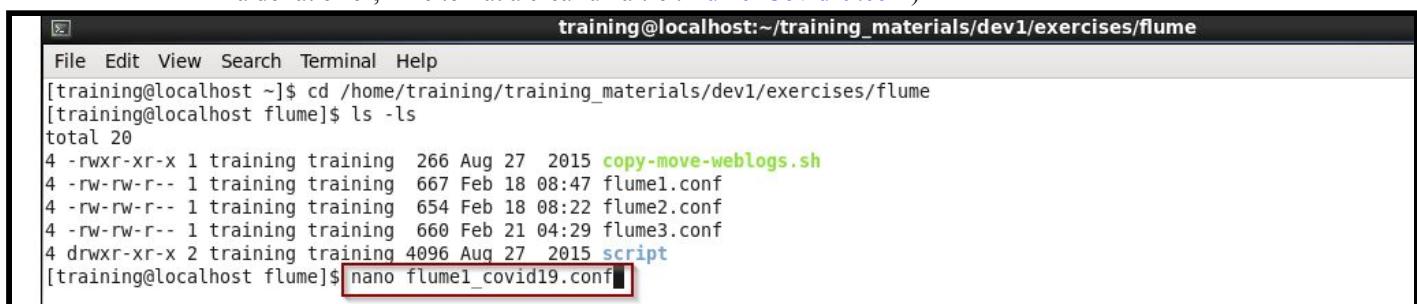
```
cloudera-training-capspark-student-rev_cdh5.4.3a - VMware Workstation 15 Player (Non-commercial use only)
File Edit View Search Terminal Help
[training@localhost ~]$ cd /flume
[training@localhost flume]$ ls -ls
total 20
drwxrwxrwx 3 root root 20480 Feb 18 08:52 covid19_spooldir
[training@localhost flume]$ sudo mkdir -p /flume/covid19_spooldir
[training@localhost flume]$ ls -ls
total 24
drwxr-xr-x 2 root root 4096 Apr 9 02:26 covid19_spooldir
drwxrwxrwx 3 root root 20480 Feb 18 08:52 covid19_spooldir
[training@localhost flume]$
```

- Donem tots els permisos a la carpeta creada
`sudo chmod a+w -R /flume`



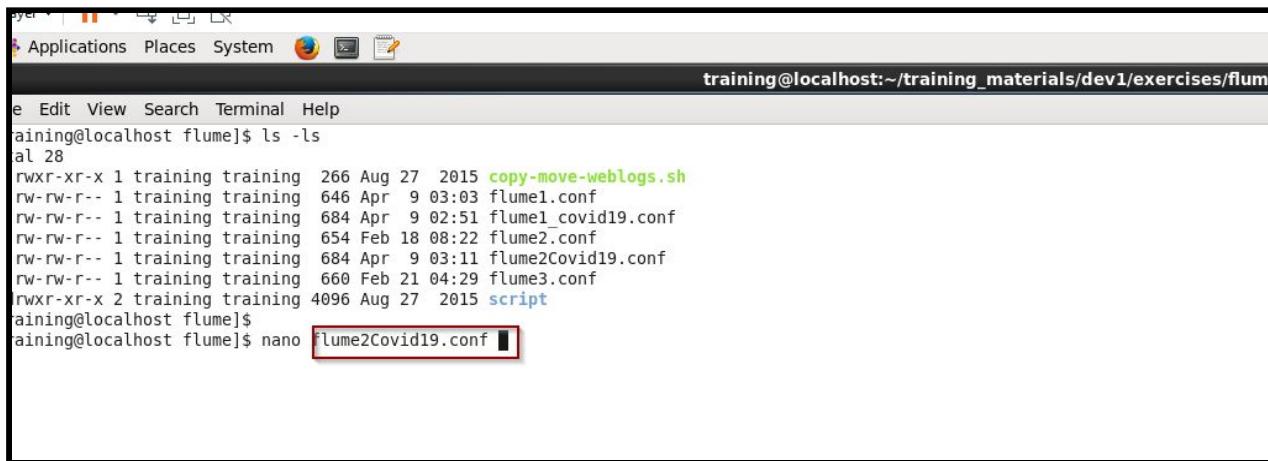
```
File Edit View Search Terminal Help
[training@localhost flume]$ ls -ls
total 24
4 drwxr-xr-x 2 root root 4096 Apr 9 02:30 covid19_spooldir
20 drwxrwxrwx 3 root root 20480 Feb 18 08:52 covid19_spooldir
[training@localhost flume]$ sudo chmod a+w -R /flume
[training@localhost flume]$
[training@localhost flume]$ ls -ls
total 24
4 drwxrwxrwx 2 root root 4096 Apr 9 02:30 covid19_spooldir
20 drwxrwxrwx 3 root root 20480 Feb 18 08:52 covid19_spooldir
[training@localhost flume]$
```

- Configuració de Flume
 - Creació de l'arxiu al mateix lloc a on hem fet les pràctiques LAB09 (l'arxiu l'he creat amb una barra baixa i m'ha donat error, n'he tornat a crear un altre : `flume2Covid19.conf`)



```
File Edit View Search Terminal Help
[training@localhost ~]$ cd /home/training/training_materials/dev1/exercises/flume
[training@localhost flume]$ ls -ls
total 20
4 -rwxr-xr-x 1 training training 266 Aug 27 2015 copy-move-weblogs.sh
4 -rw-rw-r-- 1 training training 667 Feb 18 08:47 flume1.conf
4 -rw-rw-r-- 1 training training 654 Feb 18 08:22 flume2.conf
4 -rw-rw-r-- 1 training training 660 Feb 21 04:29 flume3.conf
4 drwxr-xr-x 2 training training 4096 Aug 27 2015 script
[training@localhost flume]$ nano flume1_covid19.conf
```

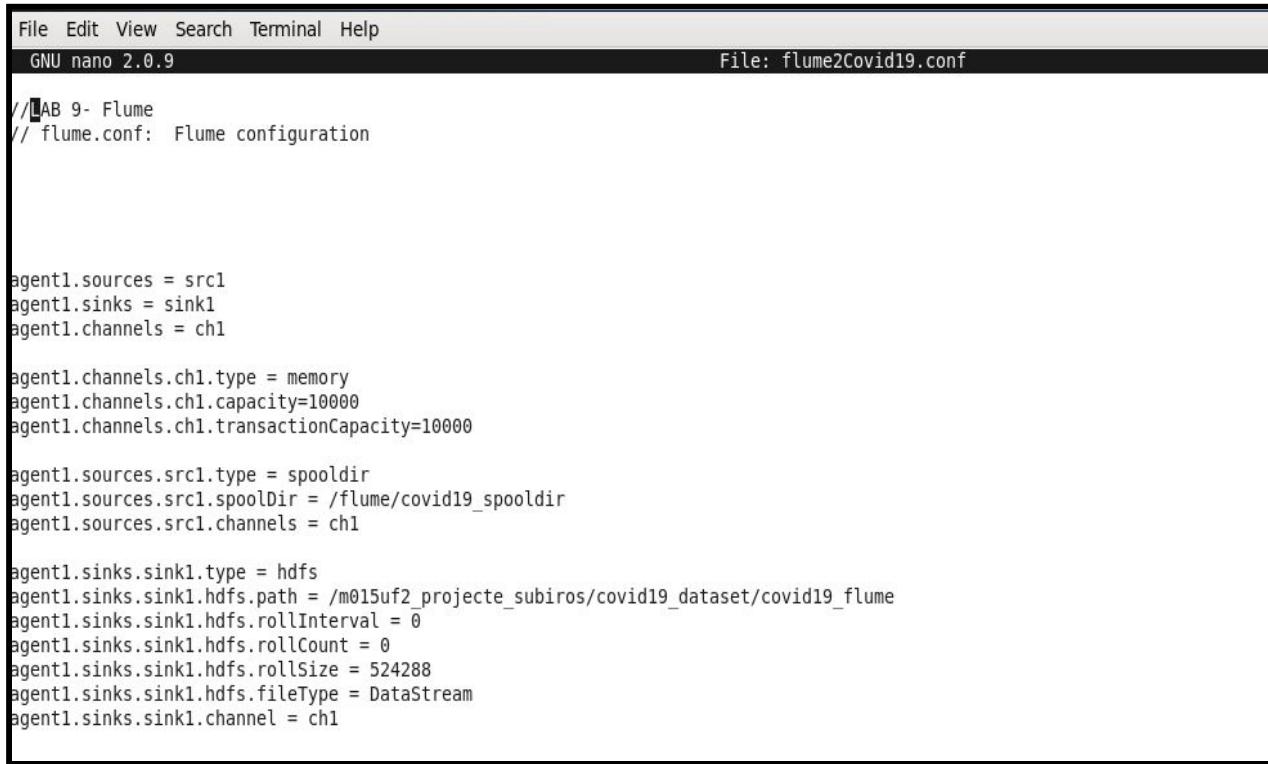
Nom i Cognoms	Data
Arnau Subirós Puigarnau	21-04-2020



```

training@localhost:~/training_materials/dev1/exercises/flume]$ ls -ls
total 28
-rwxr-xr-x 1 training training 266 Aug 27 2015 copy-move-weblogs.sh
-rw-rw-r-- 1 training training 646 Apr  9 03:03 flume1.conf
-rw-rw-r-- 1 training training 684 Apr  9 02:51 flume1_covid19.conf
-rw-rw-r-- 1 training training 654 Feb 18 08:22 flume2.conf
-rw-rw-r-- 1 training training 684 Apr  9 03:11 flume2Covid19.conf
-rw-rw-r-- 1 training training 660 Feb 21 04:29 flume3.conf
-rwxr-xr-x 2 training training 4096 Aug 27 2015 script
training@localhost flume]$ nano flume2Covid19.conf

```



```

File Edit View Search Terminal Help
GNU nano 2.0.9
File: flume2Covid19.conf

//LAB 9- Flume
// flume.conf: Flume configuration

agent1.sources = src1
agent1.sinks = sink1
agent1.channels = ch1

agent1.channels.ch1.type = memory
agent1.channels.ch1.capacity=10000
agent1.channels.ch1.transactionCapacity=10000

agent1.sources.src1.type = spooldir
agent1.sources.src1.spoolDir = /flume/covid19_spooldir
agent1.sources.src1.channels = ch1

agent1.sinks.sink1.type = hdfs
agent1.sinks.sink1.hdfs.path = /m015uf2_proyecte_subiros/covid19_dataset/covid19_flume
agent1.sinks.sink1.hdfs.rollInterval = 0
agent1.sinks.sink1.hdfs.rollCount = 0
agent1.sinks.sink1.hdfs.rollSize = 524288
agent1.sinks.sink1.hdfs.fileType = DataStream
agent1.sinks.sink1.channel = ch1

```

Nom i Cognoms	Data
Arnau Subirós Puigarnau	21-04-2020

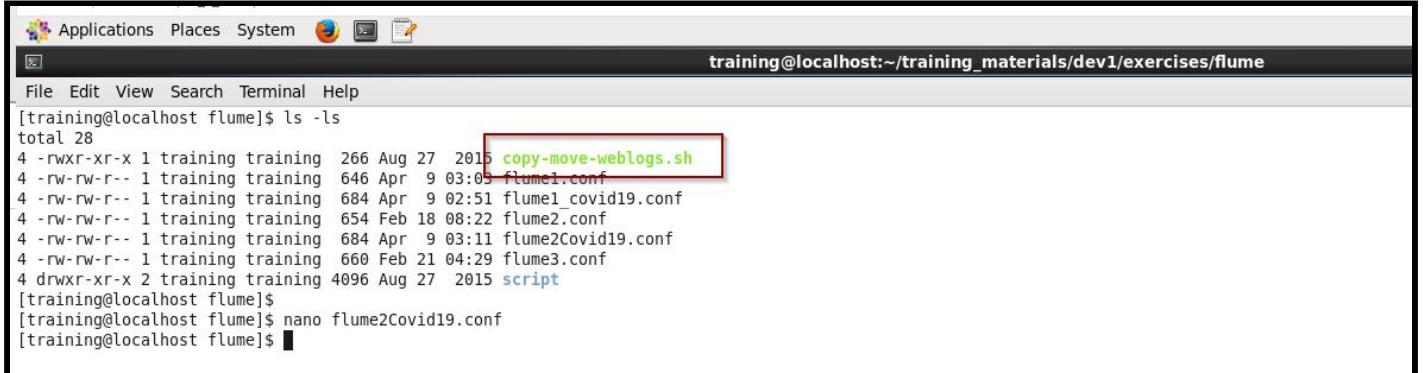
- Desde el directori a on he creat l'arxiu de configuració ,iniciem l'agent Flume

```
training@localhost flume$ flume-ng agent --conf /etc/flume-ng/conf --conf-file ./flume2Covid19.conf --name agent1 -Dflume.root.logger=INFO,console
Info: Sourcing environment configuration script /etc/flume-ng/conf/flume-env.sh
Info: Including Hadoop libraries found via (/usr/bin/hadoop) for HDFS access
Info: Excluding /usr/lib/hadoop/lib/slf4j-api-1.7.5.jar from classpath
Info: Excluding /usr/lib/hadoop/lib/slf4j-log4j12.jar from classpath
Info: Including HBASE libraries found via (/usr/bin/hbase) for HBASE access
Info: Excluding /usr/lib/hbase/bin/../* from classpath
Info: Excluding /usr/lib/hbase/bin/./lib/slf4j-log4j12.jar from classpath
Info: Excluding /usr/lib/hadoop/lib/slf4j-api-1.7.5.jar from classpath
Info: Excluding /usr/lib/hadoop/lib/slf4j-log4j12.jar from classpath
Info: Excluding /usr/lib/hadoop/lib/slf4j-api-1.7.5.jar from classpath
Info: Excluding /usr/lib/hadoop/lib/slf4j-log4j12.jar from classpath
Info: Excluding /usr/lib/zookeeper/lib/slf4j-api-1.7.5.jar from classpath
Info: Excluding /usr/lib/zookeeper/lib/slf4j-log4j12-1.7.5.jar from classpath
Info: Excluding /usr/lib/zookeeper/lib/slf4j-log4j12.jar from classpath
Info: Including Hive libraries found via () for Hive access
+ exec /usr/java/default/bin/java -Xmx500m -Dflume.root.logger=INFO,console -cp '/etc/flume-ng/conf:/usr/lib/flume-ng/lib/*:/etc/hadoop/conf:/usr/lib/hadoop/lib/activation-1.1.jar:/usr/lib/hadoop/lib/apacheds-i18n-2.0.0-M15.jar:/usr/lib/hadoop/lib/apacheds-kerberos-codec-2.0.0-M15.jar:/usr/lib/hadoop/lib/api-asn1-api-1.0.0-M20.jar:/usr/lib/hadoop/lib/api-util-1.0.0-M20.jar:/usr/lib/hadoop/lib/asm-3.2.jar:/usr/lib/hadoop/lib/avro.jar:/usr/lib/hadoop/lib/aws-java-sdk-1.7.4.jar:/usr/lib/hadoop/lib/commons-beanutils-1.7.0.jar:/usr/lib/hadoop/lib/commons-beanutils-core-1.8.0.jar:/usr/lib/hadoop/lib/commons-cli-1.2.jar:/usr/lib/hadoop/lib/commons-codec-1.4.jar:/usr/lib/hadoop/lib/commons-collections-3.2.1.jar:/usr/lib/hadoop/lib/commons-compress-1.4.1.jar:/usr/lib/hadoop/lib/commons-configuration-1.6.jar:/usr/lib/hadoop/lib/commons-digester-1.8.jar:/usr/lib/hadoop/lib/commons-el-1.0.jar:/usr/lib/hadoop/lib/commons-httpclient-3.1.jar:/usr/lib/hadoop/lib/commons-io-2.4.jar:/usr/lib/hadoop/lib/commons-lang-2.6.jar:/usr/lib/hadoop/lib/commons-logging-1.1.3.jar:/usr/lib/hadoop/lib/commons-math3-3.1.1.jar:/usr/lib/hadoop/lib/commons-net-3.1.jar:/usr/lib/hadoop/lib/curator-client-2.7.1.jar:/usr/lib/hadoop/lib/curator-framework-2.7.1.jar:/usr/lib/hadoop/lib/curator-recipes-2.7.1.jar:/usr/lib/hadoop/lib/gson-2.2.4.jar:/usr/lib/hadoop/lib/guava-11.0.2.jar:/usr/lib/hadoop/lib/hamcrest-core-1.3.jar:/usr/lib/hadoop/lib/htrace-core-3.0.4.jar:/usr/lib/hadoop/lib/httpclient-4.2.5.jar:/usr/lib/hadoop/lib/httpcore-4.2.5.jar:/usr/lib/hadoop/lib/hue-plugins-3.1.jar'
```

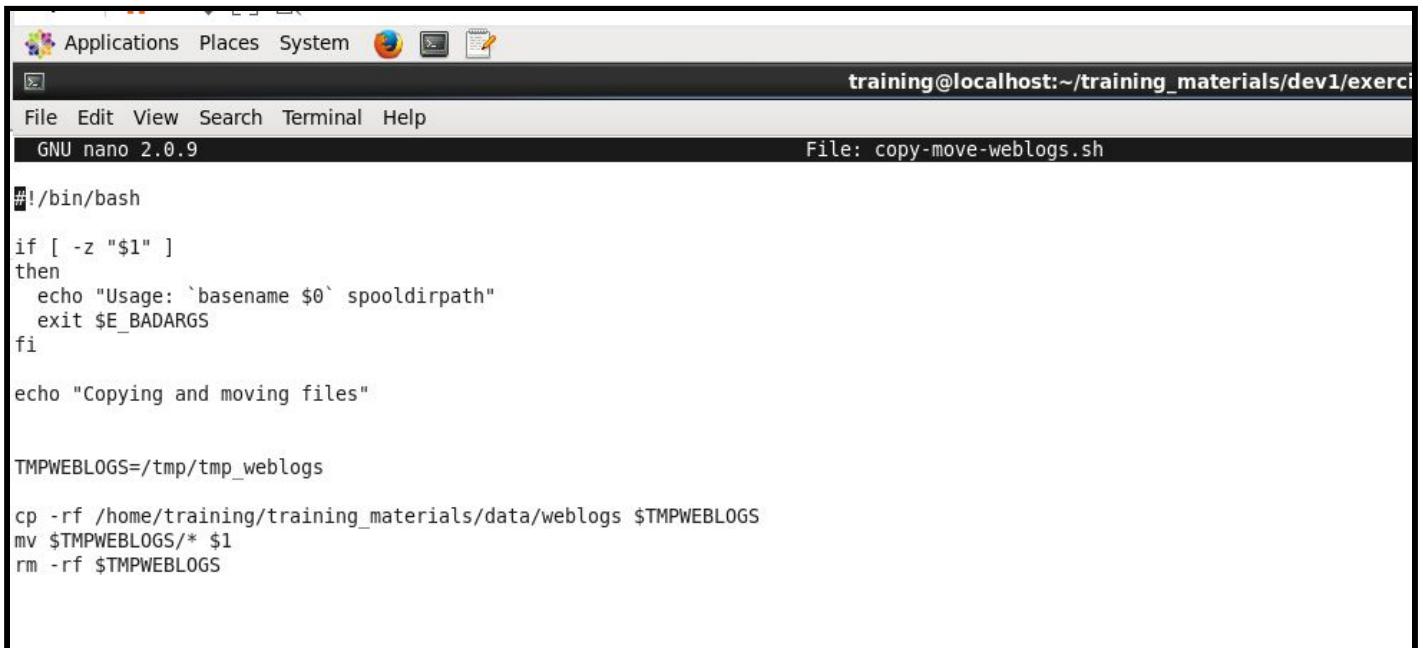
```
020-04-09 03:12:28,405 (com-tile-poller-0) [INFO - org.apache.flume.node.Application.startAllComponents(Application.java:164)] Starting source s1c1
020-04-09 03:12:28,405 (lifecycleSupervisor-1-0) [INFO - org.apache.flume.source.SpoolDirectorySource.start(SpoolDirectorySource.java:78)] SpoolDirectorySo
rce source starting with directory: /flume/covid19/spooldir
020-04-09 03:12:28,406 (lifecycleSupervisor-1-1) [INFO - org.apache.flume.instrumentation.MonitoredCounterGroup.register(MonitoredCounterGroup.java:120)] M
onitored counter group for type: SINK, name: sink1: Successfully registered new MBean.
020-04-09 03:12:28,406 (lifecycleSupervisor-1-1) [INFO - org.apache.flume.instrumentation.MonitoredCounterGroup.start(MonitoredCounterGroup.java:96)] Compo
nent type: SINK, name: sink1 started
020-04-09 03:12:28,424 (lifecycleSupervisor-1-0) [INFO - org.apache.flume.instrumentation.MonitoredCounterGroup.register(MonitoredCounterGroup.java:120)] M
onitored counter group for type: SOURCE, name: srcl: Successfully registered new MBean.
020-04-09 03:12:28,424 (lifecycleSupervisor-1-0) [INFO - org.apache.flume.instrumentation.MonitoredCounterGroup.start(MonitoredCounterGroup.java:96)] Compo
nent type: SOURCE, name: srcl started
```

Nom i Cognoms	Data
Arnau Subirós Puigarnau	21-04-2020

- *Simular els logs de sortida s d'un Servidor Web mitjançant el script “copy-move-weblogs.sh” de LAB-09 i els mou al directori spooldir*



```
Applications Places System training@localhost:~/training_materials/dev1/exercises/flume
File Edit View Search Terminal Help
[training@localhost flume]$ ls -ls
total 28
4 -rwxr-xr-x 1 training training 266 Aug 27 2015 copy-move-weblogs.sh
4 -rw-rw-r-- 1 training training 646 Apr 9 03:05 flume1.conf
4 -rw-rw-r-- 1 training training 684 Apr 9 02:51 flume1_covid19.conf
4 -rw-rw-r-- 1 training training 654 Feb 18 08:22 flume2.conf
4 -rw-rw-r-- 1 training training 684 Apr 9 03:11 flume2Covid19.conf
4 -rw-rw-r-- 1 training training 660 Feb 21 04:29 flume3.conf
4 drwxr-xr-x 2 training training 4096 Aug 27 2015 script
[training@localhost flume]$
[training@localhost flume]$ nano flume2Covid19.conf
[training@localhost flume]$
```



```
File: copy-move-weblogs.sh
GNU nano 2.0.9

#!/bin/bash

if [ -z "$1" ]
then
  echo "Usage: `basename $0` spooldirpath"
  exit $E_BADARGS
fi

echo "Copying and moving files"

TMPWEBLOGS=/tmp/tmp_weblogs

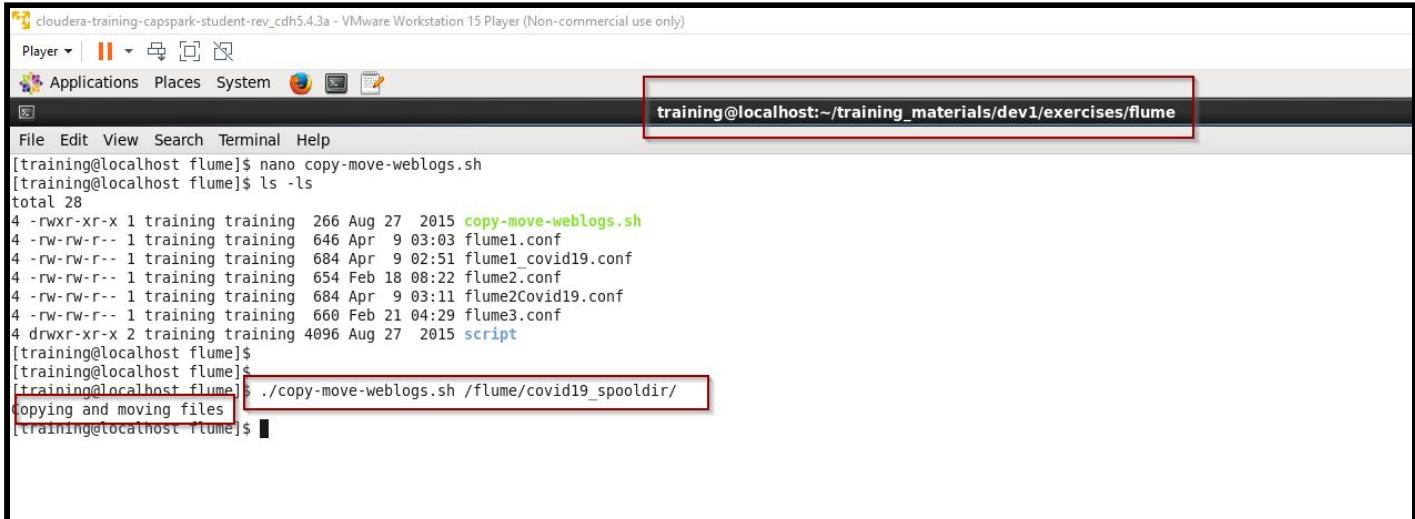
cp -rf /home/training/training_materials/data/weblogs $TMPWEBLOGS
mv $TMPWEBLOGS/* $1
rm -rf $TMPWEBLOGS
```

Nom i Cognoms

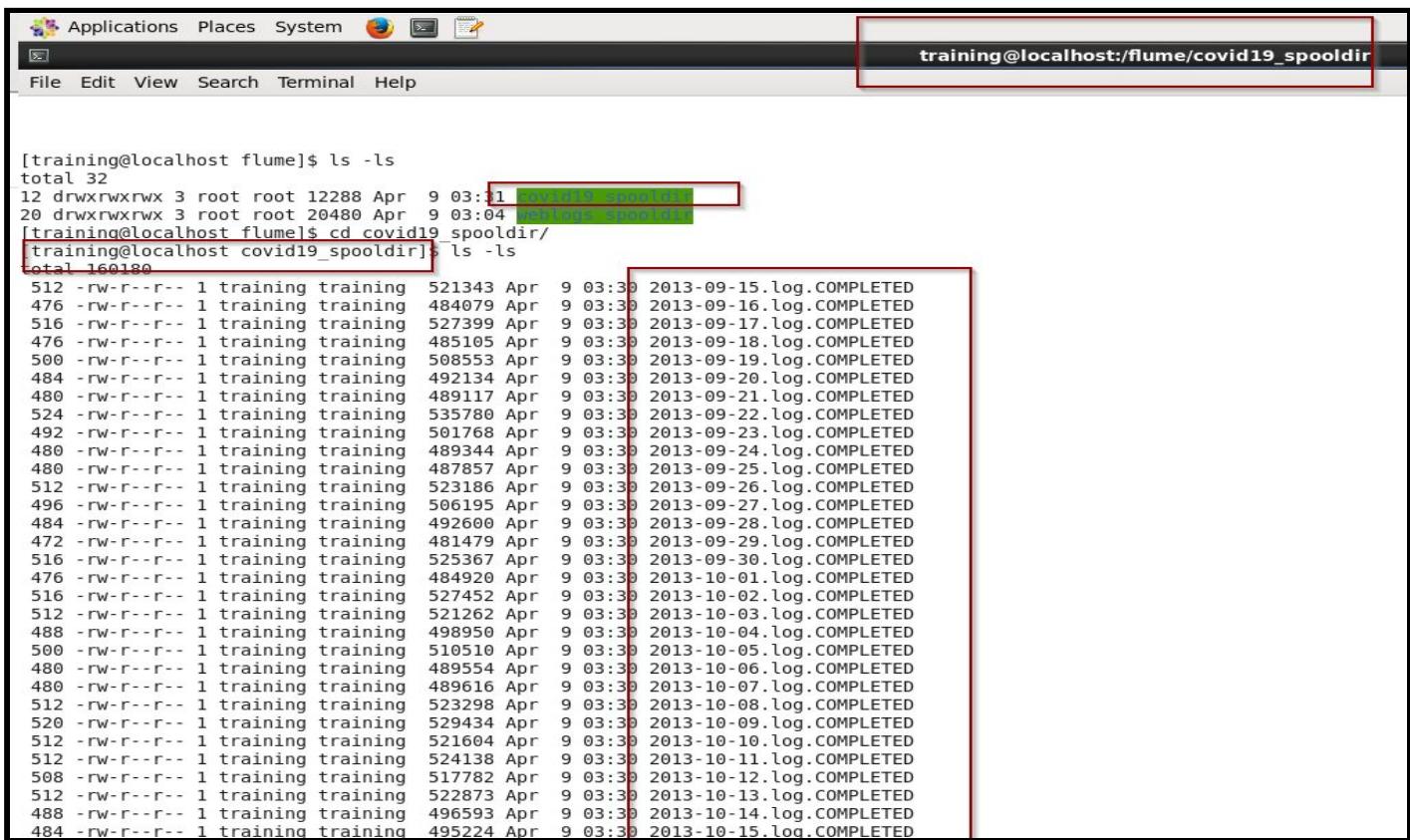
Arnaud Subirós Puigarnau

Data

21-04-2020



```
[training@localhost flume]$ nano copy-move-weblogs.sh
[training@localhost flume]$ ls -ls
total 28
4 -rwxr-xr-x 1 training training 266 Aug 27 2015 copy-move-weblogs.sh
4 -rw-rw-r-- 1 training training 646 Apr 9 03:03 flume1.conf
4 -rw-rw-r-- 1 training training 684 Apr 9 02:51 flume1_covid19.conf
4 -rw-rw-r-- 1 training training 654 Feb 18 08:22 flume2.conf
4 -rw-rw-r-- 1 training training 684 Apr 9 03:11 flume2Covid19.conf
4 -rw-rw-r-- 1 training training 660 Feb 21 04:29 flume3.conf
4 drwxr-xr-x 2 training training 4096 Aug 27 2015 script
[training@localhost flume]$
[training@localhost flume]$ ./copy-move-weblogs.sh /flume/covid19_spooldir/
Copying and moving files
[training@localhost flume]$
```



```
[training@localhost flume]$ ls -ls
total 32
12 drwxrwxrwx 3 root root 12288 Apr 9 03:11 covid19_spooldir
20 drwxrwxrwx 3 root root 20480 Apr 9 03:04 logdata_spooldir
[training@localhost flume]$ cd covid19_spooldir/
[training@localhost covid19_spooldir]$ ls -ls
total 160180
512 -rw-r--r-- 1 training training 521343 Apr 9 03:30 2013-09-15.log.COMPLETED
476 -rw-r--r-- 1 training training 484079 Apr 9 03:30 2013-09-16.log.COMPLETED
516 -rw-r--r-- 1 training training 527399 Apr 9 03:30 2013-09-17.log.COMPLETED
476 -rw-r--r-- 1 training training 485105 Apr 9 03:30 2013-09-18.log.COMPLETED
500 -rw-r--r-- 1 training training 508553 Apr 9 03:30 2013-09-19.log.COMPLETED
484 -rw-r--r-- 1 training training 492134 Apr 9 03:30 2013-09-20.log.COMPLETED
480 -rw-r--r-- 1 training training 489117 Apr 9 03:30 2013-09-21.log.COMPLETED
524 -rw-r--r-- 1 training training 535780 Apr 9 03:30 2013-09-22.log.COMPLETED
492 -rw-r--r-- 1 training training 501768 Apr 9 03:30 2013-09-23.log.COMPLETED
480 -rw-r--r-- 1 training training 489344 Apr 9 03:30 2013-09-24.log.COMPLETED
480 -rw-r--r-- 1 training training 487857 Apr 9 03:30 2013-09-25.log.COMPLETED
512 -rw-r--r-- 1 training training 523186 Apr 9 03:30 2013-09-26.log.COMPLETED
496 -rw-r--r-- 1 training training 506195 Apr 9 03:30 2013-09-27.log.COMPLETED
484 -rw-r--r-- 1 training training 492600 Apr 9 03:30 2013-09-28.log.COMPLETED
472 -rw-r--r-- 1 training training 481479 Apr 9 03:30 2013-09-29.log.COMPLETED
516 -rw-r--r-- 1 training training 525367 Apr 9 03:30 2013-09-30.log.COMPLETED
476 -rw-r--r-- 1 training training 484920 Apr 9 03:30 2013-10-01.log.COMPLETED
516 -rw-r--r-- 1 training copyng 527452 Apr 9 03:30 2013-10-02.log.COMPLETED
512 -rw-r--r-- 1 training training 521262 Apr 9 03:30 2013-10-03.log.COMPLETED
488 -rw-r--r-- 1 training training 498950 Apr 9 03:30 2013-10-04.log.COMPLETED
500 -rw-r--r-- 1 training training 510510 Apr 9 03:30 2013-10-05.log.COMPLETED
480 -rw-r--r-- 1 training training 489554 Apr 9 03:30 2013-10-06.log.COMPLETED
480 -rw-r--r-- 1 training training 489616 Apr 9 03:30 2013-10-07.log.COMPLETED
512 -rw-r--r-- 1 training training 523298 Apr 9 03:30 2013-10-08.log.COMPLETED
520 -rw-r--r-- 1 training training 529434 Apr 9 03:30 2013-10-09.log.COMPLETED
512 -rw-r--r-- 1 training training 521604 Apr 9 03:30 2013-10-10.log.COMPLETED
512 -rw-r--r-- 1 training training 524138 Apr 9 03:30 2013-10-11.log.COMPLETED
508 -rw-r--r-- 1 training training 517782 Apr 9 03:30 2013-10-12.log.COMPLETED
512 -rw-r--r-- 1 training training 522873 Apr 9 03:30 2013-10-13.log.COMPLETED
488 -rw-r--r-- 1 training training 496593 Apr 9 03:30 2013-10-14.log.COMPLETED
484 -rw-r--r-- 1 training training 495224 Apr 9 03:30 2013-10-15.log.COMPLETED
```

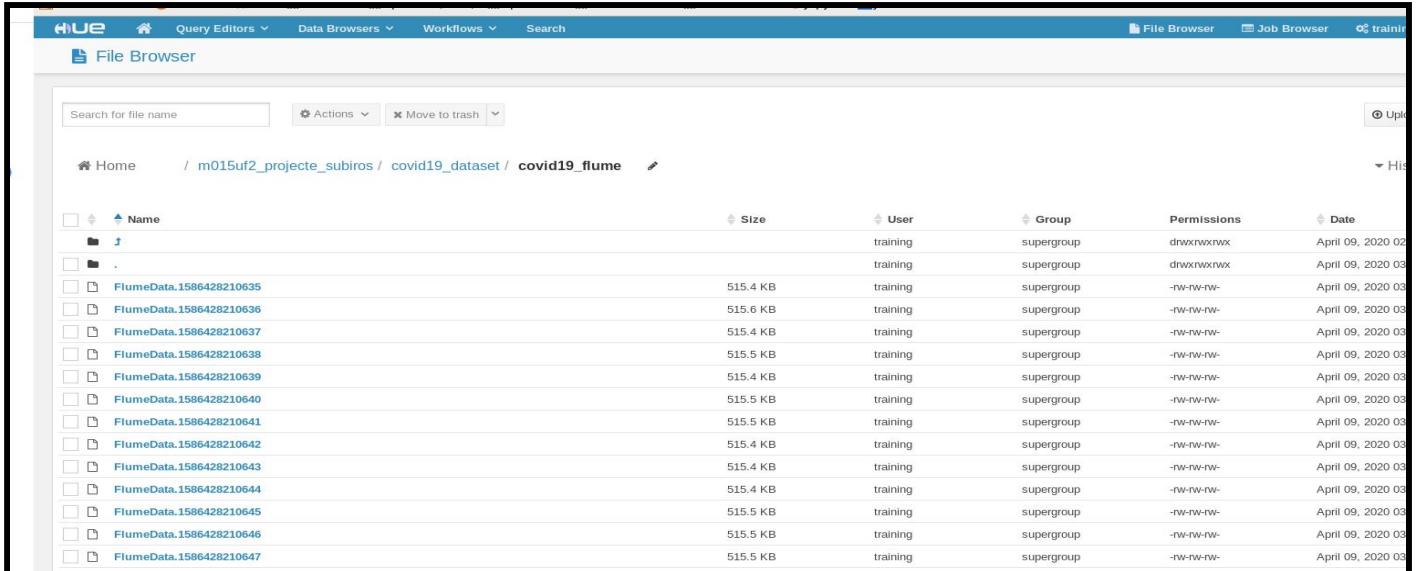
Nom i Cognoms

Arnaud Subirós Puigarnau

Data

21-04-2020

```
[training@localhost covid19_spooldir]$ ls
2013-09-16.log.COMPLETED 2013-10-09.log.COMPLETED 2013-10-31.log.COMPLETED 2013-11-23.log.COMPLETED 2013-12-16.log.COMPLETED 2014-01-08.log.COMPLETED 2014-01-31.log.COMPLETED 2014-02-23.log.COMPLETED
2013-09-17.log.COMPLETED 2013-10-10.log.COMPLETED 2013-11-01.log.COMPLETED 2013-11-24.log.COMPLETED 2013-12-17.log.COMPLETED 2014-01-09.log.COMPLETED 2014-02-01.log.COMPLETED 2014-02-24.log.COMPLETED
2013-09-18.log.COMPLETED 2013-10-11.log.COMPLETED 2013-11-02.log.COMPLETED 2013-11-25.log.COMPLETED 2013-12-18.log.COMPLETED 2014-01-10.log.COMPLETED 2014-02-02.log.COMPLETED 2014-02-25.log.COMPLETED
2013-09-19.log.COMPLETED 2013-10-12.log.COMPLETED 2013-11-03.log.COMPLETED 2013-11-26.log.COMPLETED 2013-12-19.log.COMPLETED 2014-01-11.log.COMPLETED 2014-02-03.log.COMPLETED 2014-02-26.log.COMPLETED
2013-09-20.log.COMPLETED 2013-10-13.log.COMPLETED 2013-11-04.log.COMPLETED 2013-11-27.log.COMPLETED 2013-12-20.log.COMPLETED 2014-01-12.log.COMPLETED 2014-02-04.log.COMPLETED 2014-02-27.log.COMPLETED
2013-09-21.log.COMPLETED 2013-10-14.log.COMPLETED 2013-11-05.log.COMPLETED 2013-11-28.log.COMPLETED 2013-12-21.log.COMPLETED 2014-01-13.log.COMPLETED 2014-02-05.log.COMPLETED 2014-02-28.log.COMPLETED
2013-09-22.log.COMPLETED 2013-10-15.log.COMPLETED 2013-11-06.log.COMPLETED 2013-11-29.log.COMPLETED 2013-12-22.log.COMPLETED 2014-01-14.log.COMPLETED 2014-02-06.log.COMPLETED 2014-02-31.log.COMPLETED
2013-09-23.log.COMPLETED 2013-10-16.log.COMPLETED 2013-11-07.log.COMPLETED 2013-11-30.log.COMPLETED 2013-12-23.log.COMPLETED 2014-01-15.log.COMPLETED 2014-02-07.log.COMPLETED 2014-03-02.log.COMPLETED
2013-09-24.log.COMPLETED 2013-10-17.log.COMPLETED 2013-11-08.log.COMPLETED 2013-12-01.log.COMPLETED 2013-12-24.log.COMPLETED 2014-01-16.log.COMPLETED 2014-02-08.log.COMPLETED 2014-03-03.log.COMPLETED
2013-09-25.log.COMPLETED 2013-10-18.log.COMPLETED 2013-11-09.log.COMPLETED 2013-12-03.log.COMPLETED 2013-12-26.log.COMPLETED 2014-01-17.log.COMPLETED 2014-02-09.log.COMPLETED 2014-03-04.log.COMPLETED
2013-09-26.log.COMPLETED 2013-10-19.log.COMPLETED 2013-11-10.log.COMPLETED 2013-12-04.log.COMPLETED 2013-12-27.log.COMPLETED 2014-01-18.log.COMPLETED 2014-02-10.log.COMPLETED 2014-03-05.log.COMPLETED
2013-09-27.log.COMPLETED 2013-10-20.log.COMPLETED 2013-11-11.log.COMPLETED 2013-12-05.log.COMPLETED 2013-12-28.log.COMPLETED 2014-01-19.log.COMPLETED 2014-02-11.log.COMPLETED 2014-03-06.log.COMPLETED
2013-09-28.log.COMPLETED 2013-10-21.log.COMPLETED 2013-11-12.log.COMPLETED 2013-12-06.log.COMPLETED 2013-12-29.log.COMPLETED 2014-01-20.log.COMPLETED 2014-02-12.log.COMPLETED 2014-03-07.log.COMPLETED
2013-09-29.log.COMPLETED 2013-10-22.log.COMPLETED 2013-11-13.log.COMPLETED 2013-12-07.log.COMPLETED 2013-12-30.log.COMPLETED 2014-01-21.log.COMPLETED 2014-02-13.log.COMPLETED 2014-03-08.log.COMPLETED
2013-09-30.log.COMPLETED 2013-10-23.log.COMPLETED 2013-11-14.log.COMPLETED 2013-12-08.log.COMPLETED 2013-12-31.log.COMPLETED 2014-01-22.log.COMPLETED 2014-02-14.log.COMPLETED 2014-03-09.log.COMPLETED
2013-10-01.log.COMPLETED 2013-10-24.log.COMPLETED 2013-11-09.log.COMPLETED 2013-12-09.log.COMPLETED 2014-01-01.log.COMPLETED 2014-01-24.log.COMPLETED 2014-02-16.log.COMPLETED 2014-03-11.log.COMPLETED
2013-10-02.log.COMPLETED 2013-10-25.log.COMPLETED 2013-11-17.log.COMPLETED 2013-12-10.log.COMPLETED 2014-01-02.log.COMPLETED 2014-01-25.log.COMPLETED 2014-02-17.log.COMPLETED 2014-03-12.log.COMPLETED
2013-10-03.log.COMPLETED 2013-10-26.log.COMPLETED 2013-11-18.log.COMPLETED 2013-12-11.log.COMPLETED 2014-01-03.log.COMPLETED 2014-01-26.log.COMPLETED 2014-02-18.log.COMPLETED 2014-03-13.log.COMPLETED
2013-10-04.log.COMPLETED 2013-10-27.log.COMPLETED 2013-11-19.log.COMPLETED 2013-12-12.log.COMPLETED 2014-01-04.log.COMPLETED 2014-01-27.log.COMPLETED 2014-02-19.log.COMPLETED 2014-03-14.log.COMPLETED
2013-10-05.log.COMPLETED 2013-10-28.log.COMPLETED 2013-11-20.log.COMPLETED 2013-12-13.log.COMPLETED 2014-01-05.log.COMPLETED 2014-01-28.log.COMPLETED 2014-02-20.log.COMPLETED 2014-03-15.log.COMPLETED
2013-10-06.log.COMPLETED 2013-10-29.log.COMPLETED 2013-11-21.log.COMPLETED 2013-12-14.log.COMPLETED 2014-01-06.log.COMPLETED 2014-01-29.log.COMPLETED 2014-02-21.log.COMPLETED 2014-03-20.log.COMPLETED
2013-10-07.log.COMPLETED 2013-10-30.log.COMPLETED 2013-11-22.log.COMPLETED 2013-12-15.log.COMPLETED 2014-01-07.log.COMPLETED 2014-01-30.log.COMPLETED 2014-02-22.log.COMPLETED 2014-03-21.log.COMPLETED
[training@localhost covid19_spooldir]$ hdfs dfs -ls /m015uf2/projecte_subiros/covid19_dataset/covid19_flume
Found 311 items
-rw-rw-rw- 1 training supergroup 527789 2028-04-09 03:38 /m015uf2/projecte_subiros/covid19_dataset/covid19_flume/FlumeData.1586428210635
-rw-rw-rw- 1 training supergroup 527943 2028-04-09 03:38 /m015uf2/projecte_subiros/covid19_dataset/covid19_flume/FlumeData.1586428210636
-rw-rw-rw- 1 training supergroup 527798 2028-04-09 03:38 /m015uf2/projecte_subiros/covid19_dataset/covid19_flume/FlumeData.1586428210637
-rw-rw-rw- 1 training supergroup 527782 2028-04-09 03:38 /m015uf2/projecte_subiros/covid19_dataset/covid19_flume/FlumeData.1586428210638
-rw-rw-rw- 1 training supergroup 527787 2028-04-09 03:38 /m015uf2/projecte_subiros/covid19_dataset/covid19_flume/FlumeData.1586428210639
-rw-rw-rw- 1 training supergroup 527918 2028-04-09 03:38 /m015uf2/projecte_subiros/covid19_dataset/covid19_flume/FlumeData.1586428210640
-rw-rw-rw- 1 training supergroup 527867 2028-04-09 03:38 /m015uf2/projecte_subiros/covid19_dataset/covid19_flume/FlumeData.1586428210641
-rw-rw-rw- 1 training supergroup 527783 2028-04-09 03:38 /m015uf2/projecte_subiros/covid19_dataset/covid19_flume/FlumeData.1586428210642
-rw-rw-rw- 1 training supergroup 527889 2028-04-09 03:38 /m015uf2/projecte_subiros/covid19_dataset/covid19_flume/FlumeData.1586428210643
-rw-rw-rw- 1 training supergroup 527880 2028-04-09 03:38 /m015uf2/projecte_subiros/covid19_dataset/covid19_flume/FlumeData.1586428210644
-rw-rw-rw- 1 training supergroup 527872 2028-04-09 03:38 /m015uf2/projecte_subiros/covid19_dataset/covid19_flume/FlumeData.1586428210645
-rw-rw-rw- 1 training supergroup 527915 2028-04-09 03:38 /m015uf2/projecte_subiros/covid19_dataset/covid19_flume/FlumeData.1586428210646
-rw-rw-rw- 1 training supergroup 527879 2028-04-09 03:38 /m015uf2/projecte_subiros/covid19_dataset/covid19_flume/FlumeData.1586428210647
-rw-rw-rw- 1 training supergroup 527791 2028-04-09 03:38 /m015uf2/projecte_subiros/covid19_dataset/covid19_flume/FlumeData.1586428210648
-rw-rw-rw- 1 training supergroup 527865 2028-04-09 03:38 /m015uf2/projecte_subiros/covid19_dataset/covid19_flume/FlumeData.1586428210649
-rw-rw-rw- 1 training supergroup 527863 2028-04-09 03:38 /m015uf2/projecte_subiros/covid19_dataset/covid19_flume/FlumeData.1586428210650
-rw-rw-rw- 1 training supergroup 527794 2028-04-09 03:38 /m015uf2/projecte_subiros/covid19_dataset/covid19_flume/FlumeData.1586428210651
-rw-rw-rw- 1 training supergroup 527787 2028-04-09 03:38 /m015uf2/projecte_subiros/covid19_dataset/covid19_flume/FlumeData.1586428210652
-rw-rw-rw- 1 training supergroup 527856 2028-04-09 03:38 /m015uf2/projecte_subiros/covid19_dataset/covid19_flume/FlumeData.1586428210653
-rw-rw-rw- 1 training supergroup 527927 2028-04-09 03:38 /m015uf2/projecte_subiros/covid19_dataset/covid19_flume/FlumeData.1586428210654
```



The screenshot shows the HUE File Browser interface. The top navigation bar includes Home, Query Editors, Data Browsers, Workflows, and Search. Below the navigation is a search bar labeled "File Browser". The main area displays a file tree under the path "/m015uf2/projecte_subiros/covid19_dataset/covid19_flume". The tree structure shows numerous files named "FlumeData.XXXXXXX" where XXXXXX represents a unique identifier. Each file entry includes columns for Name, Size, User, Group, Permissions, and Date. All files are owned by "training" and belong to the "supergroup". The permissions are consistently set to "drwxrwxrwx" (read, write, execute for all). The date for all files is "April 09, 2020 02:45:45". The file sizes vary slightly between 515.4 KB and 515.5 KB.

Name	Size	User	Group	Permissions	Date
.	515.4 KB	training	supergroup	drwxrwxrwx	April 09, 2020 02:45:45
FlumeData.1586428210635	515.6 KB	training	supergroup	-rw-rw-rw-	April 09, 2020 03:00:00
FlumeData.1586428210636	515.4 KB	training	supergroup	-rw-rw-rw-	April 09, 2020 03:00:00
FlumeData.1586428210637	515.5 KB	training	supergroup	-rw-rw-rw-	April 09, 2020 03:00:00
FlumeData.1586428210638	515.4 KB	training	supergroup	-rw-rw-rw-	April 09, 2020 03:00:00
FlumeData.1586428210639	515.4 KB	training	supergroup	-rw-rw-rw-	April 09, 2020 03:00:00
FlumeData.1586428210640	515.5 KB	training	supergroup	-rw-rw-rw-	April 09, 2020 03:00:00
FlumeData.1586428210641	515.5 KB	training	supergroup	-rw-rw-rw-	April 09, 2020 03:00:00
FlumeData.1586428210642	515.4 KB	training	supergroup	-rw-rw-rw-	April 09, 2020 03:00:00
FlumeData.1586428210643	515.4 KB	training	supergroup	-rw-rw-rw-	April 09, 2020 03:00:00
FlumeData.1586428210644	515.4 KB	training	supergroup	-rw-rw-rw-	April 09, 2020 03:00:00
FlumeData.1586428210645	515.5 KB	training	supergroup	-rw-rw-rw-	April 09, 2020 03:00:00
FlumeData.1586428210646	515.5 KB	training	supergroup	-rw-rw-rw-	April 09, 2020 03:00:00
FlumeData.1586428210647	515.5 KB	training	supergroup	-rw-rw-rw-	April 09, 2020 03:00:00

Nom i Cognoms	Data
Arnau Subirós Puigarnau	21-04-2020

● *Flume Sources :HTTP (tòric)*

Creació d'un directori HDFS

```
$ hdfs dfs -mkdir /m015uf2_projecte_subiros/covid19_dataset/covid19_flume/events_HTTP
$ hdfs dfs -chmod -R 777 /m015uf2_projecte_subiros/covid19_dataset/covid19_flume
$ hdfs dfs -chown -R flume /m015uf2_projecte_subiros/covid19_dataset/covid19_flume
```

- Configuració de Flume
 - Creació de l'arxiu al mateix lloc a on hem creat l'arxiu anterior ([flume3HHTPCovid19.conf](#))

Sources, channels, and sinks are defined per # agent name, in this case 'tier1'.

```
tier1.sources = source1
tier1.channels = channel1
tier1.sinks = sink1
tier1.sources.source1.interceptors = i1 i2
tier1.sources.source1.interceptors.i1.type = host
tier1.sources.source1.interceptors.i1.preserveExisting = false
tier1.sources.source1.interceptors.i1.hostHeader = host
tier1.sources.source1.interceptors.i2.type = timestamp
```

For each source, channel, and sink, set # standard properties.

```
tier1.sources.source1.type = http
tier1.sources.source1.bind = 0.0.0.0
tier1.sources.source1.port = 5140
```

JSONHandler is the default for the httpsource

```
tier1.sources.source1.handler = org.apache.flume.source.http.JSONHandler
tier1.sources.source1.channels = channel1
tier1.channels.channel1.type = memory
tier1.sinks.sink1.type = hdfs
tier1.sinks.sink1.hdfs.path = /m015uf2_projecte_subiros/covid19_dataset/covid19_flume/events_HTTP/%y-%m-%d/%H%M%S
tier1.sinks.sink1.hdfs.filePrefix = event-file-prefix-
tier1.sinks.sink1.hdfs.round = false
tier1.sinks.sink1.channel = channel1
```

Other properties are specific to each type of # source, channel, or sink. In this case, we # specify the capacity of the memory channel.

```
tier1.channels.channel1.capacity = 1000
```

Nom i Cognoms	Data
Arnau Subirós Puigarnau	21-04-2020

- Es necessari crear un Flume Client que pugui enviar els esdeveniments JSON al Flume HTTP

```
[  
 {  
   "headers": {  
     "timestamp": "434324343",  
     "host": "localhost",  
   },  
   "body": "No matter what, this must be a String, not a list or a JSON object", },  
   { ... following events take the same format as the one above ...}  
 ]
```

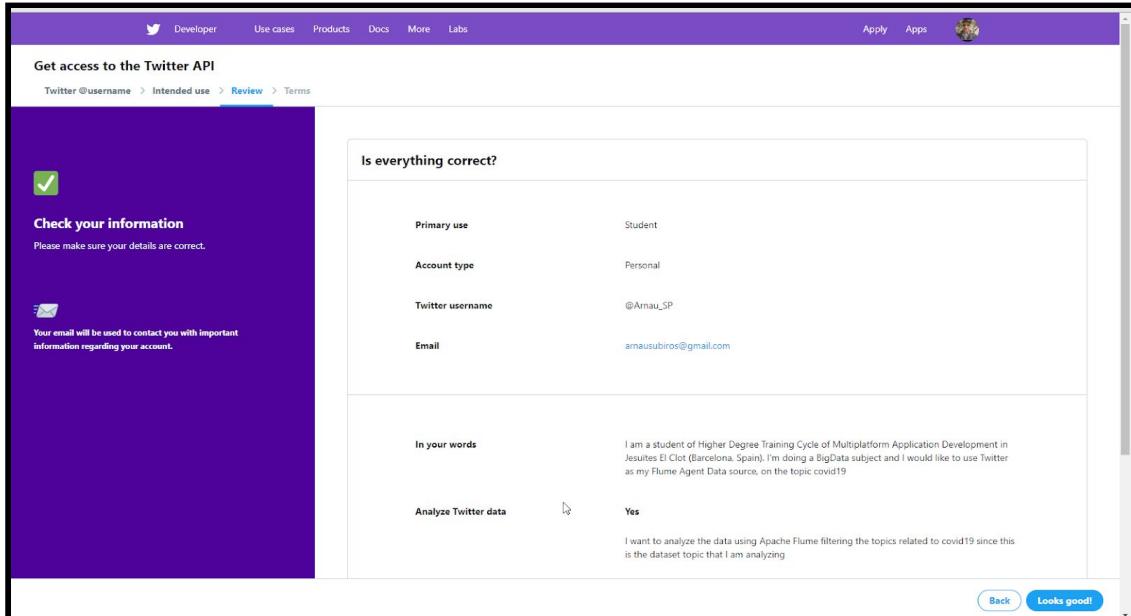
- L'agent HTTP accepta dades en format JSON, s'hauria de provar enviar un cURL amb un formulari com aquest :

```
curl -X POST -H 'Content-Type: application/json; charset=UTF-8' -d '[ {"username":"xrqwrqwryzas","password":"12124sfsfsfas123"} ]' http://yourdomain.com:81/
```

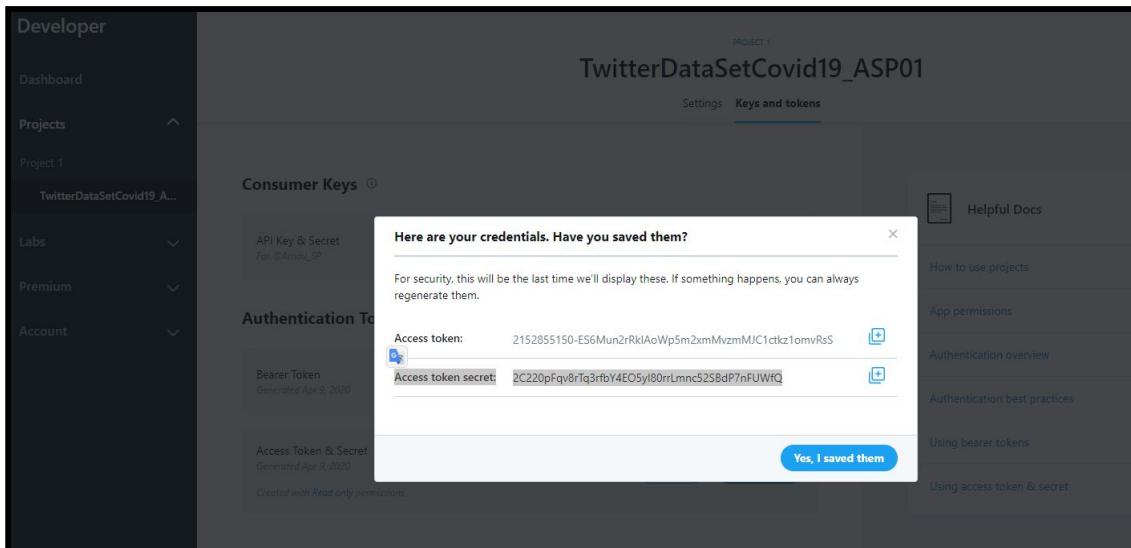
Nom i Cognoms	Data
Arnau Subirós Puigarnau	21-04-2020

● *Flume Sources : Twitter*

- Abans de tot hem d'anar a <https://developer.twitter.com/> i registrant-nos. Ens preguntaran el motiu



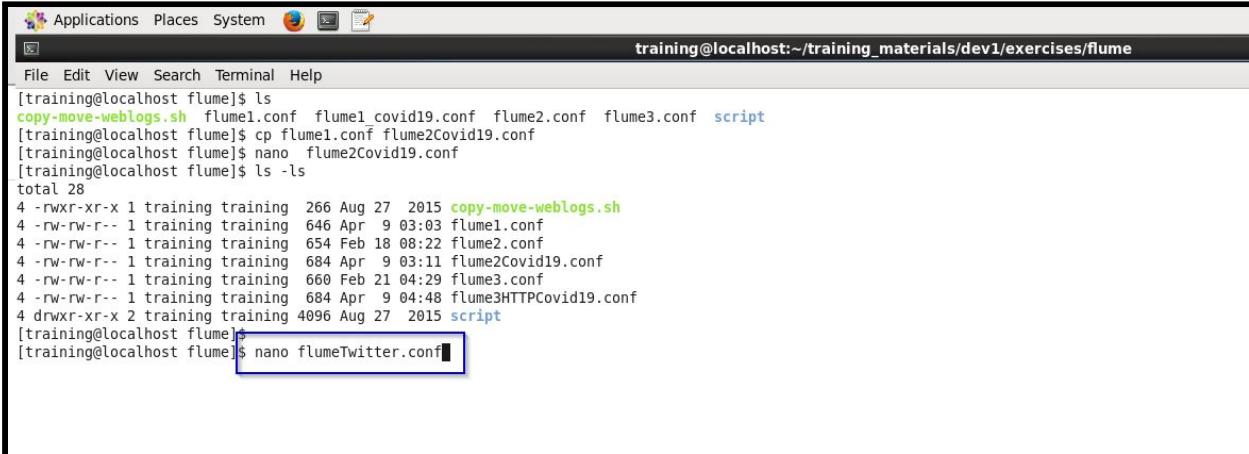
The screenshot shows the 'Get access to the Twitter API' review step. It displays account details: Primary use (Student), Account type (Personal), Twitter username (@Arnau_SP), and Email (arnausubiros@gmail.com). Below this, there's a section for 'In your words' where a message is typed: 'I am a student of Higher Degree Training Cycle of Multiplatform Application Development in Jesuites El Clot (Barcelona, Spain). I'm doing a BigData subject and I would like to use Twitter as my Flume Agent Data source, on the topic covid19'. There are 'Analyze Twitter data' and 'Yes' buttons below the message. At the bottom right are 'Back' and 'Looks good!' buttons.



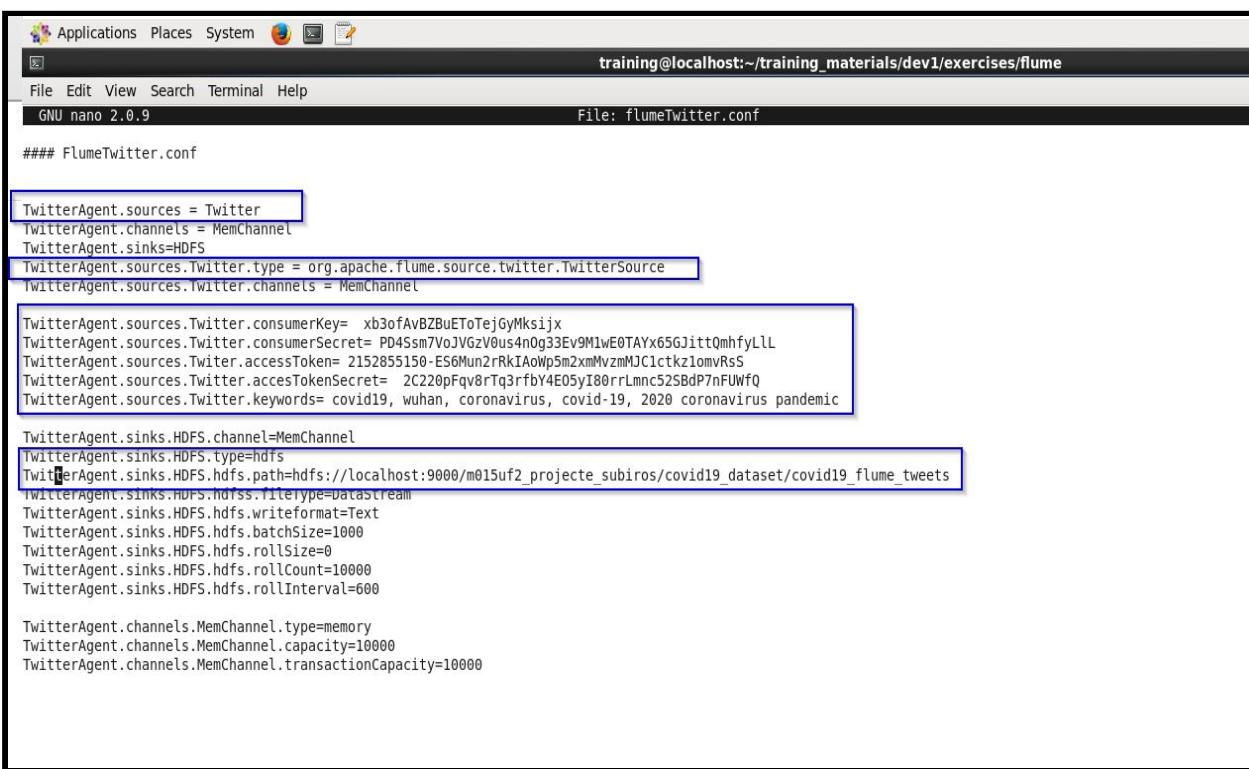
The screenshot shows the Twitter Developer dashboard for the project 'TwitterDataSetCovid19_ASP01'. Under 'Consumer Keys', it lists an 'API Key & Secret' for the user '@Arnau_SP'. A modal window titled 'Here are your credentials. Have you saved them?' shows the 'Access token:' (2152855150-ES6Mun2rRklAoWp5m2xmMvzmMJC1ctkz1omvRsS) and 'Access token secret:' (2C220pFqv8Tq3rbY4EO5yl80rrLmnc52SbdP7nFUWfQ). A 'Yes, I saved them' button is at the bottom right of the modal. To the right, there's a sidebar with 'Helpful Docs' links: How to use projects, App permissions, Authentication overview, Authentication best practices, Using bearer tokens, and Using access token & secret.

Nom i Cognoms	Data
Arnau Subirós Puigarnau	21-04-2020

- Configurem l'arxiu de configuració



```
[training@localhost flume]$ ls
copy-move-weblogs.sh  flume1.conf  flume1 covid19.conf  flume2.conf  flume3.conf  script
[training@localhost flume]$ cp flume1.conf flume2Covid19.conf
[training@localhost flume]$ nano flume2Covid19.conf
[training@localhost flume]$ ls -ls
total 28
4 -rwxr-xr-x 1 training training 266 Aug 27 2015 copy-move-weblogs.sh
4 -rw-rw-r-- 1 training training 646 Apr 9 03:03 flume1.conf
4 -rw-rw-r-- 1 training training 654 Feb 18 08:22 flume2.conf
4 -rw-rw-r-- 1 training training 684 Apr 9 03:11 flume2Covid19.conf
4 -rw-rw-r-- 1 training training 660 Feb 21 04:29 flume3.conf
4 -rw-rw-r-- 1 training training 684 Apr 9 04:48 flume3HTTPCovid19.conf
4 drwxr-xr-x 2 training training 4096 Aug 27 2015 script
[training@localhost flume]$
[training@localhost flume]$ nano flumeTwitter.conf
```



```
#### FlumeTwitter.conf

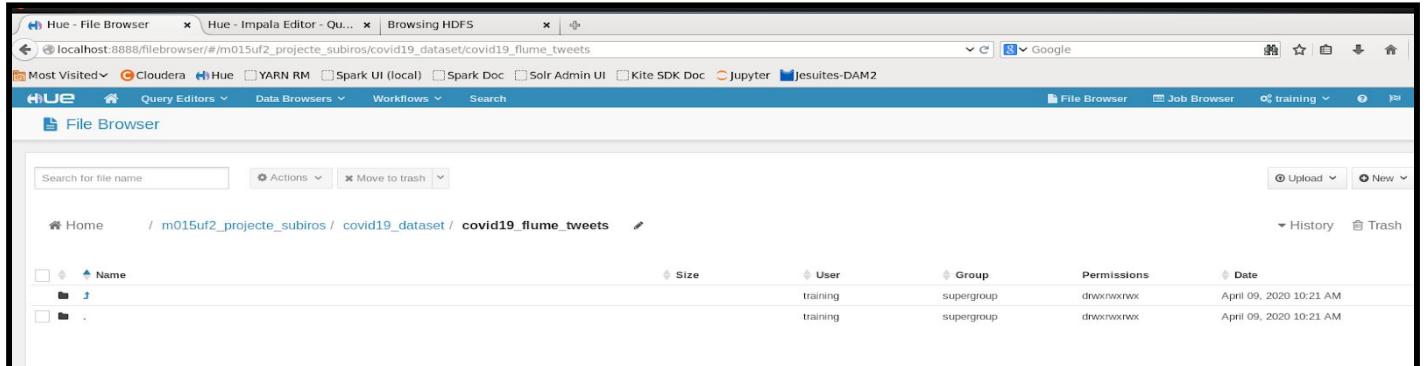
TwitterAgent.sources = Twitter
TwitterAgent.channels = MemChannel
TwitterAgent.sinks=HDFS
TwitterAgent.sources.Twitter.type = org.apache.flume.source.twitter.TwitterSource
TwitterAgent.sources.Twitter.consumerKey= xb3ofAvBZBuEToTejGyMksijx
TwitterAgent.sources.Twitter.consumerSecret= PD45sm7V0JVGzV0us4n0g33Ev9M1wE0TAYx65GJittQmhfyLlL
TwitterAgent.sources.Twitter.accessToken= 2152855150-E56Mun2rRkIAoWp5m2xmMvzmMJClctkz1omvRsS
TwitterAgent.sources.Twitter.accessTokenSecret= 2C220pfqv8rTq3rfbY4E05yI8orrLmnc525BdP7nFUWfQ
TwitterAgent.sources.Twitter.keywords= covid19, wuhan, coronavirus, covid-19, 2020 coronavirus pandemic

TwitterAgent.sinks.HDFS.channel=MemChannel
TwitterAgent.sinks.HDFS.type=hdfs
TwitterAgent.sinks.HDFS.hdfs.path=dfs://localhost:9000/m015uf2_projecte_subiros/covid19_dataset/covid19_flume_tweets
TwitterAgent.sinks.HDFS.hdfs.fileType=DataStream
TwitterAgent.sinks.HDFS.hdfs.writeFormat=Text
TwitterAgent.sinks.HDFS.hdfs.batchSize=1000
TwitterAgent.sinks.HDFS.hdfs.rollSize=0
TwitterAgent.sinks.HDFS.hdfs.rollCount=10000
TwitterAgent.sinks.HDFS.hdfs.rollInterval=600

TwitterAgent.channels.MemChannel.type=memory
TwitterAgent.channels.MemChannel.capacity=10000
TwitterAgent.channels.MemChannel.transactionCapacity=10000
```

Nom i Cognoms	Data
Arnau Subirós Puigarnau	21-04-2020

- Creacio d'un nou directori HDFS anomenat covid19_flume_tweets



- ### • *Iniciem l'agent flume*

```
[training@localhost flume]$ flume-ng agent --conf /etc/flume-ng/conf --conf-file ./flumeTwitter.conf --name TwitterAgent -Dflume.root.logger=INFO,console
```

Nom i Cognoms

Data

Arnau Subirós Puigarnau

21-04-2020

- Ens dona un error en Access Token i Access Token Secret ja que li consta que és null , tot i s'havia afegit.

The screenshot shows the Hue - File Browser - Mozilla Firefox window. The address bar displays the URL: `localhost:8888/filebrowser/#/m015uf2_proyecto_subiros/covid19_dataset/covid19_flume_tweets`. The browser toolbar includes icons for Applications, Places, System, Home, Stop, Back, Forward, and Refresh. The main interface has a top navigation bar with links to Cloudera, Hue, YARN RM, Spark UI (local), Spark Doc, Solr Admin UI, Kite SDK Doc, Jupyter, and Jesuites-DAM2. Below this is a secondary navigation bar with links to File Browser, Job Browser, training, and other Hue components like Query Editors, Data Browsers, Workflows, and Search. The main content area is titled "File Browser" and shows a search bar with placeholder "Search for file name". It features a breadcrumb trail: Home / m015uf2_proyecto_subiros / covid19_dataset / covid19_flume_tweets. On the right, there are buttons for Upload and New. A table lists files in the directory:

Name	Size	User	Group	Permissions	Date
covid19_flume_tweets		training	supergroup	drwxrwxrwx	April 09, 2020 10:21 AM
.		training	supergroup	drwxrwxrwx	April 09, 2020 10:21 AM

Nom i Cognoms	Data
Arnau Subirós Puigarnau	21-04-2020

• *Flume Sources : Twitter (version 2)*

arxiu : flumeTwitter02.conf

```
rés                                training@localhost:~/training_materials/dev1/exercises/flume
File Edit View Search Terminal Help
GNU nano 2.0.9                         File: flumeTwitter02.conf

#### FlumeTwitter.conf

##versio 02

TwitterAgent.sources = Twitter
TwitterAgent.channels = MemChannel
TwitterAgent.sinks=HDFS
TwitterAgent.sources.Twitter.type = com.cloudera.flume.source.TwitterSource
TwitterAgent.sources.Twitter.channels = Memchannel

TwitterAgent.sources.Twitter.consumerKey= xb3ofAvBZBuEToTejGyMksijx
TwitterAgent.sources.Twitter.consumerSecret= PD4Ssm7VoJVGzV0us4n0g33Ev9M1wE0TAYx65GJittQmhfyLL
TwitterAgent.sources.Twitter.accessToken= 2152855150-E56Mun2rRkIAoWp5m2xmVmzMC1ctkz1omvRsS
TwitterAgent.sources.Twitter.accessTokenSecret= 2C220pfFqv8Tq3rbfY4E05yI80rrLmc525BdP7nfUWFQ
TwitterAgent.sources.Twitter.keywords= covid19, wuhan, coronavirus, covid-19, 2020 coronavirus pandemic

TwitterAgent.sinks.HDFS.channel=MemChannel
TwitterAgent.sinks.HDFS.type=hdfs
TwitterAgent.sinks.HDFS.hdfs.path=dfs://localhost:9000/m015uf2_proiecte_subiros/covid19_dataset/covid19_flume_tweets
TwitterAgent.sinks.HDFS.hdfs.fileType=DataStream
TwitterAgent.sinks.HDFS.hdfs.writeformat=Text
TwitterAgent.sinks.HDFS.hdfs.batchSize=1000
TwitterAgent.sinks.HDFS.hdfs.rollSize=0
TwitterAgent.sinks.HDFS.hdfs.rollCount=10000
TwitterAgent.sinks.HDFS.hdfs.rollInterval=600

TwitterAgent.channels.MemChannel.type=memory
TwitterAgent.channels.MemChannel.capacity=10000
TwitterAgent.channels.MemChannel.transactionCapacity=10000
```

```
2020-04-10 03:32:42,491 (lifecycleSupervisor-1-0) [INFO - org.apache.flume.node.PollingPropertiesFileConfigurationProvider.start(PollingPropertiesFileConfigurationProvider.java:61)] Configuration provider started
2020-04-10 03:32:42,495 (conf-file-poller-0) [INFO - org.apache.flume.node.PollingPropertiesFileConfigurationProvider$FileWatcherRunnable.run(PollingPropertiesFileConfigurationProvider.java:133)] Reloading configuration file: /flumeTwitter02.conf
2020-04-10 03:32:42,499 (conf-file-poller-0) [INFO - org.apache.flume.conf.FlumeConfiguration$AgentConfiguration.addProperty(FlumeConfiguration.java:1017)] Processing:HDFS
2020-04-10 03:32:42,500 (conf-file-poller-0) [INFO - org.apache.flume.conf.FlumeConfiguration$AgentConfiguration.addProperty(FlumeConfiguration.java:1017)] Added sinks: HDFS Agent: TwitterAgent
2020-04-10 03:32:42,500 (conf-file-poller-0) [INFO - org.apache.flume.conf.FlumeConfiguration$AgentConfiguration.addProperty(FlumeConfiguration.java:1017)] Processing:HDFS
2020-04-10 03:32:42,500 (conf-file-poller-0) [INFO - org.apache.flume.conf.FlumeConfiguration$AgentConfiguration.addProperty(FlumeConfiguration.java:1017)] Post-validation flume configuration contains configuration for agents: [TwitterAgent]
2020-04-10 03:32:42,634 (conf-file-poller-0) [INFO - org.apache.flume.node.AbstractConfigurationProvider.loadChannels(AbstractConfigurationProvider.java:145)] Creating channels
2020-04-10 03:32:42,639 (conf-file-poller-0) [INFO - org.apache.flume.channel.DefaultChannelFactory.create(DefaultChannelFactory.java:42)] Creating instance of channel Memchannel type memory
2020-04-10 03:32:42,642 (conf-file-poller-0) [INFO - org.apache.flume.node.AbstractConfigurationProvider.loadChannels(AbstractConfigurationProvider.java:200)] Created channel Memchannel
2020-04-10 03:32:42,643 (conf-file-poller-0) [INFO - org.apache.flume.source.DefaultSourceFactory.create(DefaultSourceFactory.java:41)] Creating instance of source Twitter, type com.cloudera.flume.source.TwitterSource
2020-04-10 03:32:42,644 (conf-file-poller-0) [ERROR - org.apache.flume.node.PollingPropertiesFileConfigurationProvider$FileWatcherRunnable.run(PollingPropertiesFileConfigurationProvider.java:142)] Failed to load configuration data. Exception follows.
org.apache.flume.FlumeException: Unable to load source type: com.cloudera.flume.source.TwitterSource, class: com.cloudera.flume.source.TwitterSource
    at org.apache.flume.source.DefaultSourceFactory.getClazz(DefaultSourceFactory.java:69)
    at org.apache.flume.source.DefaultSourceFactory.create(DefaultSourceFactory.java:42)
    at org.apache.flume.node.AbstractConfigurationProvider.loadSources(AbstractConfigurationProvider.java:22)
    at org.apache.flume.node.PollingPropertiesFileConfigurationProvider$FileWatcherRunnable.run(PollingPropertiesFileConfigurationProvider.java:97)
    at org.apache.flume.node.PollingPropertiesFileConfigurationProvider$FileWatcherRunnable.run(PollingPropertiesFileConfigurationProvider.java:140)
    at java.util.concurrent.Executors$RunnableAdapter.call(Executors.java:471)
    at java.util.concurrent.FutureTask.runAndReset(FutureTask.java:304)
    at java.util.concurrent.ScheduledThreadPoolExecutor$ScheduledFutureTask.access$301(ScheduledThreadPoolExecutor.java:178)
    at java.util.concurrent.ScheduledThreadPoolExecutor$ScheduledFutureTask.run(ScheduledThreadPoolExecutor.java:293)
    at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1145)
    at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:615)
    at java.lang.Thread.run(Thread.java:745)
Caused by: java.lang.ClassNotFoundException: com.cloudera.flume.source.TwitterSource
    at java.net.URLClassLoader$1.run(URLClassLoader.java:360)
    at java.net.URLClassLoader.findClass(URLClassLoader.java:355)
    at java.net.FactoryURLClassLoader.findClass(FactoryURLClassLoader.java:354)
    at java.lang.ClassLoader.loadClass(ClassLoader.java:425)
    at sun.misc.Launcher$AppClassLoader.loadClass(Launcher.java:308)
    at java.lang.ClassLoader.loadClass(ClassLoader.java:358)
    at java.lang.Class.forName0(Native Method)
    at java.lang.Class.forName(Class.java:190)
    at org.apache.flume.source.DefaultSourceFactory.getClazz(DefaultSourceFactory.java:67)
    ... 11 more
```

Nom i Cognoms	Data
Arnau Subirós Puigarnau	21-04-2020

3. Transformació i persistència d'un RDD amb spark shell

Desde el terminal executem pyspark(està configurat perquè treballem desde Jupyter Notebook

```
[training@localhost data]$ pyspark
[I 02:14:29.189 NotebookApp] Using MathJax:
[I 02:14:29.200 NotebookApp] The port 2222 is already in use.
```



The screenshot shows a Jupyter Notebook interface with the following content:

- Kernel:** Python 2
- In [1]:** `logfile="/m015uf2_proyecto_subiros/covid19_dataset/COVID19_line_list_data.csv"`
- Text:** Creacio d'una variable per l'arxiu de dades per no repetir-ho cada moment ubicat al HDFS
- In [2]:** `logs = sc.textFile(logfile)`
- Text:** Creacio d'un RDD de l'arxiu de dades
- In [3]:** `csvLog = logs.filter(lambda line : "Wuhan" in line)`
- Text:** Creacio d'un RDD que contigui nomes les línies amb la paraula "Wuhan"
- In [4]:** `csvLog.take(5)`
- Out[4]:** A list of five JSON-like strings representing COVID-19 cases. One entry is shown below:

```

[u'id,case_in_country,reporting
date,,summary,location,country,gender,age,symptom_onset>If_onset_approximated,hosp_visit_date,exposure_start,exposure_end
,visiting_Wuhan,from_Wuhan,death,recovered,symptom,source,link.....',
u'1,1/20/2020,First confirmed imported COVID-19 pneumonia patient in Shenzhen (from Wuhan); male, 66, shenzhen
residence, visited relatives in Wuhan on 12/29/2019, symptoms onset on 01/03/2020, returned to Shenzhen and seek medical
care on 01/04/2020, hospitalized on 01/11/2020, sample sent to China CDC for testing on 01/18/2020, confirmed on
01/19/2020. 8 others under medical observation, contact tracing ongoing.,"Shenzhen, Guangdong",China,male,66,01/03
/20,0,01/11/20,12/29/2019,01/04/20,1,0,0,0.,Shenzhen Municipal Health Commission,http://wjw.sz.gov.cn/wzx/202001
/20200120_18987787.htm.....',
u'2,1/20/2020,First confirmed imported COVID-19 pneumonia patient in Shanghai (from Wuhan); female, 56, Wuhan
]

```

Nom i Cognoms	Data
Arnau Subirós Puigarnau	21-04-2020

Jupyter Notebook02-CovidHDFS Last Checkpoint: 24 minutes ago (autosaved)

File Edit View Insert Cell Kernel Help

Python 2

Cell Toolbar: None

Contar el numero de peticions "Wuhan" , executem una comanda sense guardar objectes intermediaris en una variable

```
In [5]: sc.textFile(logfile).filter(lambda line : "Wuhan" in line).count()
Out[5]: 326
```

Utilitzarem la funcio "map per definir un nou RDD. Comencem amb un simple mapa que retorna la longitud de cada línia en el arxiu

```
In [6]: logs.map(lambda line : line.split()).take(5)
Out[6]: [[u'id,case_in_country,reporting',
          u'date,,summary,location,country,gender,age,symptom_onset>If_onset_approximated,hosp_visit_date,exposure_start,exposure_end,visiting',
          u'Wuhan,from',
          u'Wuhan,death,recovered,symptom,source,link,,,...'],
         [u'1,,1/20/2020,,First',
          u'confirmed',
          u'imported',
          u'COVID-19',
          u'pneumonia',
          u'patient',
          u'in',
          u' Shenzhen',
          u'(from',
          u'Wuhan):',
          u'male:',
          u'66:',
          u'shenzhen']]
```

Ara definim un nou RDD

```
In [7]: log02 = logs.map (lambda line : line.split() [0])
In [8]: log02.take(7)
Out[8]: [u'id,case_in_country,reporting',
          u'1,,1/20/2020,,First',
          u'2,,1/20/2020,,First',
          u'3,,1/21/2020,,First',
          u'4,,1/21/2020,,new',
          u'5,,1/21/2020,,new',
          u'6,,1/21/2020,,First']
```

In []:

Nom i Cognoms	Data
Arnau Subirós Puigarnau	21-04-2020

jupyter Notebook02-CovidHDFS Last Checkpoint: an hour ago (unsaved changes)

```

File Edit View Insert Cell Kernel Help
Cell Toolbar: None | Python

##### Arxiu covid_19_data.csv

In [1]: logfile02="/m015uf2_proyecte_subiros/covid19_dataset/covid_19_data.csv"

In [2]: print logfile02
/m015uf2_proyecte_subiros/covid19_dataset/covid_19_data.csv

In [3]: logs02 = sc.textFile(logfile02)

In [4]: logs02.count()
Out[4]: 9118

In [12]: for line in logs02.take(6) : print line
SNo,ObservationDate,Province/State,Country/Region,Last Update,Confirmed,Deaths,Recovered
1,01/22/2020,Anhui,Mainland China,1/22/2020 17:00,1.0,0.0,0.0
2,01/22/2020,Beijing,Mainland China,1/22/2020 17:00,14.0,0.0,0.0
3,01/22/2020,Chongqing,Mainland China,1/22/2020 17:00,6.0,0.0,0.0
4,01/22/2020,Fujian,Mainland China,1/22/2020 17:00,1.0,0.0,0.0
5,01/22/2020,Gansu,Mainland China,1/22/2020 17:00,0.0,0.0,0.0

# Creació d'una funció per convertir totes les paraules a majusculas

In [7]: def toUpper(s):
         return s.upper()

In [8]: logs02 = sc.textFile(logfile02)

In [11]: logs02.map(toUpper).take(6)
Out[11]: [u'SNO,OBSERVATIONDATE,PROVINCE/STATE,COUNTRY/REGION,LAST UPDATE,CONFIRMED,DEATHS,RECOVERED',
u'1,01/22/2020,ANHUI,MAINLAND CHINA,1/22/2020 17:00,1.0,0.0,0.0',
u'2,01/22/2020,BEIJING,MAINLAND CHINA,1/22/2020 17:00,14.0,0.0,0.0',
u'3,01/22/2020,CHONGQING,MAINLAND CHINA,1/22/2020 17:00,6.0,0.0,0.0',
u'4,01/22/2020,FUJIAN,MAINLAND CHINA,1/22/2020 17:00,1.0,0.0,0.0',
u'5,01/22/2020,GANSU,MAINLAND CHINA,1/22/2020 17:00,0.0,0.0,0.0'],

```

Creació d'una funció per convertir totes les paraules a majusculas

```

In [14]: def toUpper(s):
           return s.upper()

In [15]: logs02 = sc.textFile(logfile02)

In [16]: logs02.map(toUpper).take(6)
Out[16]: [u'SNO,OBSERVATIONDATE,PROVINCE/STATE,COUNTRY/REGION,LAST UPDATE,CONFIRMED,DEATHS,RECOVERED',
u'1,01/22/2020,ANHUI,MAINLAND CHINA,1/22/2020 17:00,1.0,0.0,0.0',
u'2,01/22/2020,BEIJING,MAINLAND CHINA,1/22/2020 17:00,14.0,0.0,0.0',
u'3,01/22/2020,CHONGQING,MAINLAND CHINA,1/22/2020 17:00,6.0,0.0,0.0',
u'4,01/22/2020,FUJIAN,MAINLAND CHINA,1/22/2020 17:00,1.0,0.0,0.0',
u'5,01/22/2020,GANSU,MAINLAND CHINA,1/22/2020 17:00,0.0,0.0,0.0']



### Creacció RDD de col·leccions en lloc de fitxers



```

In [17]: logs02 = ["Anhui","Beijing","Fujian","Chongqing","Fujian","Gansu"]

In [18]: logs02RDD = sc.parallelize(logs02)

In [19]: logs02RDD.take(4)
Out[19]: ['Anhui', 'Beijing', 'Fujian', 'Chongqing']

In []:

```


```

Nom i Cognoms	Data
Arnau Subirós Puigarnau	21-04-2020

Nom i Cognoms	Data
Arnaud Subirós Puigarnau	21-04-2020

CREANT PAIR RDD - FEM PROVES

```
In [12]: PAIR_RDD1 = logfile03_RDD1.map(lambda line : line.split('[') and line.split('/') and line.split(',') )
In [13]: PAIR_RDD1 = logfile03_RDD1.map(lambda line : line.split(','))
In [14]: PAIR_RDD1 = logfile03_RDD1.map(lambda fields : ( fields[2],fields[1],fields[2]))
In [15]: PAIR_RDD1.take(6)
Out[15]: [(u'o', u'r', u'o'),
(u'f', u'A', u'f'),
(u'l', u'A', u'l'),
(u'l', u'A', u'l'),
(u'n', u'A', u'n'),
(u'n', u'A', u'n')]
```

Fem un altre Notebook anomena CovidVersio2



The screenshot shows a Jupyter Notebook interface with the following details:

- Title:** Covidversio2
- Kernel:** Python 2
- Toolbar:** File, Edit, View, Insert, Cell, Kernel, Help, Cell Toolbar: None
- Content:**
 - Header:** JESUITES educació
 - Section:** APACHE HADOOP I APACHE SPARK (i ECOSISTEMA)
 - Text:** M015 uf2 Big Data, Curs: DAM2 2019-20, Alumne: Arnaud Subirós
 - Image:** Apache Spark logo
 - Text:** Creacio d'una variable per l'arxiu de dades per no repetir-ho cada moment ubicat al HDFS Arxiu covid_19_data.csv"
 - Code:**

```
In [2]: covifile01 = "/m015uf2_projecte_subiros/covid19_dataset/covid_19_data.csv"
In [3]: print covifile01
/m015uf2_projecte_subiros/covid19_dataset/covid_19_data.csv
In [4]: data = sc.textFile(covifile01)
```

Nom i Cognoms	Data
Arnau Subirós Puigarnau	21-04-2020

Creacio d'una variable per l'arxiu de dades per no repetir-ho cada moment ubicat al HDFS Arxiu covid_19_data.csv"

```
In [2]: covifile01 ="/m015uf2_proyecte_subiros/covid19_dataset/covid_19_data.csv"

In [3]: print covifile01
/m015uf2_proyecte_subiros/covid19_dataset/covid_19_data.csv

In [4]: data = sc.textFile(covifile01)

In [5]: data.count()
Out[5]: 9118

In [11]: #filtra les entrades i noems selecciona les que continguin "Hubei"
dataHubei = data.filter(lambda line : "Hubei" in line )

In [12]: #visualitzem les 5 primeres entrades
dataHubei.take(5)

Out[12]: [u'14,01/22/2020,Hubei,Mainland China,1/22/2020 17:00,444.0,17.0,28.0',
u'52,01/23/2020,Hubei,Mainland China,1/23/20 17:00,444.0,17.0,28.0',
u'85,01/24/2020,Hubei,Mainland China,1/24/20 17:00,549.0,24.0,31.0',
u'126,01/25/2020,Hubei,Mainland China,1/25/20 17:00,761.0,40.0,32.0',
u'170,01/26/2020,Hubei,Mainland China,1/26/20 16:00,1058.0,52.0,42.0']

In [14]: #Fem un count del filtre sense variables intermitges
sc.textFile(covifile01).filter(lambda line: "Hubei" in line).count()

Out[14]: 65
```

Nom i Cognoms	Data
Arnau Subirós Puigarnau	21-04-2020

```
In [8]: #Definim un nou RDD que contingui els caracters de les 5 primeres entrades (númeric)
data .map(lambda line: len(line)).take(5)

Out[8]: [88, 61, 64, 65, 62]

In [9]: #Definim un nou RDD que contingui els caracters de les 5 primeres entrades (arrays)
data.map(lambda line: line.split()).take(5)

Out[9]: [[u'SNo,ObservationDate,Province/State,Country/Region,Last',
  u'Update,Confirmed,Deaths,Recovered'],
[u'1,01/22/2020,Anhui,Mainland', u'China,1/22/2020', u'17:00,1.0,0.0,0.0'],
[u'2,01/22/2020,Beijing,Mainland', u'China,1/22/2020', u'17:00,14.0,0.0,0.0'],
[u'3,01/22/2020,Chongqing,Mainland',
  u'China,1/22/2020',
  u'17:00,6.0,0.0,0.0'],
[u'4,01/22/2020,Fujian,Mainland', u'China,1/22/2020', u'17:00,1.0,0.0,0.0']]

In [10]: #Aga noms data 2 ,ls qual son la primera paraula e cada entrada. Es crea un RDD i noms es mostres els 5 primers .
data2 = data.map(lambda line: line.split()[0])
data2.take(5)

Out[10]: [u'SNo,ObservationDate,Province/State,Country/Region,Last',
  u'1,01/22/2020,Anhui,Mainland',
  u'2,01/22/2020,Beijing,Mainland',
  u'3,01/22/2020,Chongqing,Mainland',
  u'4,01/22/2020,Fujian,Mainland']

In [11]: #Podem iterar les entrades per fer-les més legibles
for iteraData2 in data2.take(5): print iteraData2

SNo,ObservationDate,Province/State,Country/Region,Last
1,01/22/2020,Anhui,Mainland
2,01/22/2020,Beijing,Mainland
3,01/22/2020,Chongqing,Mainland
4,01/22/2020,Fujian,Mainland
```

Nom i Cognoms
Data

Arnaud Subirós Puigarnau

21-04-2020

 Jupyter Covidversio2 Last Checkpoint: an hour ago (autosaved)

File Edit View Insert Cell Kernel Help Python 2

```
In [12]: # Exemple de PAIR RDD
rdd1 = data.map(lambda line : line.split(','))
rdd1.map(lambda fields : (fields[0] , fields[1]))
rdd1.take(3)

Out[12]: [(u'SNo',
u'ObservationDate',
u'Province/State',
u'Country/Region',
u'Last Update',
u'Confirmed',
u'Deaths',
u'Recovered'],
[u'1',
u'01/22/2020',
u'Anhui',
u'Mainland China',
u'1/22/2020 17:00',
u'1.0',
u'0.0',
u'0.0'],
[u'2',
u'01/22/2020',
u'Beijing',
u'Mainland China',
u'1/22/2020 17:00',
u'14.0',
u'0.0',
u'0.0'])

In [13]: def funcioDemo():
rdd2 = rdd1.map(lambda fields: (fields[0],fields[1]))
rdd2.take(5)
print("prova")

In [14]: funcioDemo()
prova

In [16]: #Guardem el RDD en un fitxer ubicat al HDFS
data2.saveAsTextFile("/m015uf2_proyecto_subiros/covid19_dataset/covid_notebook2")
```

Most Visited ▾ Cloudera Hue YARN RM Spark UI (local) Spark Doc Solr Admin UI Kite SDK Doc Jupyter Jesuites-DAM2

HUE Home Query Editors Data Browsers Workflows Search File Browser Job Bl

File Browser

Search for file name Actions Move to trash

Home / m015uf2_proyecto_subiros / covid19_dataset / covid_notebook2

Name	Size	User	Group	Permissions
._SUCCESS	0 bytes	training	supergroup	drwxrwxrwx
part-00000	165.2 KB	training	supergroup	-rw-rw-rw-
part-00001	195.0 KB	training	supergroup	-rw-rw-rw-

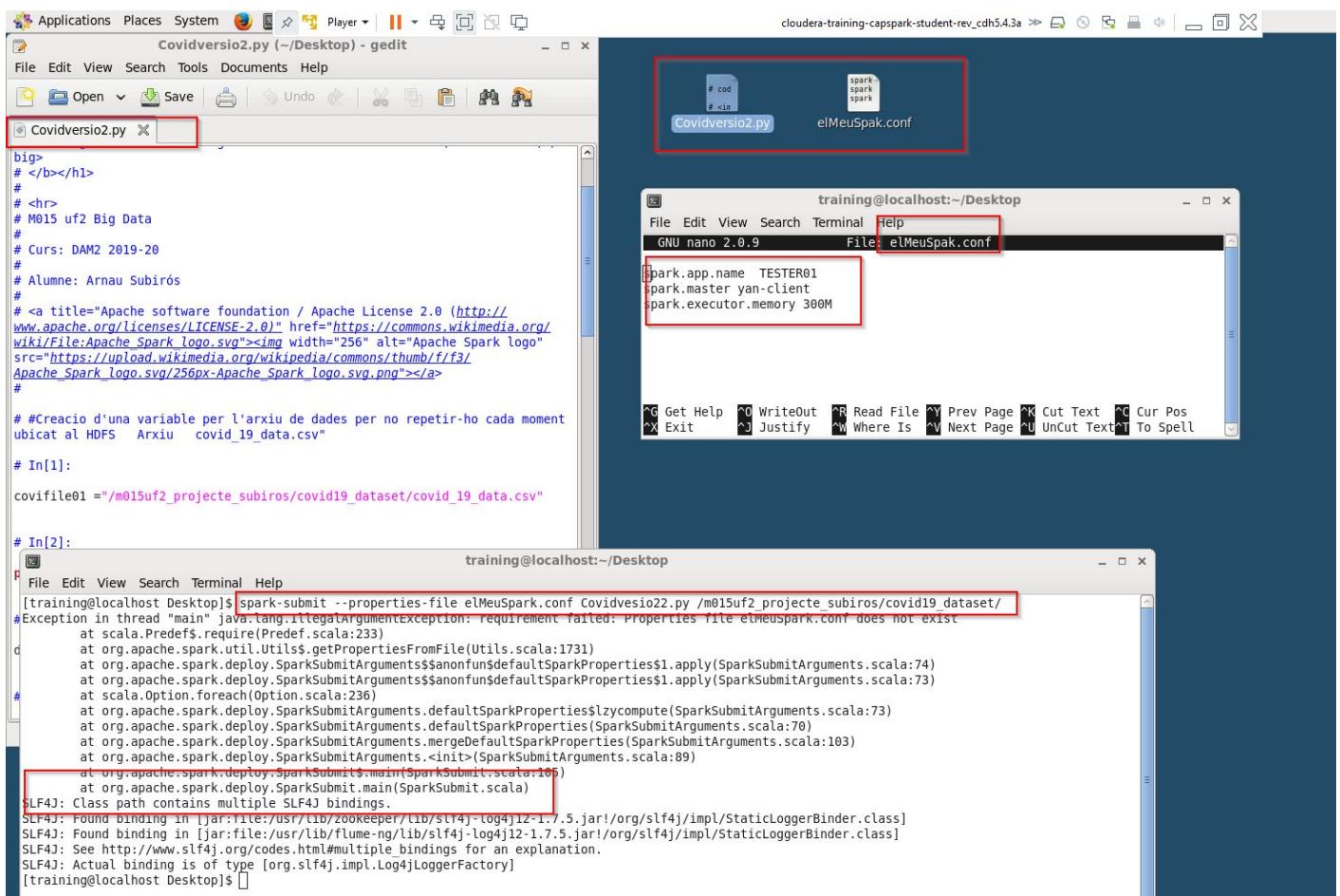
Nom i Cognoms

Arnaud Subirós Puigarnau

Data

21-04-2020

• Intentem crear una aplicació amb SPARK



The screenshot shows a Linux desktop environment with three windows:

- Code Editor:** A window titled "Covidversio2.py (~/Desktop) - gedit" containing Python code. The code includes imports, a class definition, and a main function that reads a CSV file from HDFS.
- Terminal 1:** A window titled "training@localhost:~/Desktop" showing the configuration file "elMeuSpark.conf". It contains properties for the Spark application: app.name set to TESTER01, master set to yarn-client, and executor.memory set to 300M.
- Terminal 2:** A window titled "training@localhost:~/Desktop" showing the command "spark-submit --properties-file elMeuSpark.conf Covidversio2.py /m015uf2_projecte_subiros/covid19_dataset/" being run. The output shows an exception due to a missing properties file, followed by multiple SLF4J binding errors.

Nom i Cognoms	Data
Arnau Subirós Puigarnau	21-04-2020

```
training@localhost:~/Desktop
File Edit View Search Terminal Help

[training@localhost Desktop]$ spark-submit --properties-file elMeuSpark.conf Covidversio2.py /m015uf2 projecte subiros
Exception in thread "main" java.lang.IllegalArgumentException: requirement failed: Properties file elMeuSpark.conf does not exist
    at scala.Predef$.require(Predef.scala:233)
    at org.apache.spark.util.Utils$.getPropertiesFromFile(Utils.scala:1731)
    at org.apache.spark.deploy.SparkSubmitArguments$$anonfun$defaultSparkProperties$1.apply(SparkSubmitArguments.scala:74)
    at org.apache.spark.deploy.SparkSubmitArguments$$anonfun$defaultSparkProperties$1.apply(SparkSubmitArguments.scala:73)
    at scala.Option.foreach(Option.scala:236)
    at org.apache.spark.deploy.SparkSubmitProperties$$lzycompute$(SparkSubmitArguments.scala:73)
    at org.apache.spark.deploy.SparkSubmitArguments.defaultSparkProperties(SparkSubmitArguments.scala:70)
    at org.apache.spark.deploy.SparkSubmitArguments.mergeDefaultSparkProperties(SparkSubmitArguments.scala:103)
    at org.apache.spark.deploy.SparkSubmitArguments.<init>(SparkSubmitArguments.scala:89)
    at org.apache.spark.deploy.SparkSubmit$.main(SparkSubmit.scala:105)
    at org.apache.spark.deploy.SparkSubmit.main(SparkSubmit.scala)
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/zookeeper/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/flume-ng/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
[training@localhost Desktop]$ spark-submit --properties-file elMeuSpark.conf Covidversio2.py /m015uf2 projecte_subiros
Error: Master must start with yarn, spark, mesos, or local
Run with -h for usage help or --verbose for debug output
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/zookeeper/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/flume-ng/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
[training@localhost Desktop]$ clear
```

```
training@localhost:~/Desktop
File Edit View Search Terminal Help

client token: N/A
diagnostics: N/A
ApplicationMaster host: N/A
ApplicationMaster RPC port: -1
queue: root.root
start time: 1587142894376
final status: UNDEFINED
tracking URL: http://localhost:8088/proxy/application_1587120630708_0002/
user: root
20/04/17 10:01:36 INFO yarn.Client: Application report for application 1587120630708_0002 (state: ACCEPTED)
20/04/17 10:01:37 INFO yarn.Client: Application report for application 1587120630708_0002 (state: ACCEPTED)
20/04/17 10:01:38 INFO yarn.Client: Application report for application 1587120630708_0002 (state: ACCEPTED)
20/04/17 10:01:39 INFO yarn.Client: Application report for application 1587120630708_0002 (state: ACCEPTED)
20/04/17 10:01:40 INFO yarn.Client: Application report for application 1587120630708_0002 (state: ACCEPTED)
20/04/17 10:01:41 INFO yarn.Client: Application report for application 1587120630708_0002 (state: ACCEPTED)
20/04/17 10:01:42 INFO yarn.Client: Application report for application 1587120630708_0002 (state: ACCEPTED)
20/04/17 10:01:43 INFO yarn.Client: Application report for application 1587120630708_0002 (state: ACCEPTED)
20/04/17 10:01:44 INFO yarn.Client: Application report for application 1587120630708_0002 (state: ACCEPTED)
20/04/17 10:01:45 INFO yarn.Client: Application report for application 1587120630708_0002 (state: ACCEPTED)
20/04/17 10:01:46 INFO yarn.Client: Application report for application 1587120630708_0002 (state: ACCEPTED)
20/04/17 10:01:47 INFO yarn.Client: Application report for application 1587120630708_0002 (state: FAILED)
20/04/17 10:01:47 INFO yarn.Client:
    client token: N/A
    diagnostics: Application application_1587120630708_0002 failed 2 times due to AM Container for appattempt_1587120630708_000002 exited with exitCode: 1
For more detailed output, check application tracking page: http://localhost:8088/proxy/application_1587120630708_0002/Then, click on links to logs of each attempt.
Diagnostics: Exception from container-launch.
Container id: container_1587120630708_0002_02_000001
Exit code: 1
Stack trace: ExitCodeException exitCode=1:
    at org.apache.hadoop.util.Shell.runCommand(Shell.java:528)
```

Nom i Cognoms
Data

Arnau Subirós Puigarnau

21-04-2020

localhost:8088/cluster/app/application_1587120630708_0002

Most Visited ▾ Cloudera Hue YARN RM Spark UI (local) Spark Doc Solr Admin UI Kite SDK Doc Jupyter Jesuites-DAM2

Cluster

- About
- Nodes
- Applications
 - NEW
 - NEW_SAVING
 - SUBMITTED
 - ACCEPTED
 - RUNNING
 - FINISHED
 - FAILED
 - KILLED
- Scheduler
- Tools

User: root
 Name: TESTER01
 Application Type: SPARK
 Application Tags:
 State: FAILED
 FinalStatus: FAILED
 Started: Fri Apr 17 10:01:34 -0700 2020
 Elapsed: 12sec
 Tracking URL: History
 Diagnostics: Application application_1587120630708_0002 failed 2 times due to AM Container for appattempt_1587120630708_0002_000002 exited with exitCode: 1
 For more detailed output, check application tracking page: http://localhost:8088/proxy/application_1587120630708_0002/Then, click on links to logs of each attempt.
 Container id: container_1587120630708_0002_02_000001
 Exit code: 1
 Stack trace: ExitCodeException exitCode=1:
 at org.apache.hadoop.util.Shell.runCommand(Shell.java:538)
 at org.apache.hadoop.util.Shell.run(Shell.java:455)
 at org.apache.hadoop.util.Shell\$ShellCommandExecutor.execute(Shell.java:715)
 at org.apache.hadoop.yarn.server.nodemanager.DefaultContainerExecutor.launchContainer(DefaultContainerExecutor.java:211)
 at org.apache.hadoop.yarn.server.nodemanager.containermanager.launcher.ContainerLaunch.call(ContainerLaunch.java:302)
 at org.apache.hadoop.yarn.server.nodemanager.containermanager.launcher.ContainerLaunch.call(ContainerLaunch.java:82)
 at java.util.concurrent.FutureTask.run(FutureTask.java:262)
 at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1145)
 at java.util.concurrent.ThreadPoolExecutor\$Worker.run(ThreadPoolExecutor.java:615)
 at java.lang.Thread.run(Thread.java:745)
 Container exited with a non-zero exit code 1
 Failing this attempt. Failing the application.

Application Metrics

Total Resource Preempted:	<memory:0, vCores:0>
Total Number of Non-AM Containers Preempted:	0
Total Number of AM Containers Preempted:	0
Resource Preempted from Current Attempt:	<memory:0, vCores:0>
Number of Non-AM Containers Preempted from Current Attempt:	0
Aggregate Resource Allocation:	10264 MB-seconds, 10 vcore-seconds

localhost:8088/cluster

Most Visited ▾ Cloudera Hue YARN RM Spark UI (local) Spark Doc Solr Admin UI Kite SDK Doc Jupyter Jesuites-DAM2

Logged in as: drwho



All Applications

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	Vcores Used	Vcores Total	Vcores Reserved	Active Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes
7	0	0	7	0	0 B	2 GB	0 B	0	8	0	1	0	0	0	0

User Metrics for dr.who

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Containers Pending	Containers Reserved	Memory Used	Memory Pending	Memory Reserved	Vcores Used	Vcores Pending	Vcores Reserved
0	0	0	7	0	0	0	0 B	0 B	0 B	0	0	0

Show 20 entries Search:

ID	User	Name	Application Type	Queue	StartTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU Vcores	Allocated Memory MB	Progress	Tracking UI
application_1587120630708_0008	training	Covidversio2.py	SPARK	root.training	Fri Apr 17 11:48:35 -0700 2020	Fri Apr 17 11:48:42 -0700 2020	FAILED	FAILED	N/A	N/A	N/A		History
application_1587120630708_0007	training	Covidversio2.py	SPARK	root.training	Fri Apr 17 11:37:02 -0700 2020	Fri Apr 17 11:37:09 -0700 2020	FAILED	FAILED	N/A	N/A	N/A		History
application_1587120630708_0006	training	wordcount.py	SPARK	root.training	Fri Apr 17 11:32:14 -0700 2020	Fri Apr 17 11:33:52 -0700 2020	FINISHED	SUCCEEDED	N/A	N/A	N/A		History
application_1587120630708_0004	root	TESTER01	SPARK	root.root	Fri Apr 17 11:08:00 -0700 2020	Fri Apr 17 11:08:09 -0700 2020	FAILED	FAILED	N/A	N/A	N/A		History
application_1587120630708_0003	root	Covidversio2.py	SPARK	root.root	Fri Apr 17 11:00:47 -0700 2020	Fri Apr 17 11:00:57 -0700 2020	FAILED	FAILED	N/A	N/A	N/A		History

Nom i Cognoms	Data
Arnau Subirós Puigarnau	21-04-2020

Hue - File ... Failed redirect... Home Covidversio2 Covid19Nu... File:Apach... File:Apach... Home --proprie... Failed redirect... http://l...08_0006 +
 localhost:8042/node/containerlogs/container_1587120630708_0008_01_000001/training Google
 Most Visited Cloudera Hue YARN RM Spark UI (local) Spark Doc Solr Admin UI Kite SDK Doc Jupyter Jesuites-DAM2
 Logged in as: drwho

Failed redirect for container_1587120630708_0008_01_000001

Failed while trying to construct the redirect url to the log server. Log Server url may not be configured
 java.lang.Exception: Unknown container. Container either has not started or has already completed
 or doesn't belong to this node at all.

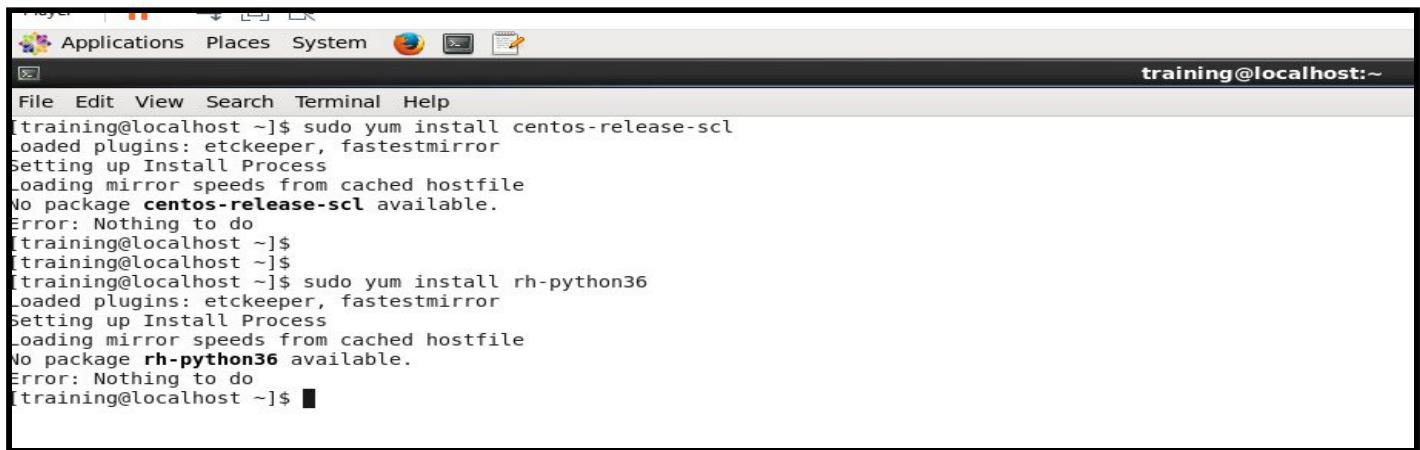
intentem instal.lar pandas

```
File Edit View Search Terminal Help
[redacted]
raise RuntimeError("Python version >= 3.5 required.")
RuntimeError: Python version >= 3.5 required.

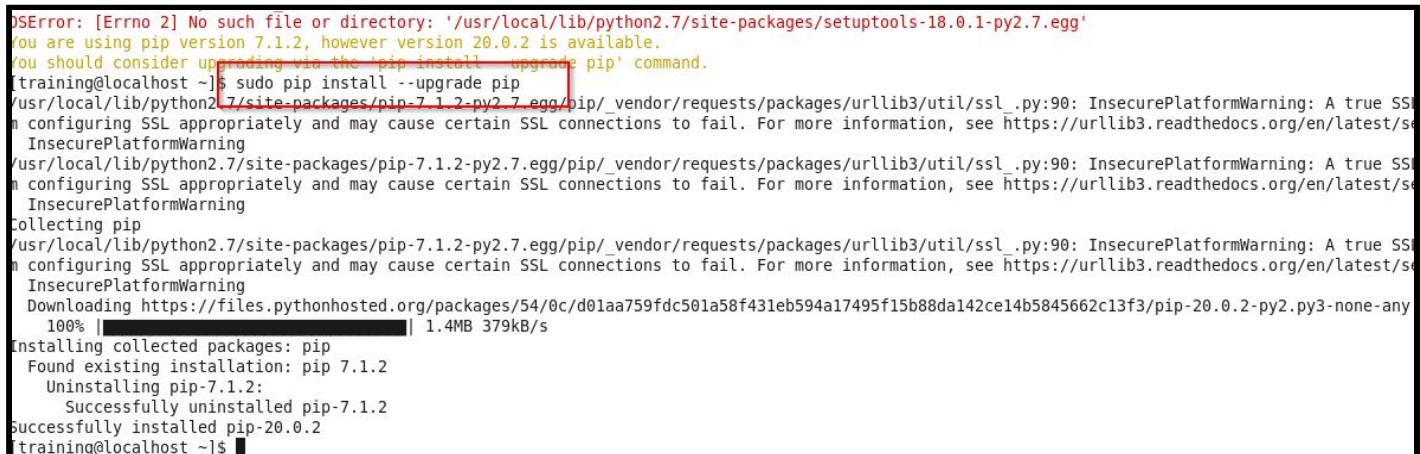
-----
Command "python setup.py egg_info" failed with error code 1 in /tmp/pip-build-hv
Ieyr/numpy
/usr/local/lib/python2.7/site-packages/pip-7.1.2-py2.7.egg/pip/_vendor/requests/
packages/urllib3/util/ssl_.py:90: InsecurePlatformWarning: A true SSLContext obj
ect is not available. This prevents urllib3 from configuring SSL appropriately a
nd may cause certain SSL connections to fail. For more information, see https://
urllib3.readthedocs.org/en/latest/security.html#insecureplatformwarning.
InsecurePlatformWarning
/usr/local/lib/python2.7/site-packages/pip-7.1.2-py2.7.egg/pip/_vendor/requests/
packages/urllib3/util/ssl_.py:90: InsecurePlatformWarning: A true SSLContext obj
ect is not available. This prevents urllib3 from configuring SSL appropriately a
nd may cause certain SSL connections to fail. For more information, see https://
urllib3.readthedocs.org/en/latest/security.html#insecureplatformwarning.
InsecurePlatformWarning
You are using pip version 7.1.2, however version 20.0.2 is available.
You should consider upgrading via the 'pip install --upgrade pip' command.
[training@localhost ~]$ python --version
Python 2.6.6
[training@localhost ~]$ ^C
[training@localhost ~]$
```

- *intentem instal.lar sense èxit Python 3*

Nom i Cognoms	Data
Arnau Subirós Puigarnau	21-04-2020



```
[training@localhost ~]$ sudo yum install centos-release-scl
Loaded plugins: etckeeper, fastestmirror
Setting up Install Process
Loading mirror speeds from cached hostfile
No package centos-release-scl available.
Error: Nothing to do
[training@localhost ~]$
[training@localhost ~]$ sudo yum install rh-python36
Loaded plugins: etckeeper, fastestmirror
Setting up Install Process
Loading mirror speeds from cached hostfile
No package rh-python36 available.
Error: Nothing to do
[training@localhost ~]$
```



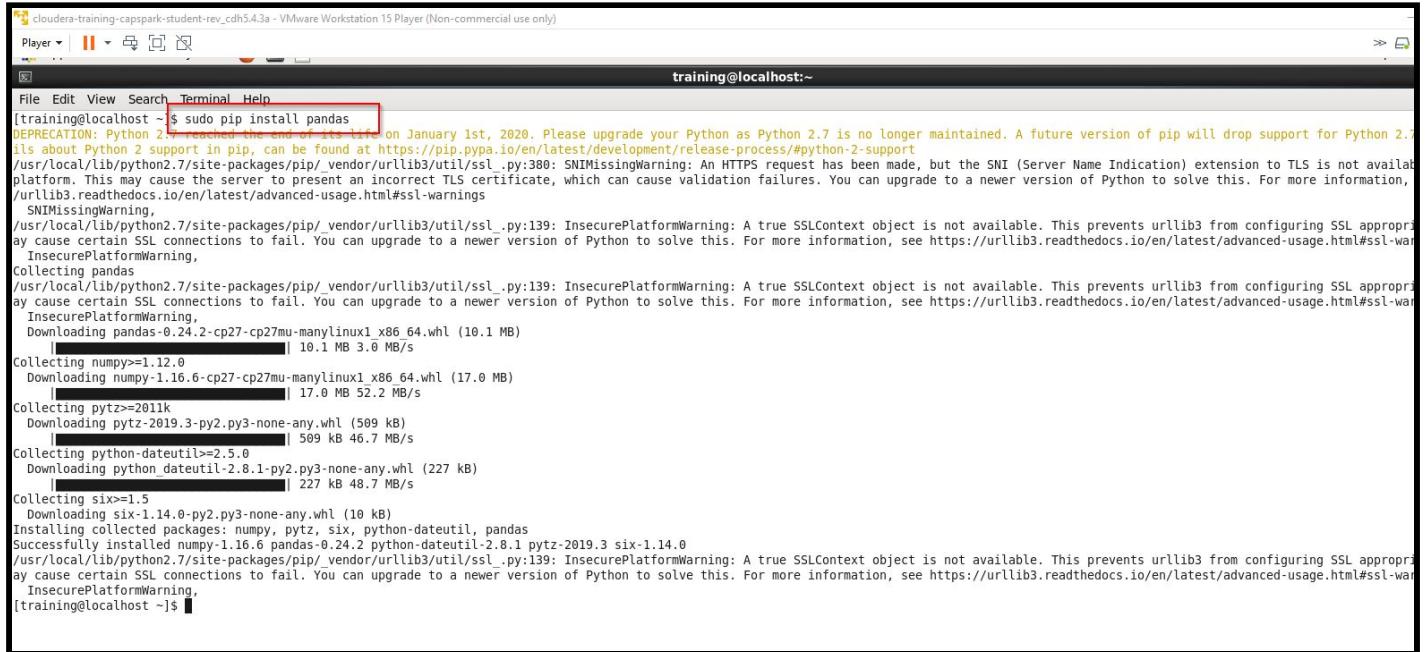
```
DSError: [Errno 2] No such file or directory: '/usr/local/lib/python2.7/site-packages/setuptools-18.0.1-py2.7.egg'
You are using pip version 7.1.2, however version 20.0.2 is available.
You should consider upgrading via 'pip install --upgrade pip'.
[training@localhost ~]$ sudo pip install --upgrade pip
/usr/local/lib/python2.7/site-packages/pip-7.1.2-py2.7.egg/pip/_vendor/requests/packages/urllib3/util/ssl_.py:90: InsecurePlatformWarning: A true SSL configuration SSL appropriately and may cause certain SSL connections to fail. For more information, see https://urllib3.readthedocs.org/en/latest/se
InsecurePlatformWarning
/usr/local/lib/python2.7/site-packages/pip-7.1.2-py2.7.egg/pip/_vendor/requests/packages/urllib3/util/ssl_.py:90: InsecurePlatformWarning: A true SSL configuration SSL appropriately and may cause certain SSL connections to fail. For more information, see https://urllib3.readthedocs.org/en/latest/se
InsecurePlatformWarning
Collecting pip
/usr/local/lib/python2.7/site-packages/pip-7.1.2-py2.7.egg/pip/_vendor/requests/packages/urllib3/util/ssl_.py:90: InsecurePlatformWarning: A true SSL configuration SSL appropriately and may cause certain SSL connections to fail. For more information, see https://urllib3.readthedocs.org/en/latest/se
InsecurePlatformWarning
  Downloading https://files.pythonhosted.org/packages/54/0c/d01aa759fdc501a58f431eb594a17495f15b88da142ce14b5845662c13f3/pip-20.0.2-py2.py3-none-any
    100% |██████████| 1.4MB 379kB/s
Installing collected packages: pip
  Found existing installation: pip 7.1.2
    Uninstalling pip-7.1.2:
      Successfully uninstalled pip-7.1.2
Successfully installed pip-20.0.2
[training@localhost ~]$
```

Nom i Cognoms

Arnaud Subirós Puigarnau

Data

21-04-2020



```
[training@localhost ~]$ sudo pip install pandas
DEPRECATION: Python 2 reached the end of its life on January 1st, 2020. Please upgrade your Python as Python 2.7 is no longer maintained. A future version of pip will drop support for Python 2.7. Python 2.7 will reach the end of its life on January 1st, 2020. Please upgrade your Python as Python 2.7 is no longer maintained. A future version of pip will drop support for Python 2.7.
/usr/local/lib/python2.7/site-packages/pip/_vendor/urllib3/util/ssl_.py:380: SNIMissingWarning: An HTTPS request has been made, but the SNI (Server Name Indication) extension to TLS is not available. This may cause the server to present an incorrect TLS certificate, which can cause validation failures. You can upgrade to a newer version of Python to solve this. For more information, see https://urllib3.readthedocs.io/en/latest/advanced-usage.html#ssl-warn
  SNIMissingWarning,
/usr/local/lib/python2.7/site-packages/pip/_vendor/urllib3/util/ssl_.py:139: InsecurePlatformWarning: A true SSLContext object is not available. This prevents urllib3 from configuring SSL appropriately cause certain SSL connections to fail. You can upgrade to a newer version of Python to solve this. For more information, see https://urllib3.readthedocs.io/en/latest/advanced-usage.html#ssl-war
  InsecurePlatformWarning,
Collecting pandas
/usr/local/lib/python2.7/site-packages/pip/_vendor/urllib3/util/ssl_.py:139: InsecurePlatformWarning: A true SSLContext object is not available. This prevents urllib3 from configuring SSL appropriately cause certain SSL connections to fail. You can upgrade to a newer version of Python to solve this. For more information, see https://urllib3.readthedocs.io/en/latest/advanced-usage.html#ssl-war
  InsecurePlatformWarning,
  Downloading pandas-0.24.2-cp27-cp27mu-manylinux1_x86_64.whl (10.1 MB)
    |
    |██████████| 10.1 MB 3.0 MB/s
Collecting numpy>=1.12.0
  Downloading numpy-1.16.6-cp27-cp27mu-manylinux1_x86_64.whl (17.0 MB)
    |
    |██████████| 17.0 MB 52.2 MB/s
Collecting pytz>=2011k
  Downloading pytz-2019.3-py2.py3-none-any.whl (509 kB)
    |
    |██████████| 509 kB 46.7 MB/s
Collecting python-dateutil>=2.5.0
  Downloading python_dateutil-2.8.1-py2.py3-none-any.whl (227 kB)
    |
    |██████████| 227 kB 48.7 MB/s
Collecting six>=1.5
  Downloading six-1.14.0-py2.py3-none-any.whl (10 kB)
Installing collected packages: numpy, pytz, six, python-dateutil, pandas
Successfully installed numpy-1.16.6 pandas-0.24.2 python-dateutil-2.8.1 pytz-2019.3 six-1.14.0
/usr/local/lib/python2.7/site-packages/pip/_vendor/urllib3/util/ssl_.py:139: InsecurePlatformWarning: A true SSLContext object is not available. This prevents urllib3 from configuring SSL appropriately cause certain SSL connections to fail. You can upgrade to a newer version of Python to solve this. For more information, see https://urllib3.readthedocs.io/en/latest/advanced-usage.html#ssl-war
  InsecurePlatformWarning,
[training@localhost ~]$
```

Nom i Cognoms	Data
Arnau Subirós Puigarnau	21-04-2020

Bibliografia :

- <https://github.com/facebook/prophet/issues/418>
- <https://datatofish.com/k-means-clustering-python/>
- <https://github.com/pypa/pip/issues/6667>
- <https://linuxize.com/post/how-to-install-python-3-on-centos-7/>