M015: Crèdits de lliure elecció

UF2-Big Data

APACHE HADOOP I APACHE SPARK (i ECOSISTEMA)

Curs: 2019-20

CFGS: DAM2

Alumne: Arnau Subirós Puigarnau

Data:

Nom i Cognoms

Arnau Subirós Puigarnau

Data

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

Arnau Subirós Puigarnau

Data

21-04-2020

INDEX

pàgina

- 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

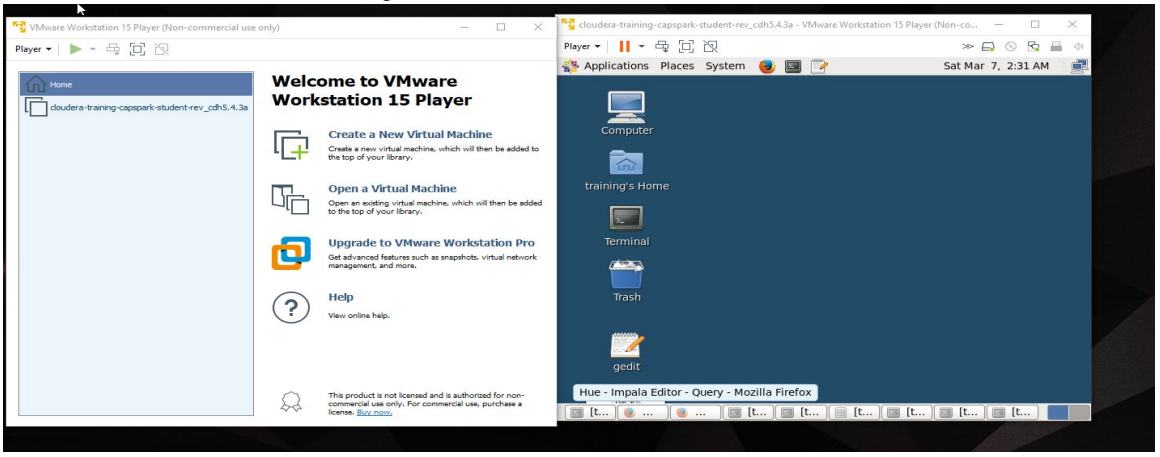
Nom i Cognoms **Data**

Arnau Subirós Puigarnau 21-04-2020

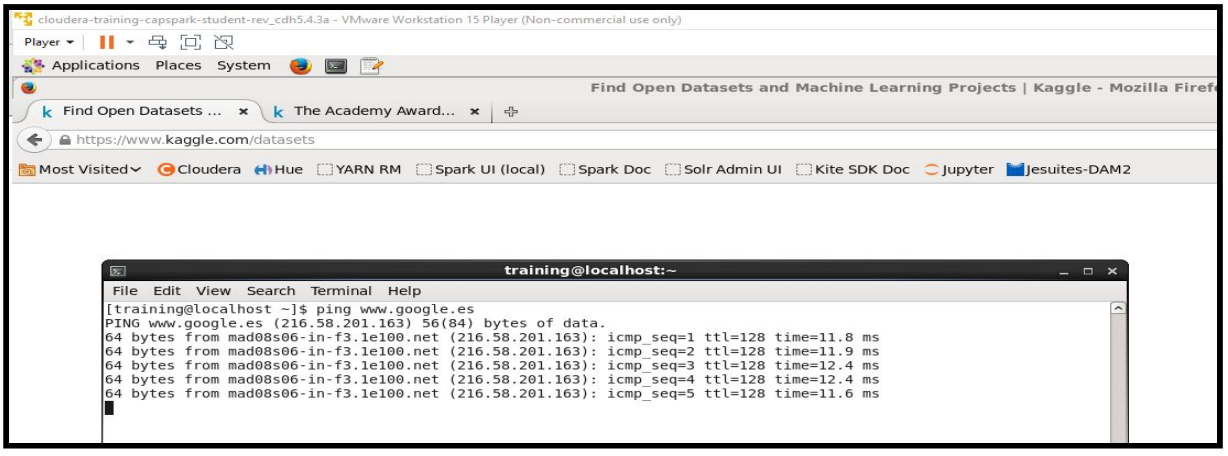
1. Importació de dades al HDFS

• Introducció

- Primer de tot obrim la màquina virtual

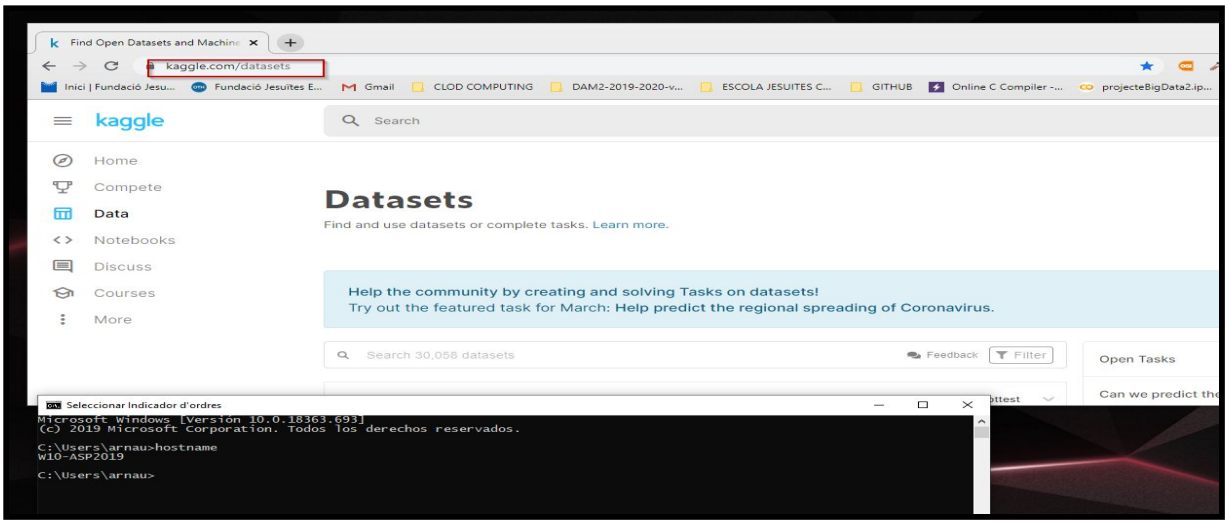


- Des de el navegador Firefox accedim a https://www.kaggle.com/datasets a la recerca de dades.En la màquina virtual, confirmo que tinc connexió a la xarxa, però la versió del navegador deu ser obsolet i no es pot visualitzar tot correctament.

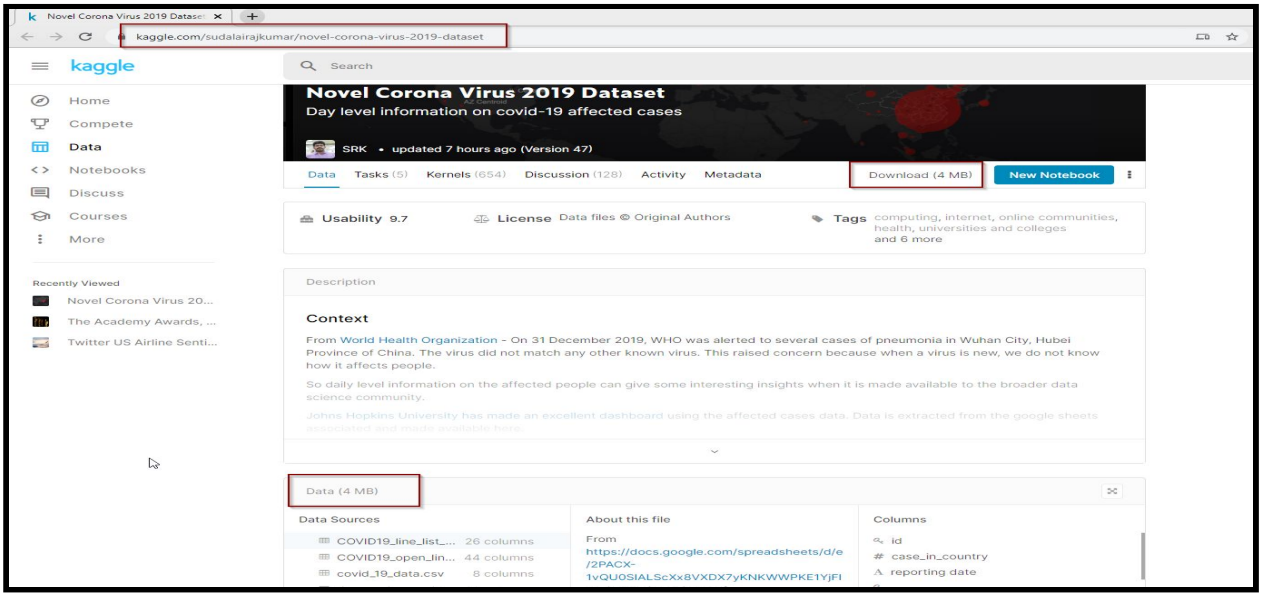


- Per tan desde el Windows 10 faré la recerca i guardarà l'arxiu csv en una carpeta compartida.

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



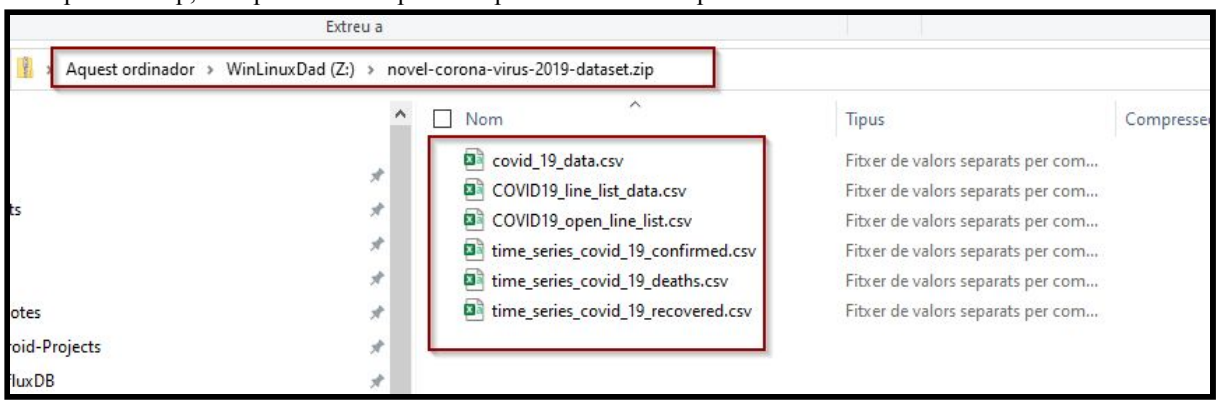
- Seleccione els datasets amb els quals voldrem treballar
  https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset

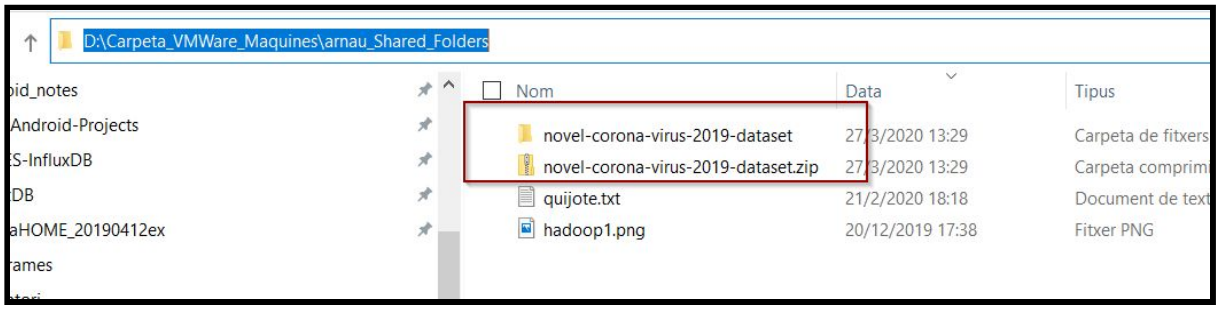


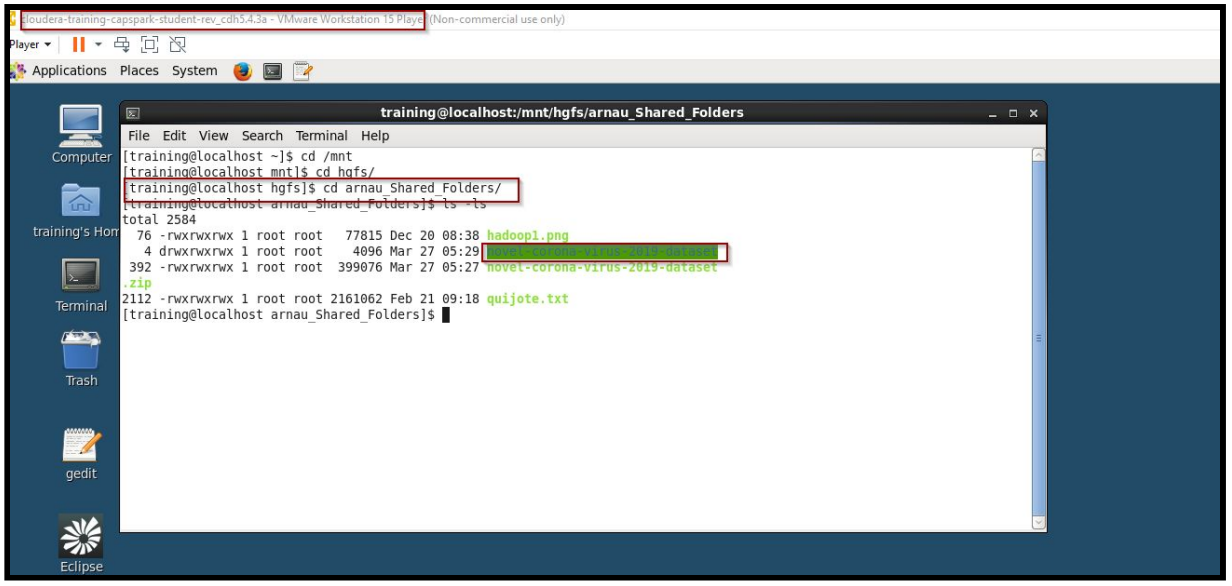
Nom i Cognoms **Data**

Arnau Subirós Puigarnau 21-04-2020

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



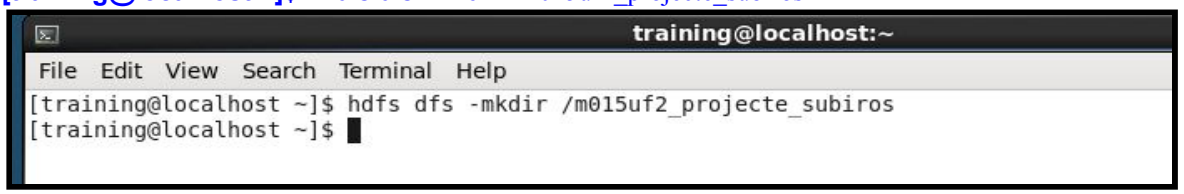




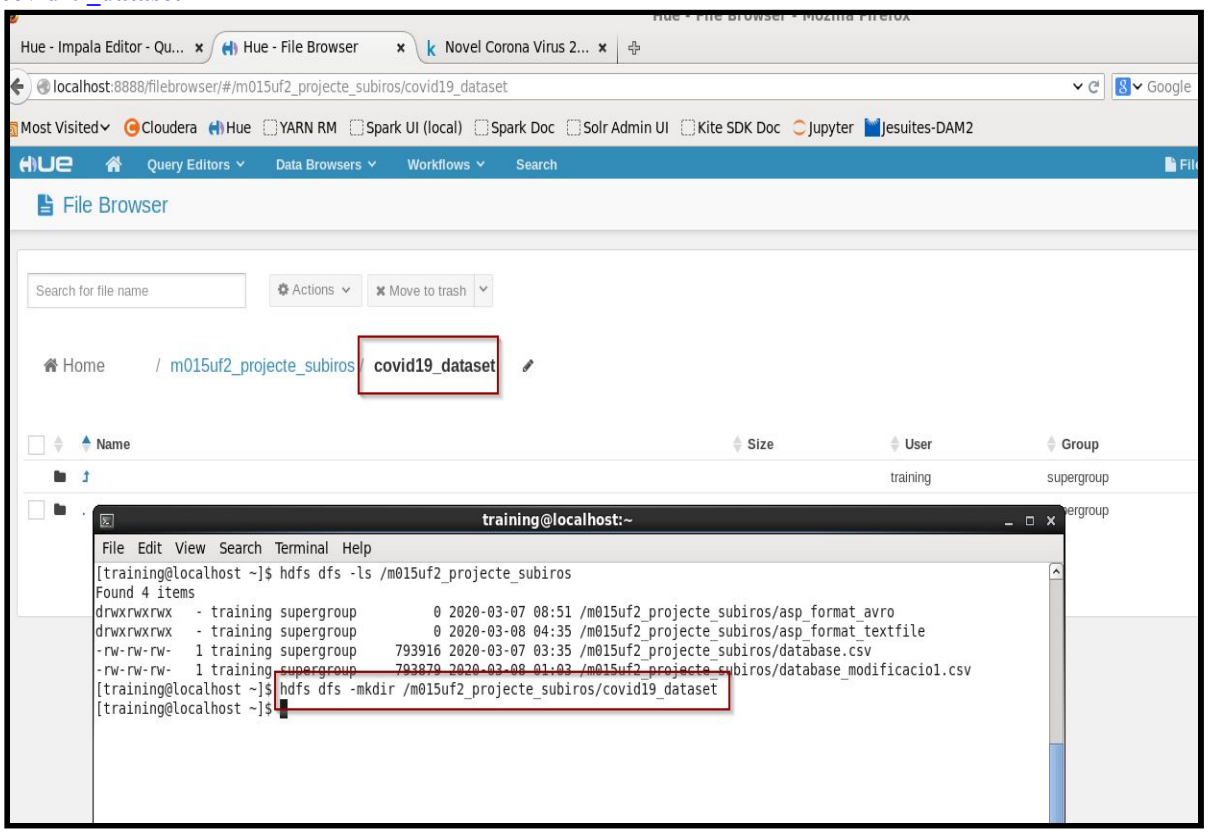
Nom i Cognoms	Data
Arnau Subirós Puigarnau	21-04-2020

- 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



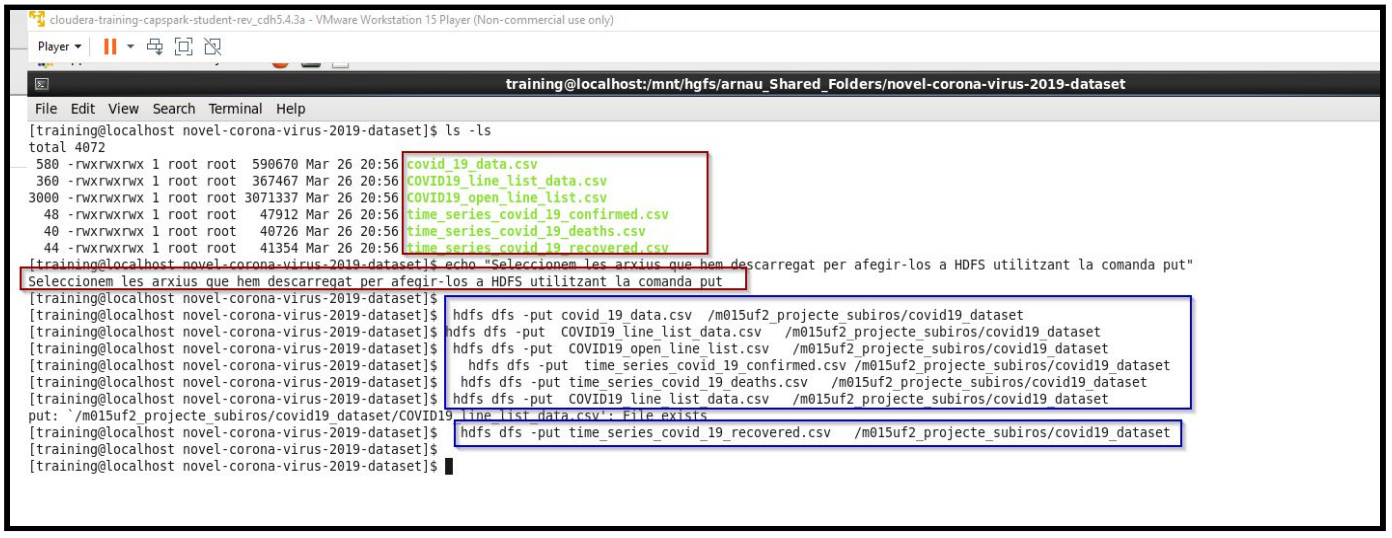
- Com que encara no vull eliminar las proves que he fet amb l'anterior dataset, creare una directori anomenat **covid19_dataset**



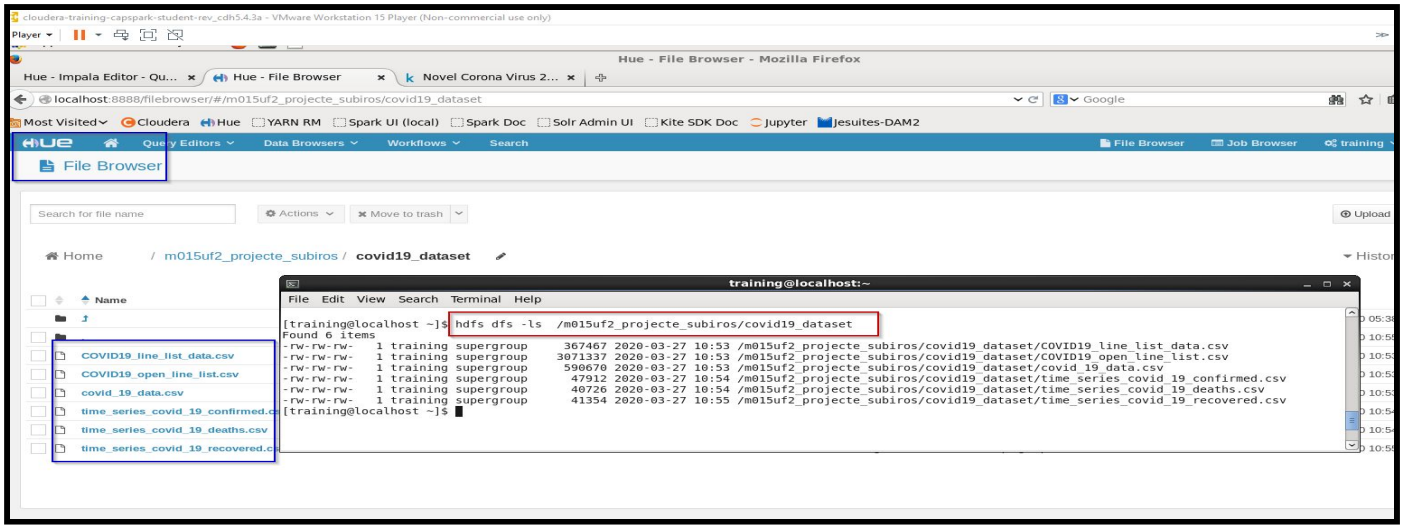
Nom i Cognoms	Data
Arnau Subirós Puigarnau	21-04-2020

- Inserció de dades a HDFS

```
[training@localhost ~]$ hdfs dfs -put covid_19_data.csv /m015uf2_projecte_subiros/covid19_dataset
[training@localhost ~]$ hdfs dfs -put COVID19_line_list_data.csv /m015uf2_projecte_subiros/covid19_dataset
[training@localhost ~]$ hdfs dfs -put COVID19_open_line_list.csv /m015uf2_projecte_subiros/covid19_dataset
[training@localhost ~]$ hdfs dfs -put time_series_covid_19_confirmed.csv /m015uf2_projecte_subiros/covid19_dataset
[training@localhost ~]$ hdfs dfs -put time_series_covid_19_deaths.csv /m015uf2_projecte_subiros/covid19_dataset
[training@localhost ~]$ hdfs dfs -put time_series_covid_19_recovered.csv /m015uf2_projecte_subiros/covid19_dataset
```



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



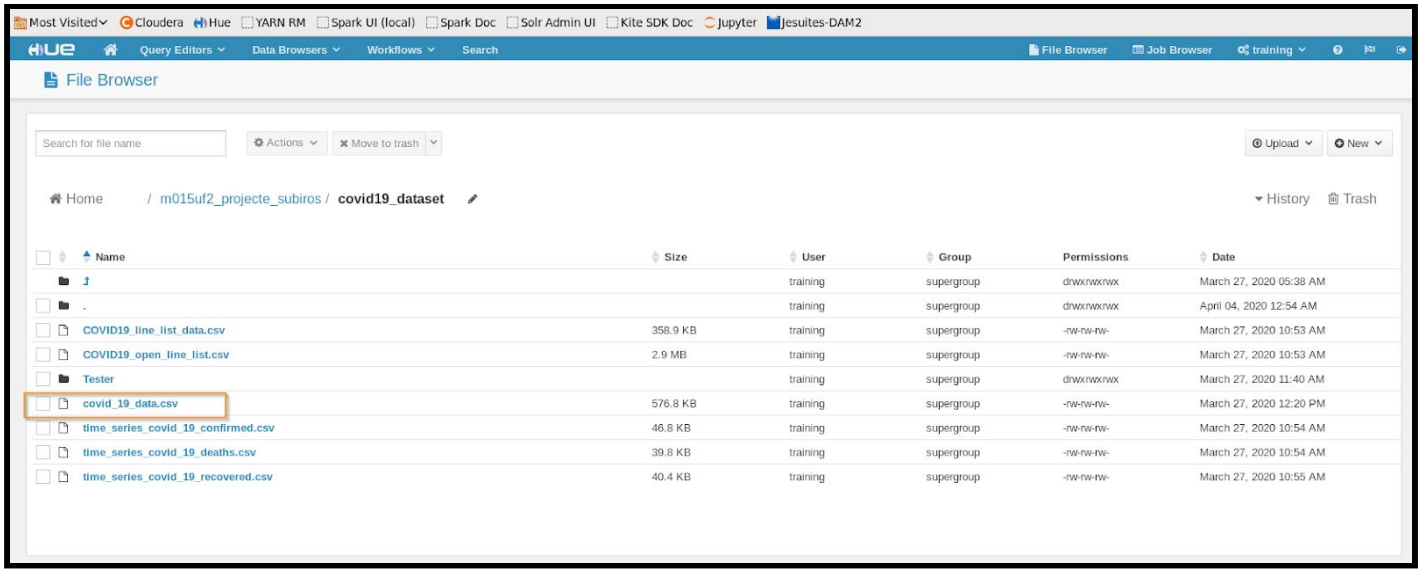
Nom i Cognoms **Data**

Arnau Subirós Puigarnau 21-04-2020

• FORMAT : TEXTFILE

Ens basarem en l' arxiu : covid_19_data.csv

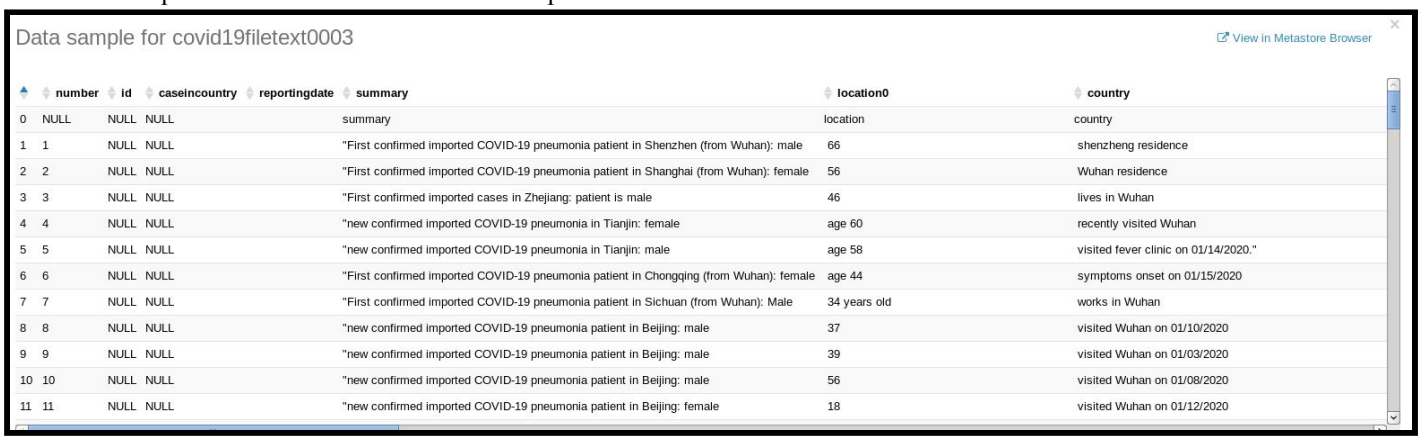
```
CREATE EXTERNAL  TABLE COVID19TESTER3 (
Serial int,
ObserbationDate String,
Province String ,
State String,
Country String,
Regiom String,
LastUpdate String,
Confirmed float,
Deaths float,
Recovered float
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','

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



Nom i Cognoms **Data**

Arnau Subirós Puigarnau 21-04-2020

- Com que hi ha diversos arxius la consulta produeix errors

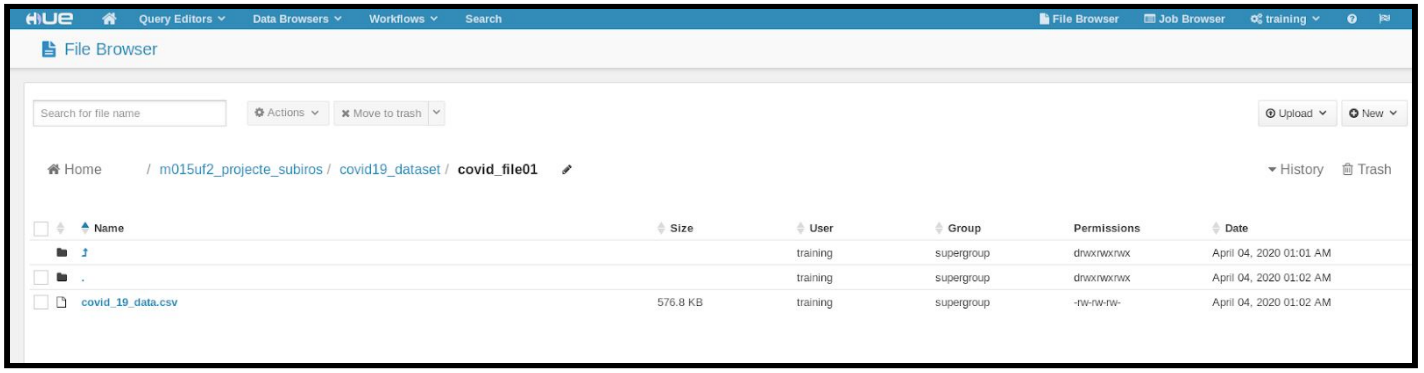


Data sample for covid19filetext0003

View in Metastore Browser

	number	id	caseincountry	reportingdate	summary	location0	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 subdirectori anomenat covid_file01 fem una copia de l'arxiu covid_19_data.csv i el dipositem dins



HUE Query Editors Data Browsers Workflows Search File Browser Job Browser training

File Browser

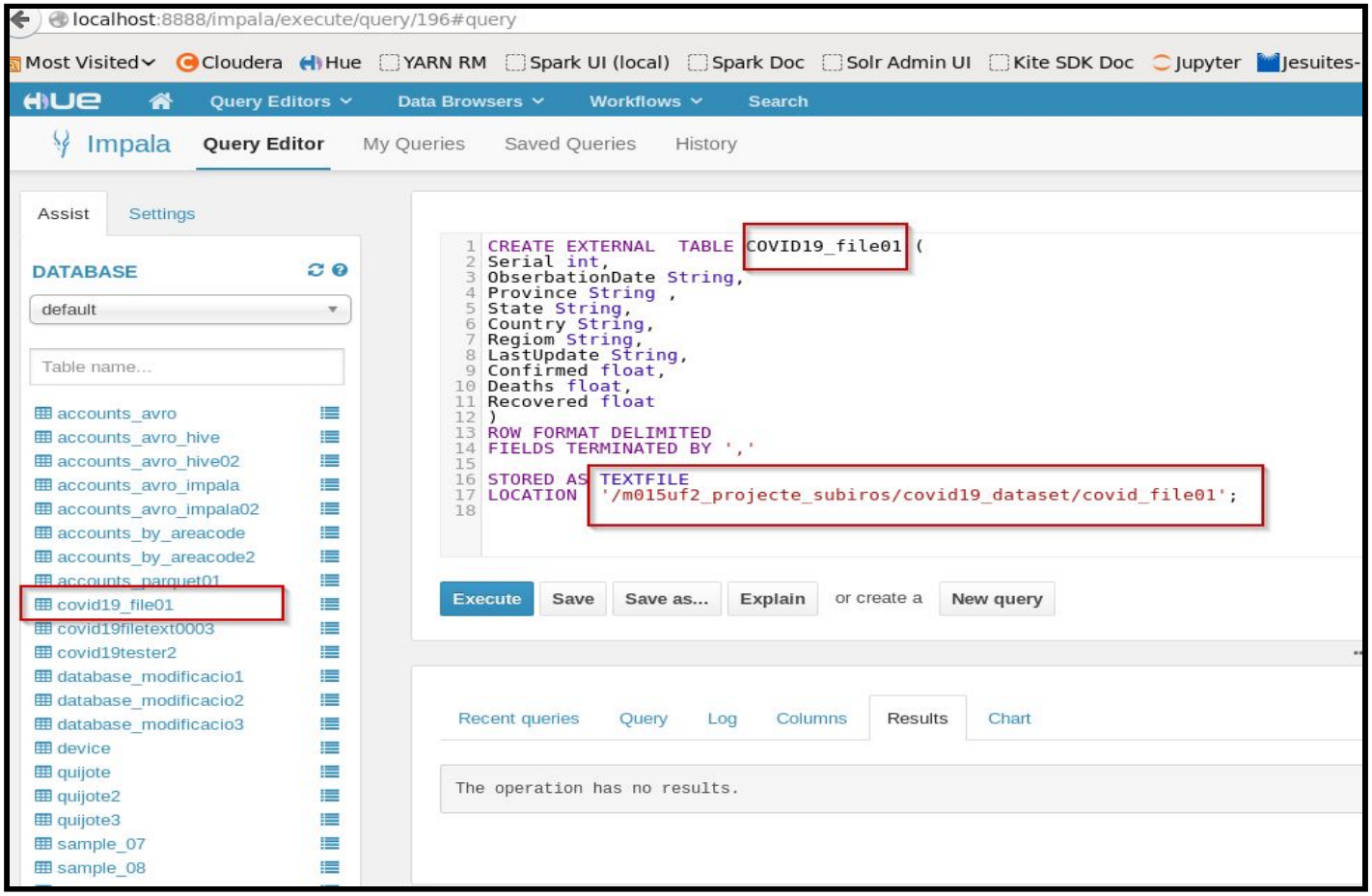
Search for file name Actions Move to trash Upload New

Home / m015uf2_projecte_subiros / covid19_dataset / covid_file01 History Trash

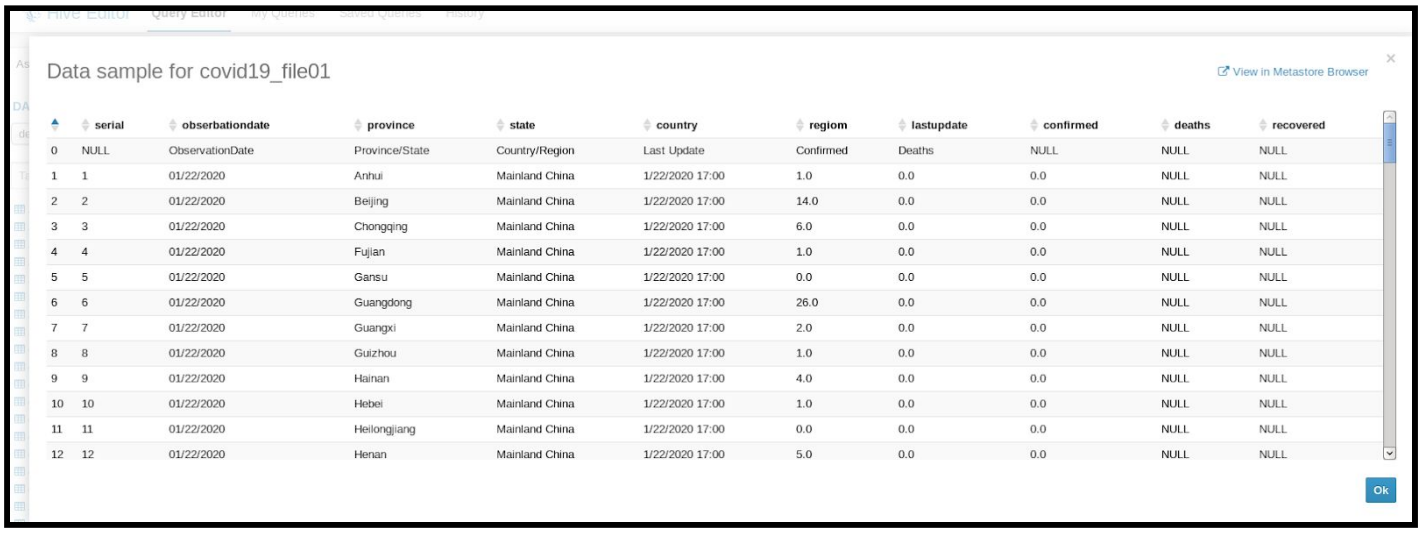
Name	Size	User	Group	Permissions	Date
..		training	supergroup	drwxrwxrwx	April 04, 2020 01:01 AM
.		training	supergroup	drwxrwxrwx	April 04, 2020 01:02 AM
covid_19_data.csv	576.8 KB	training	supergroup	-rw-rw-rw-	April 04, 2020 01:02 AM

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

```
CREATE EXTERNAL  TABLE COVID19_File01 (
Serial int,
ObserbationDate String,
Province String ,
State String,
Country String,
Regiom 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';
```

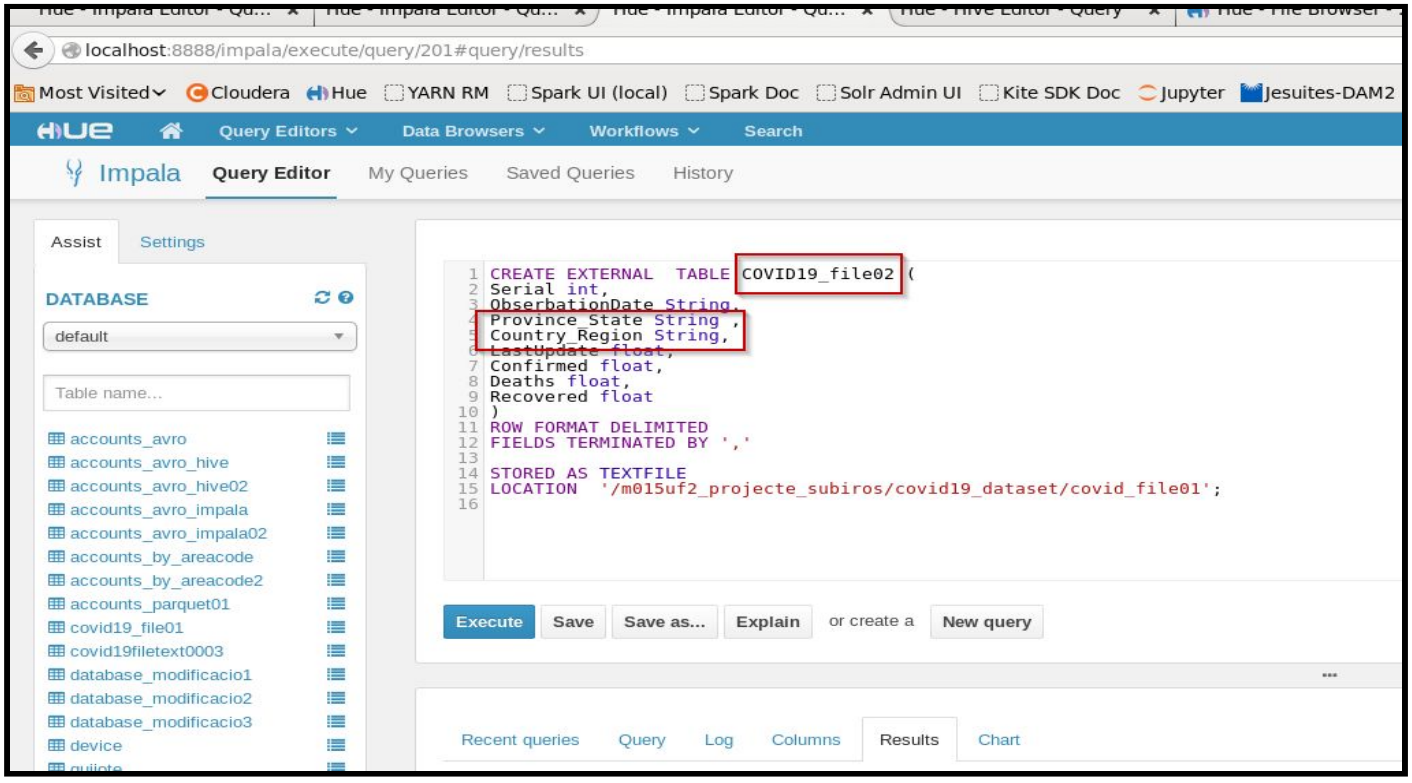


Nom i Cognoms	Data
Arnau Subirós Puigarnau	21-04-2020



	serial	obserbationdate	province	state	country	regiom	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

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



Nom i Cognoms

Arnau Subirós Puigarnau

Data

21-04-2020

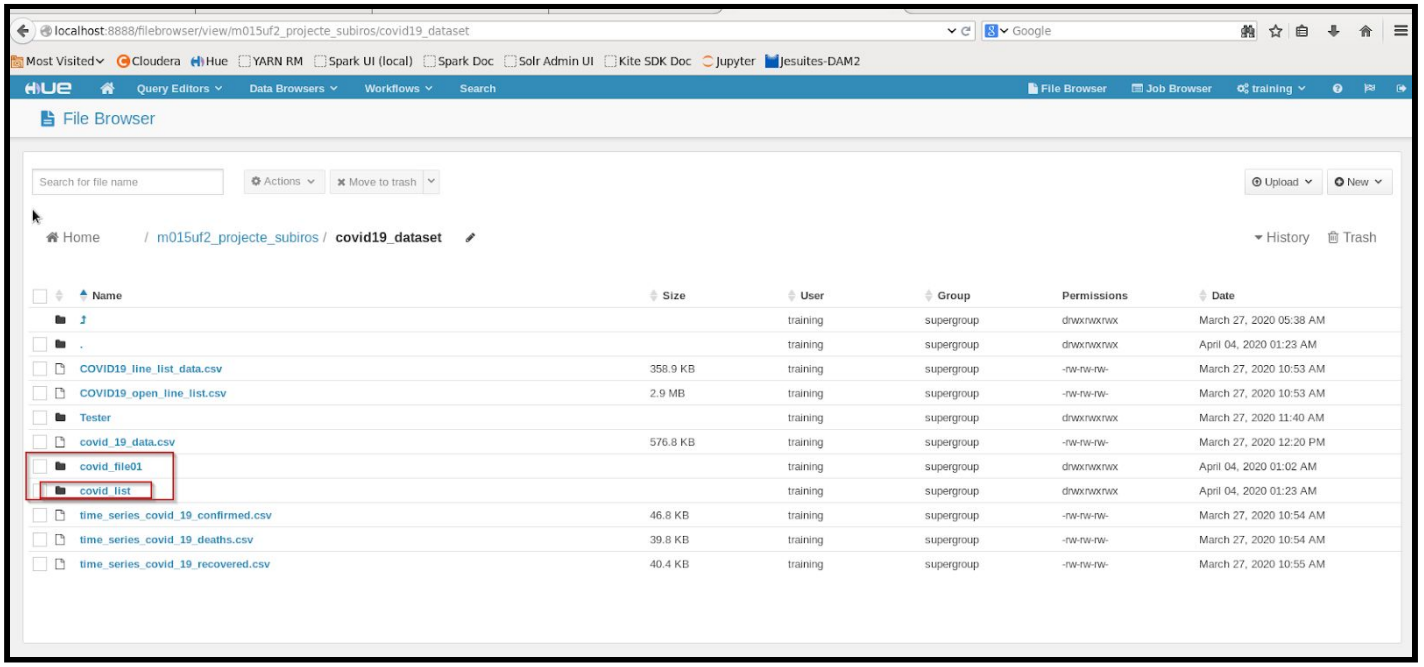
Data sample for covid19_file02

View in Metastore Browser

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

Ok

- 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 subdirectori anomenat covid_list(com he fet abans amb aquest arxiu)



localhost:8888/filebrowser/view/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 | Query Editors | Data Browsers | Workflows | Search | File Browser | Job Browser | training

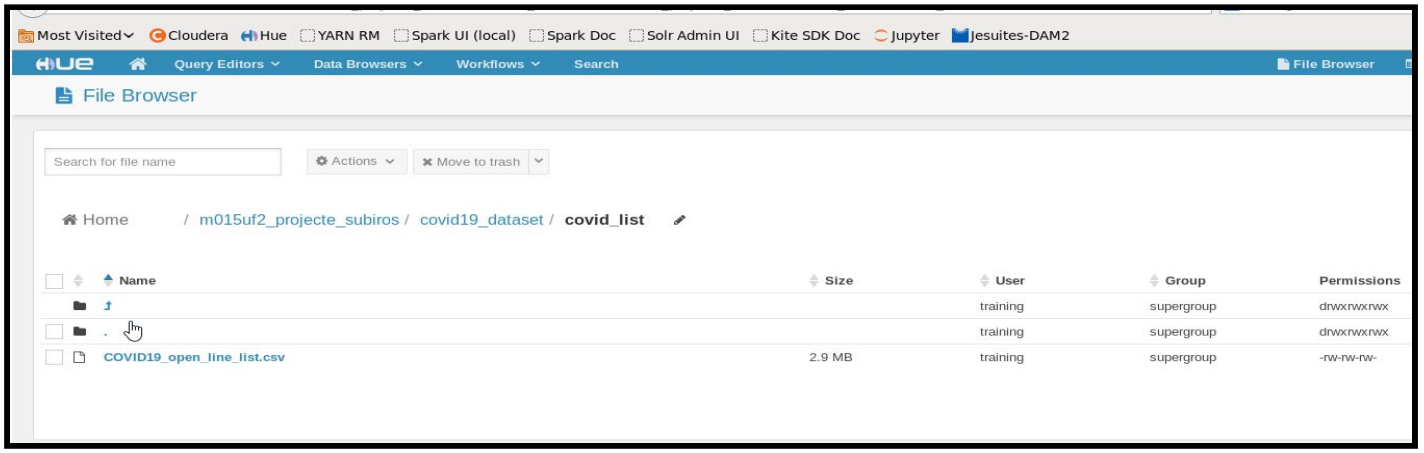
File Browser

Search for file name | Actions | Move to trash | Upload | New

Home / m015uf2_projecte_subiros / covid19_dataset | History | Trash

Name	Size	User	Group	Permissions	Date
..		training	supergroup	drwxrwxrwx	March 27, 2020 05:38 AM
.		training	supergroup	drwxrwxrwx	April 04, 2020 01:23 AM
COVID19_line_list_data.csv	358.9 KB	training	supergroup	-rw-rw-rw-	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

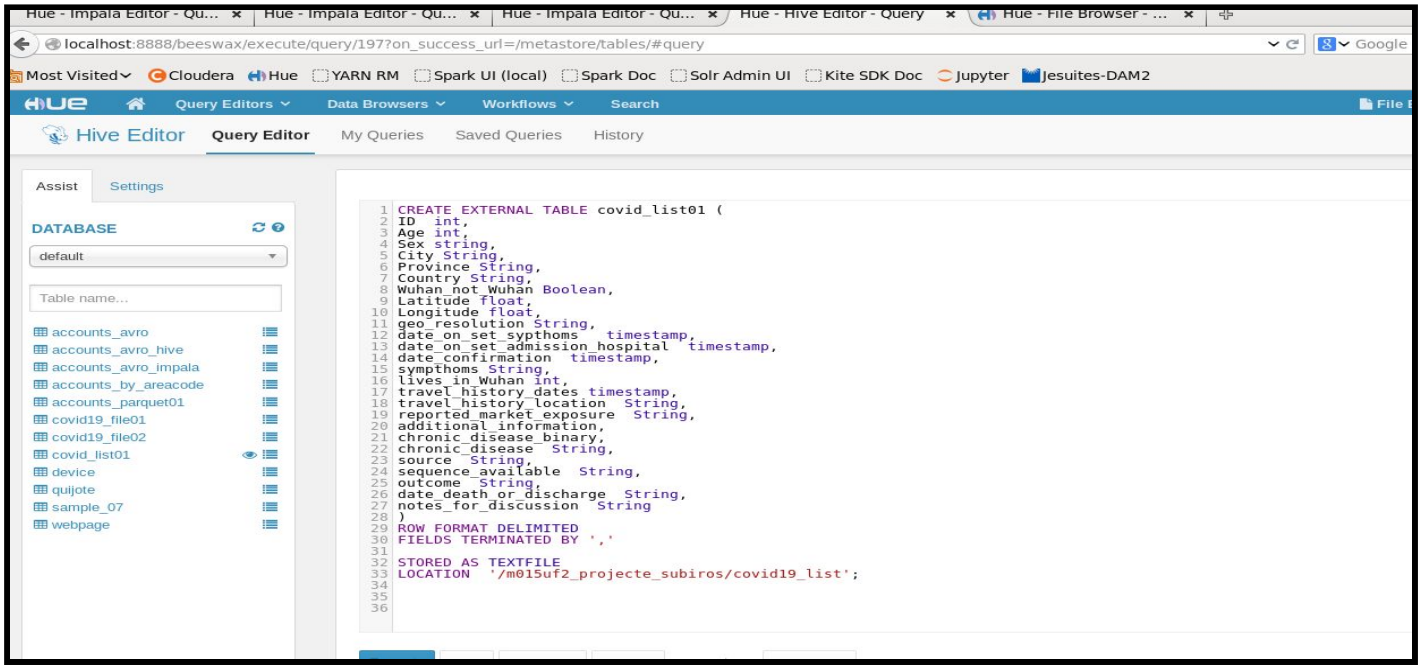
Nom i Cognoms	Data
Arnau Subirós Puigarnau	21-04-2020



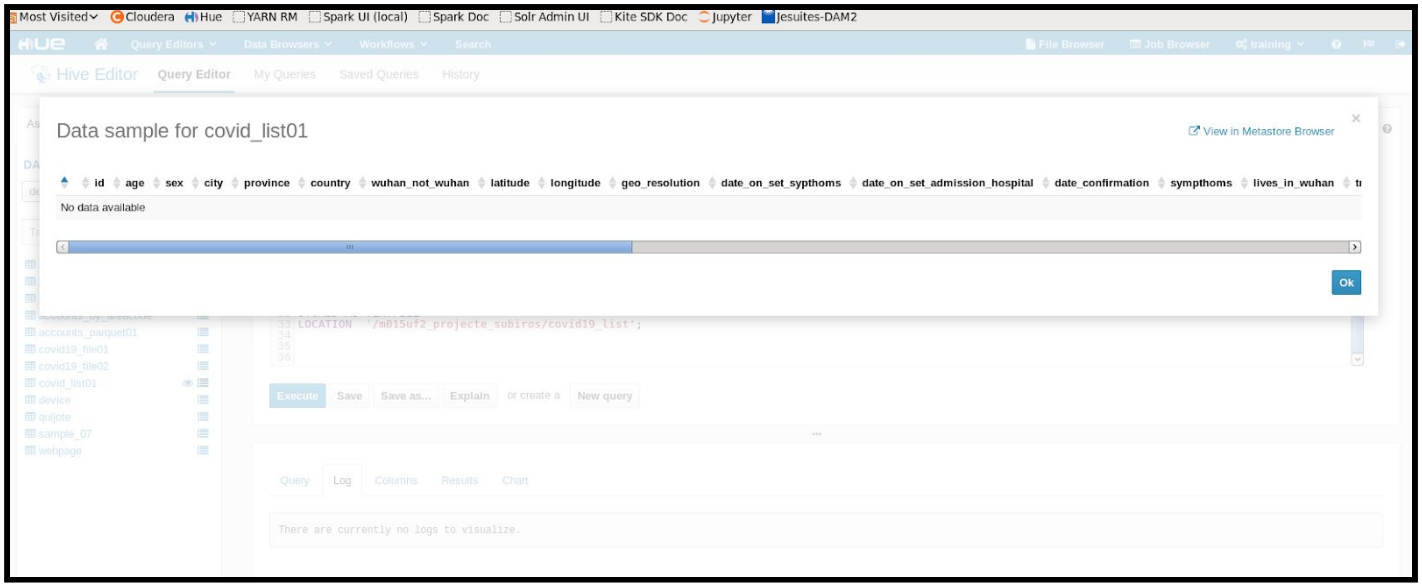
Ens basarem en l' arxiu : COVID19_open_line_list.csv

```
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_sypthoms   timestamp,
date_on_set_admission_hospital  timestamp,
date_confirmation  timestamp,
sympthoms 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';
```

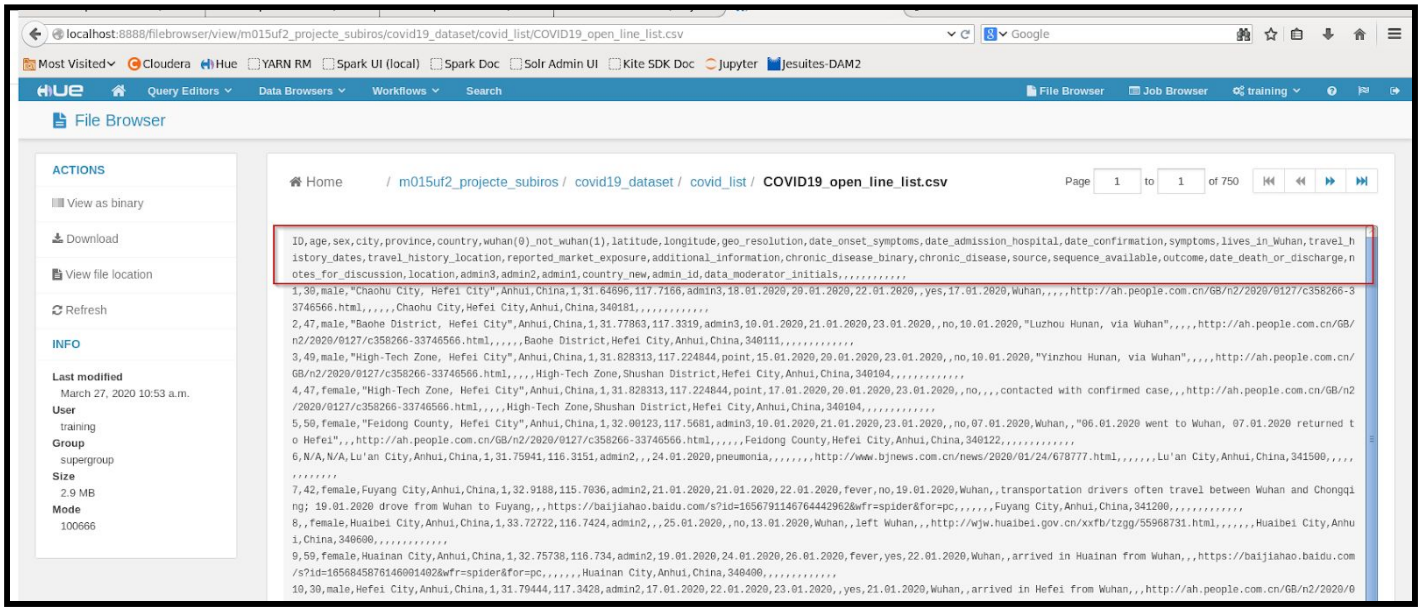
Nom i Cognoms	Data
Arnau Subirós Puigarnau	21-04-2020



Pendent de solucionar el problema



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



Nom i Cognoms

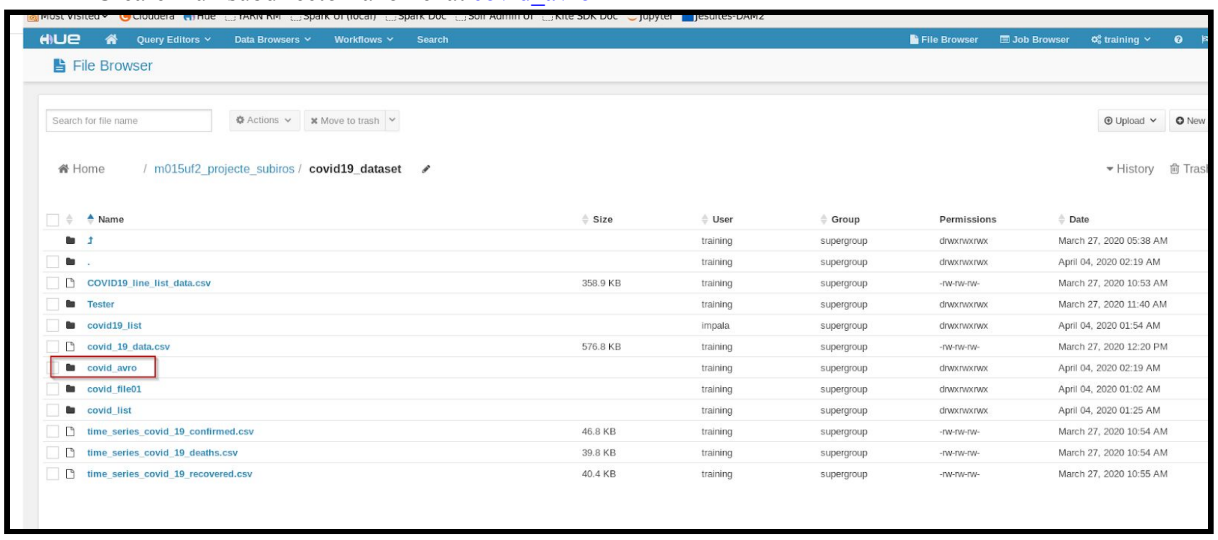
Arnau Subirós Puigarnau

Data

21-04-2020

• FORMAT : AVRO

- Ens basarem en l' arxiu : covid_19_data.csv
- Crearem un subdirectori anomenat covid_avro



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

1. Opció 1

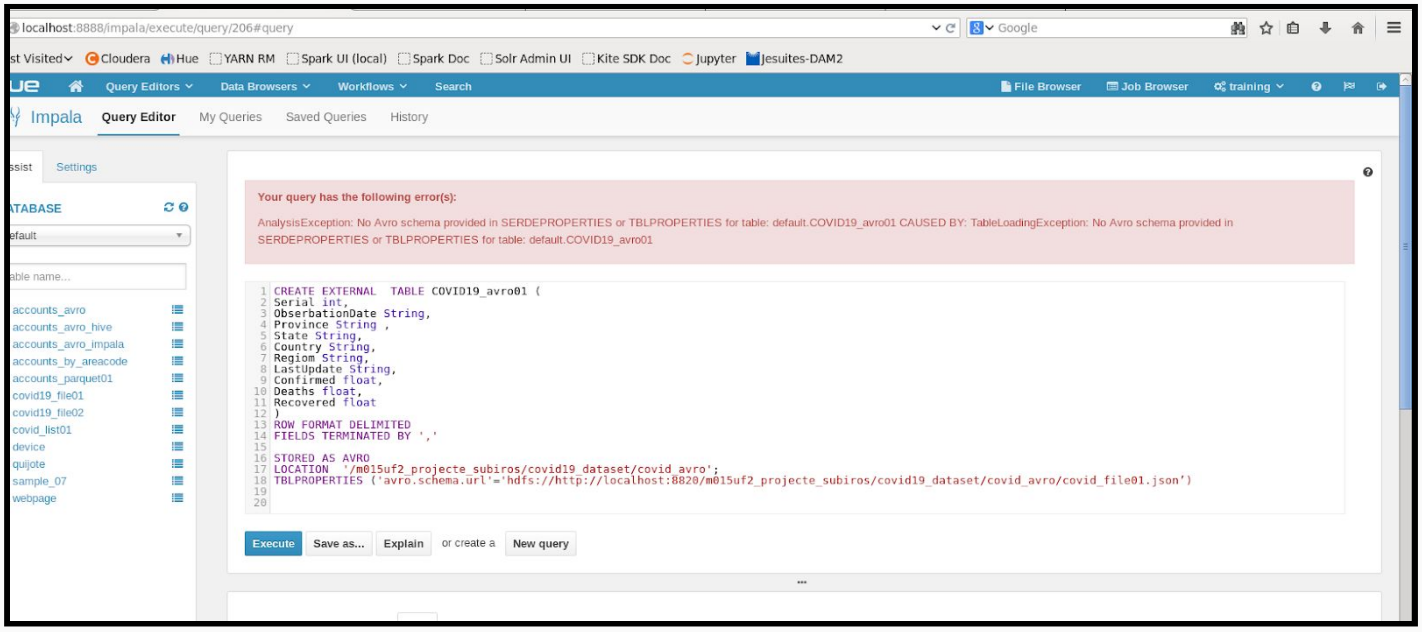
```
CREATE EXTERNAL TABLE database_covid9_avro1 (
 (Serial int, ObserbationDate String, Province String, State String, Country, Region String, LastUpdate String, Confirmed float, Deaths float, Recovered float)
 STORED AS AVRO
LOCATION  '/m015uf2_projecte_subiros/covid19_dataset/covid_avro'
TBLPROPERTIES ('avro.schema.literal' '=

'{"name": "database_covid9_avro1",
  "type": "record",
   "fields": [
            {"name":"Serial", "type":"int"},
            {"name":"ObserbationDate", "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,
ObserbationDate String,
Province String ,
State String,
Country String,
Regiom 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



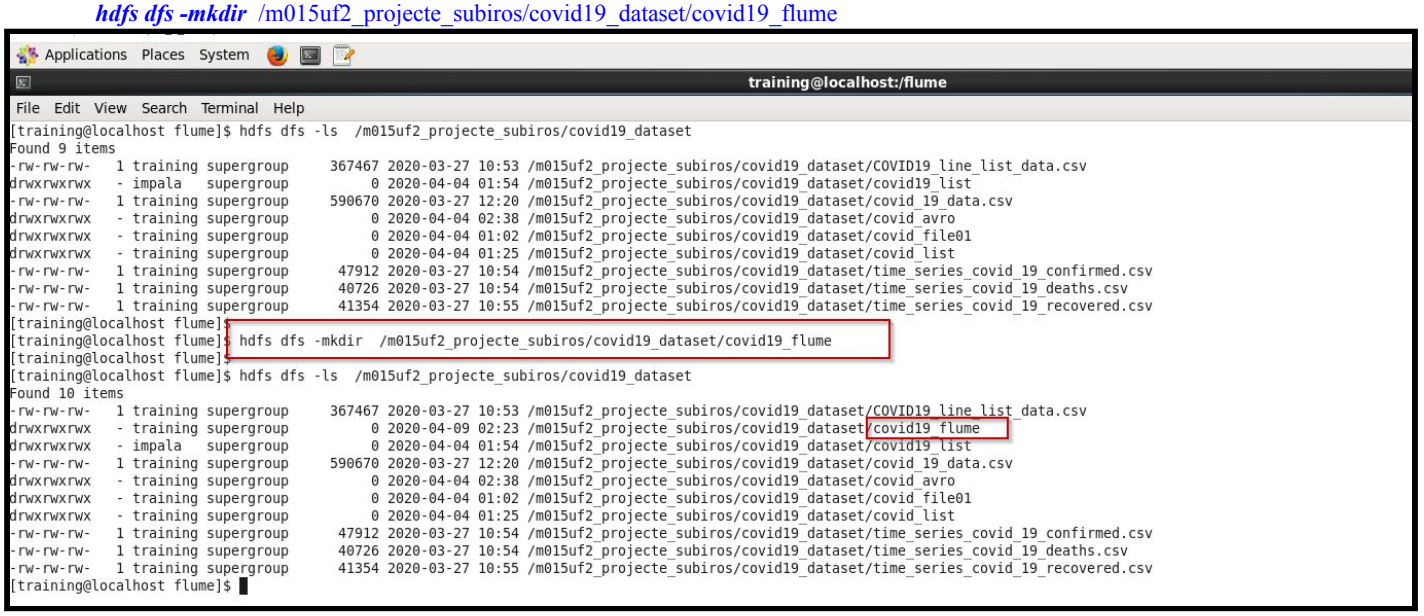
Nom i Cognoms **Data**

Arnau Subirós Puigarnau 21-04-2020

2. Recollida de dades en temps real

• *Flume Sources :Spooldir*

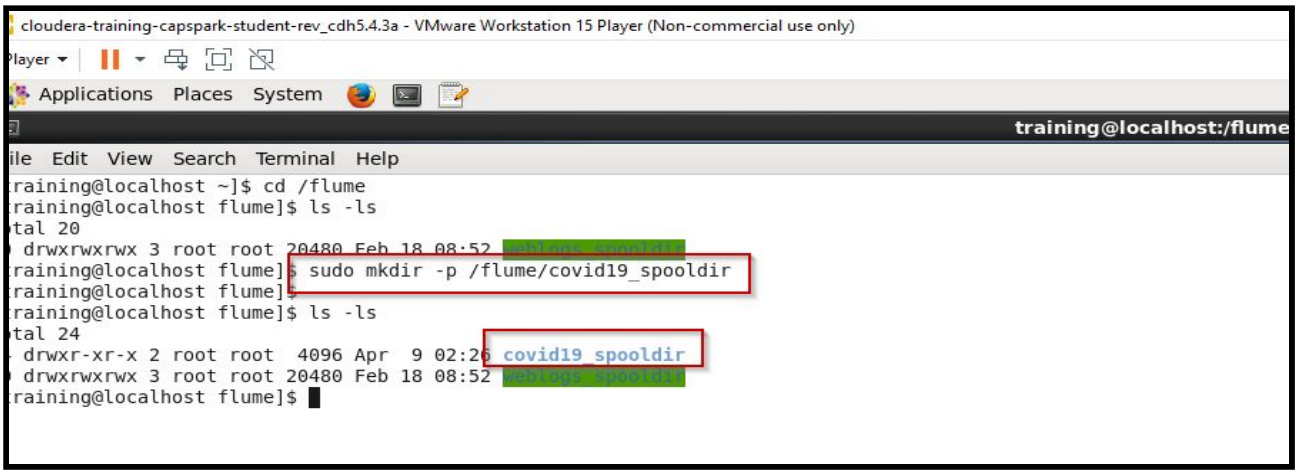
- Crearem un directori per la ingesta de dades a HDFS
  hdfs dfs -mkdir /m015uf2_projecte_subiros/covid19_dataset/covid19_flume



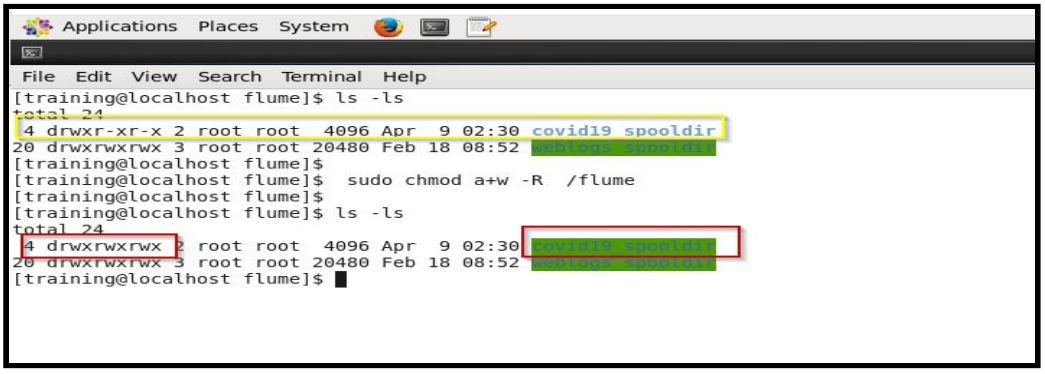
- Crearem un directori local(no al HDFS) per el log de sortid del web server
  [training@localhost flume]$ sudo mkdir -p /flume/covid19_spooldir

Nom i Cognoms **Data**

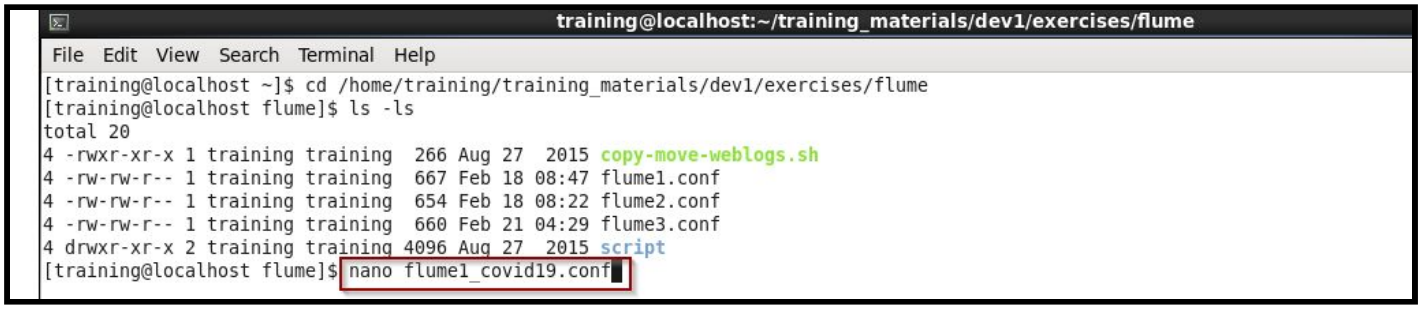
Arnau Subirós Puigarnau 21-04-2020



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

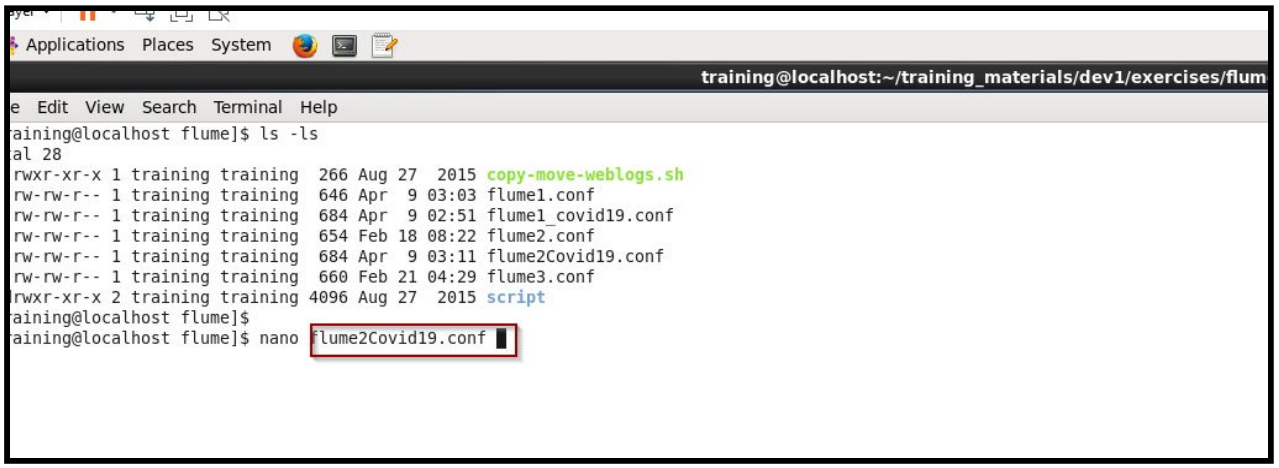


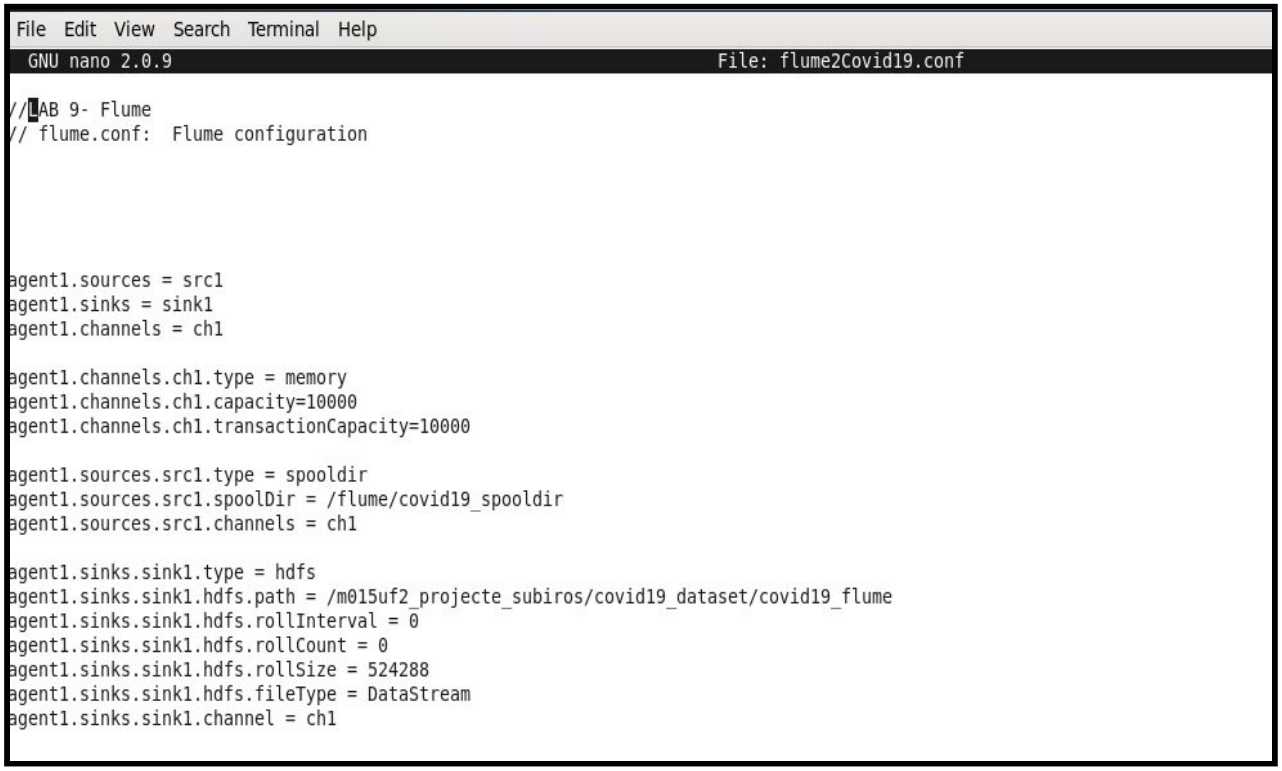
- Configuració de Flume
  - Creació de l'arxiu al mateix lloc a on hem fet les pràctiques LAB09 (l'arxiu l'he creat amb una barra baixa i m'ha donat error, n'he tornat a crear un altre : flume2Covid19.conf)



Nom i Cognoms **Data**

Arnau Subirós Puigarnau 21-04-2020

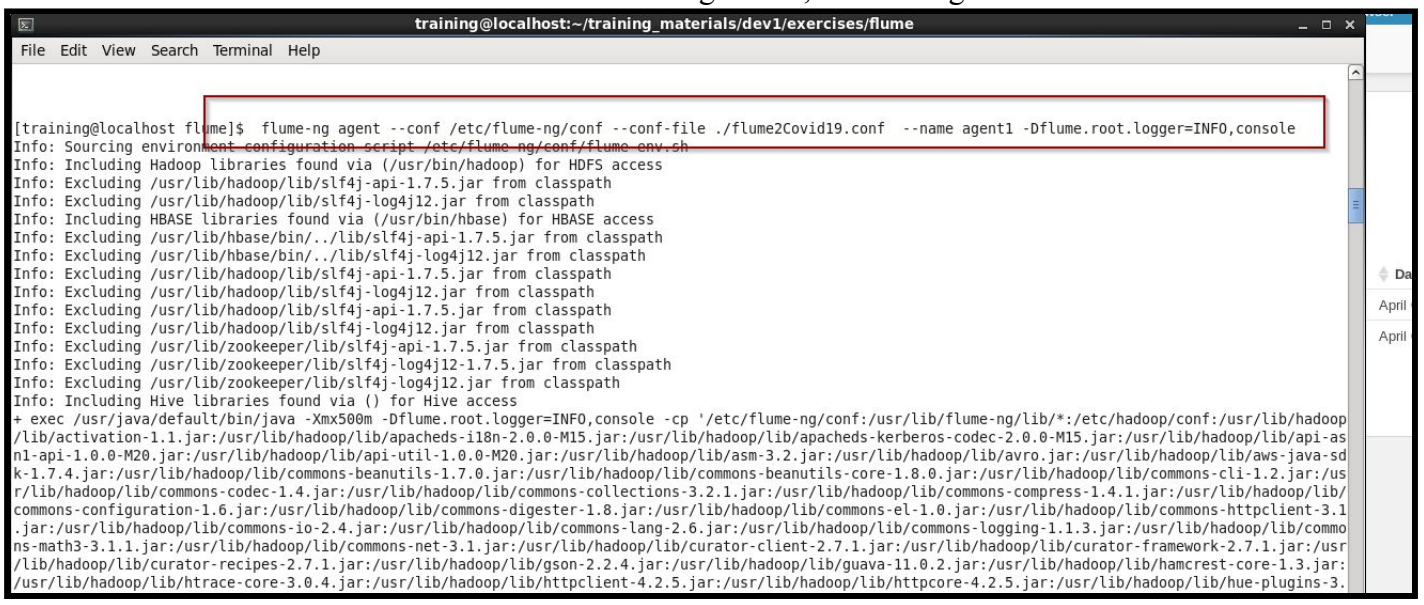


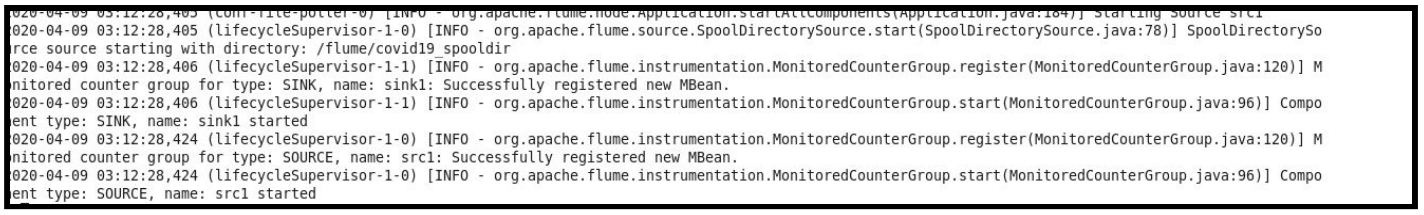


Nom i Cognoms **Data**

Arnau Subirós Puigarnau 21-04-2020

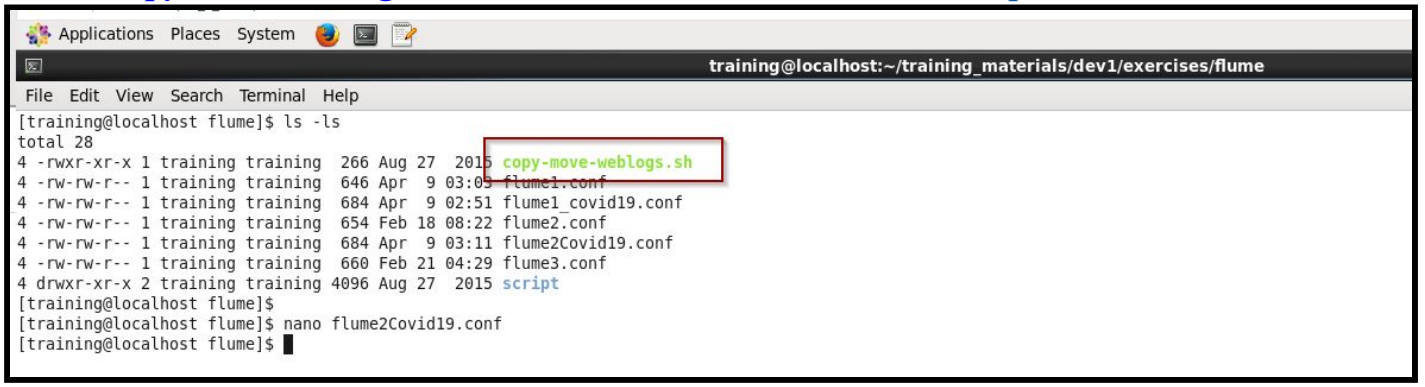
- Desde el directori a on he creat l'arxiu de configuració ,iniciem l'agent Flume

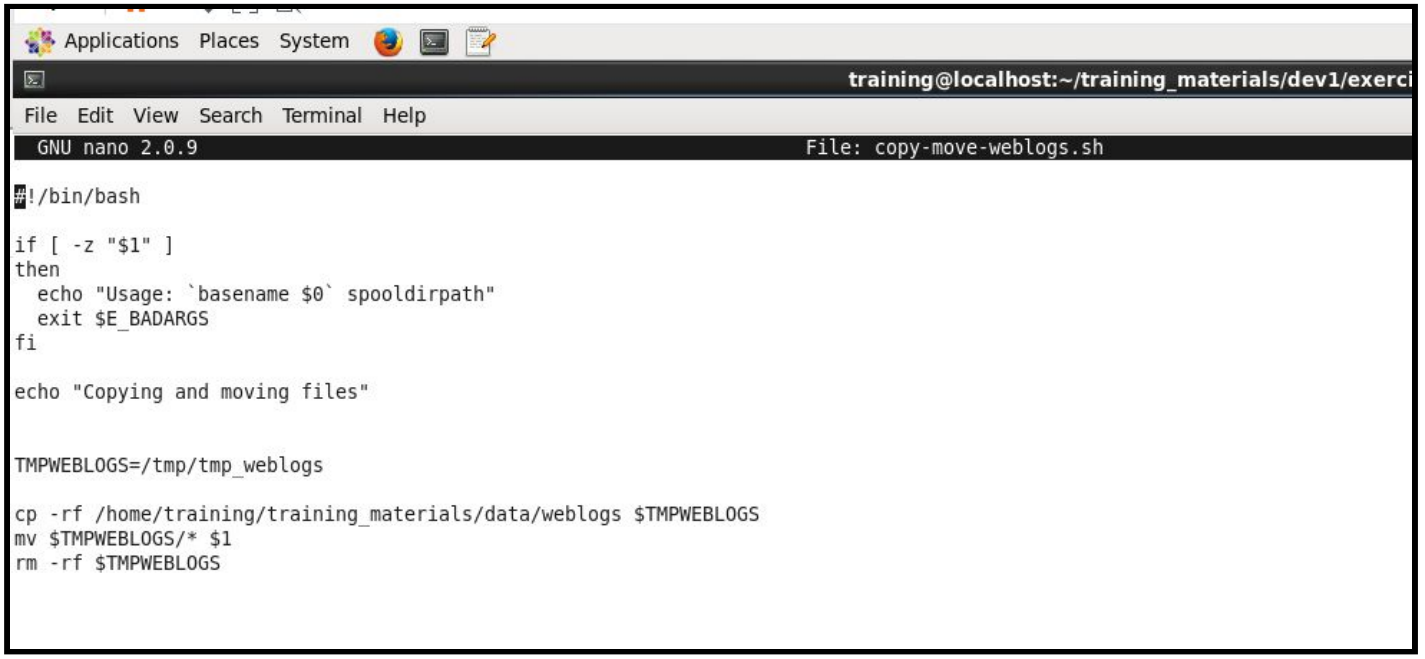




Nom i Cognoms	Data
Arnau Subirós Puigarnau	21-04-2020

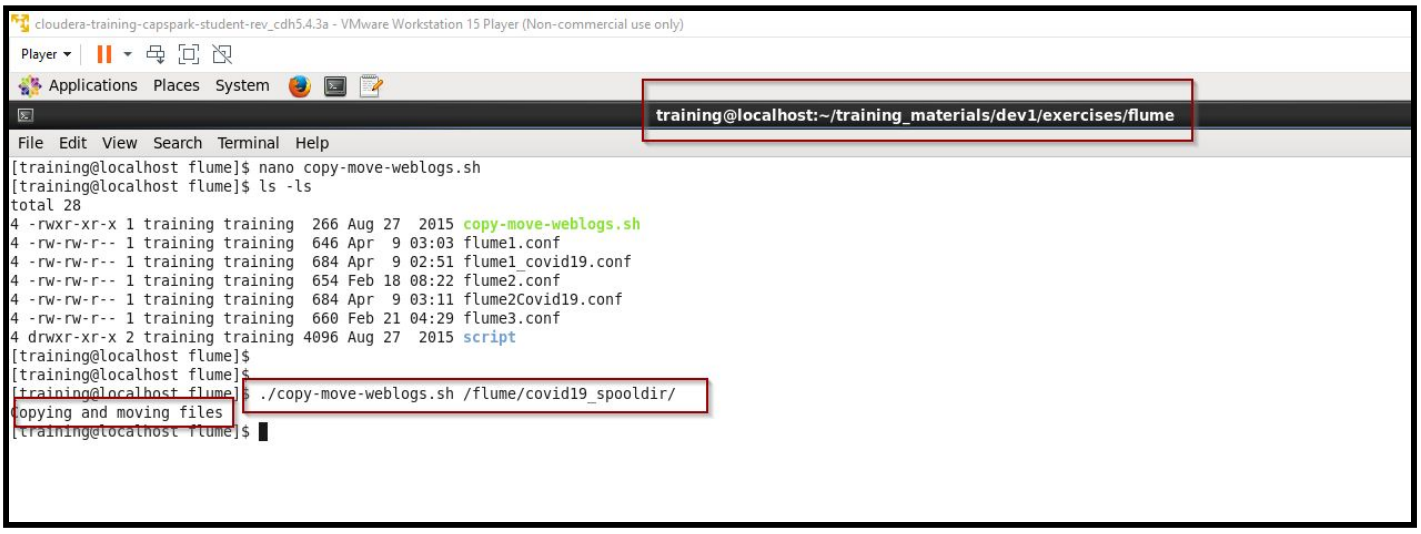
- ***Simular els logs de sortida s d'un Servidor Web mitjançant el script "copy-move-weblogs.sh" de LAB-09 i els mou al directori spooldir***

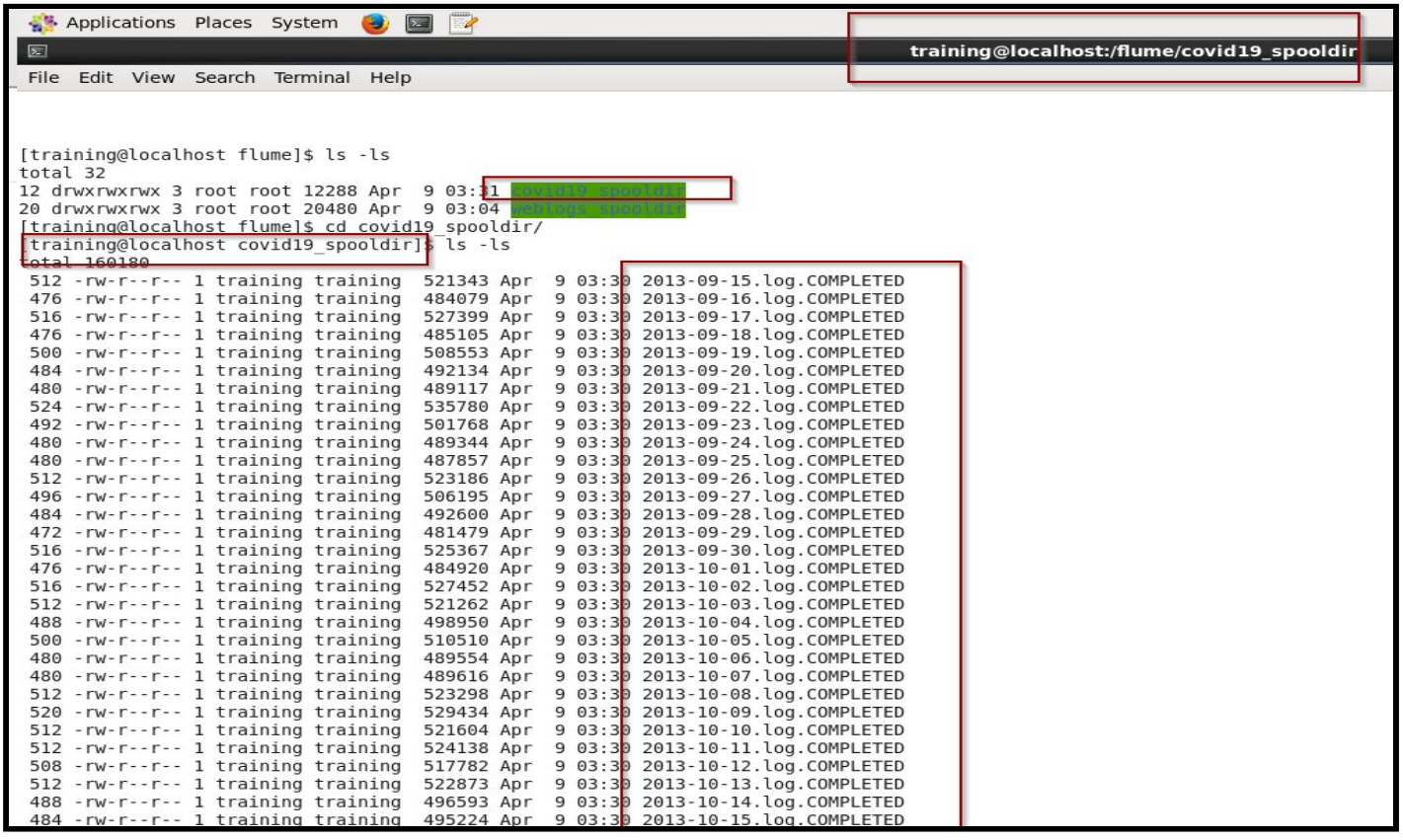




Nom i Cognoms **Data**

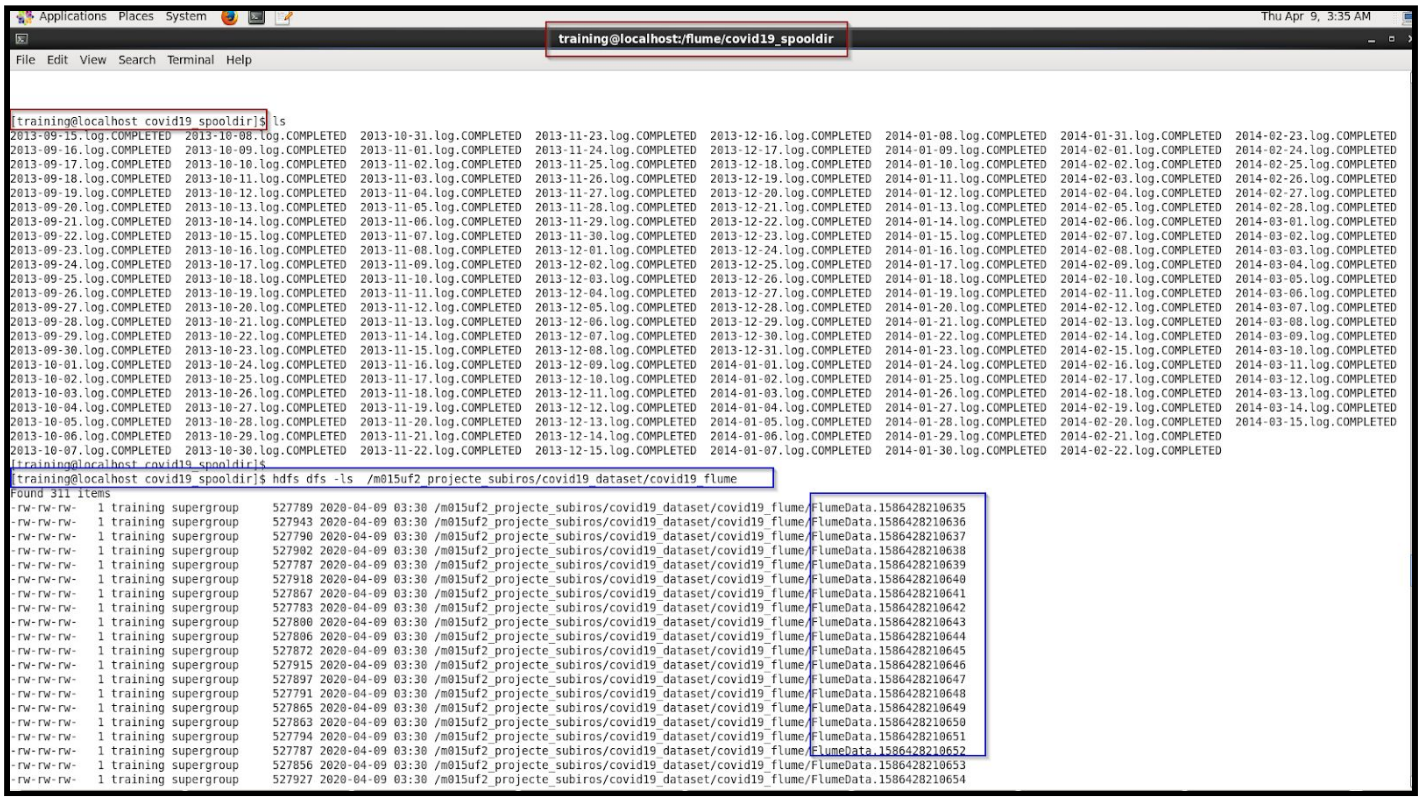
Arnau Subirós Puigarnau 21-04-2020

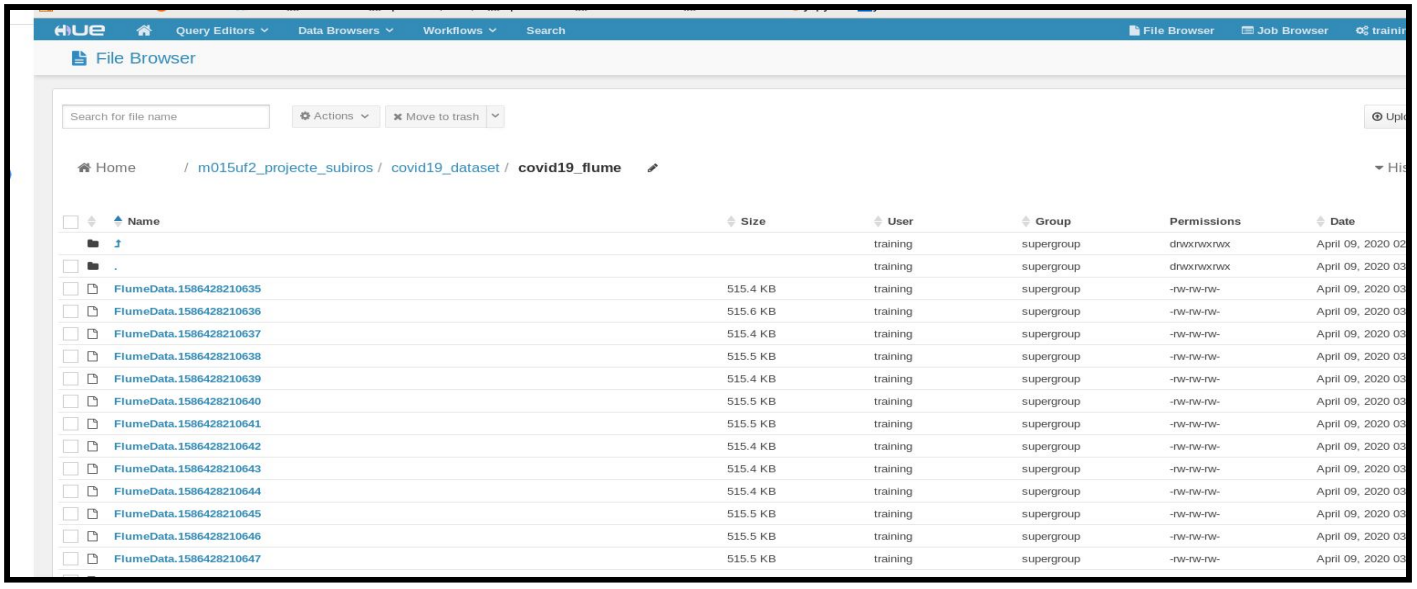




Nom i Cognoms **Data**

Arnau Subirós Puigarnau 21-04-2020





Nom i Cognoms

Arnau Subirós Puigarnau

Data

21-04-2020

- ***Flume Sources :HTTP (tèoric)***

Creacio 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 ...}
 ]
```

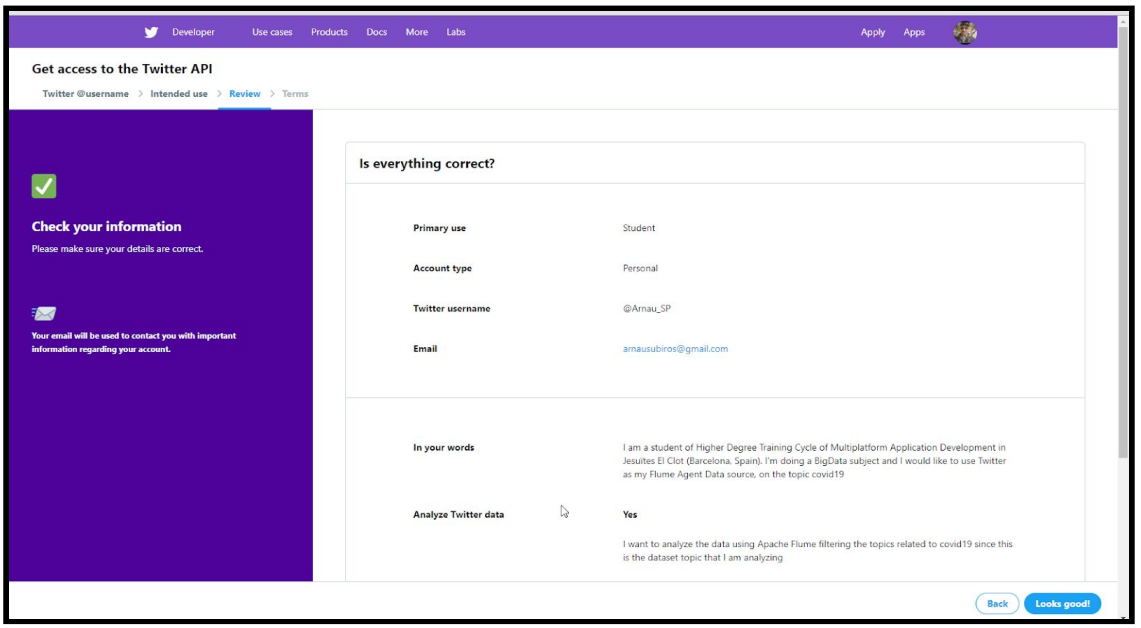
- L'agent HTTP acepta dades en format JSON, s'hauria de provar enviar un cURL amb un formulari com aquest :

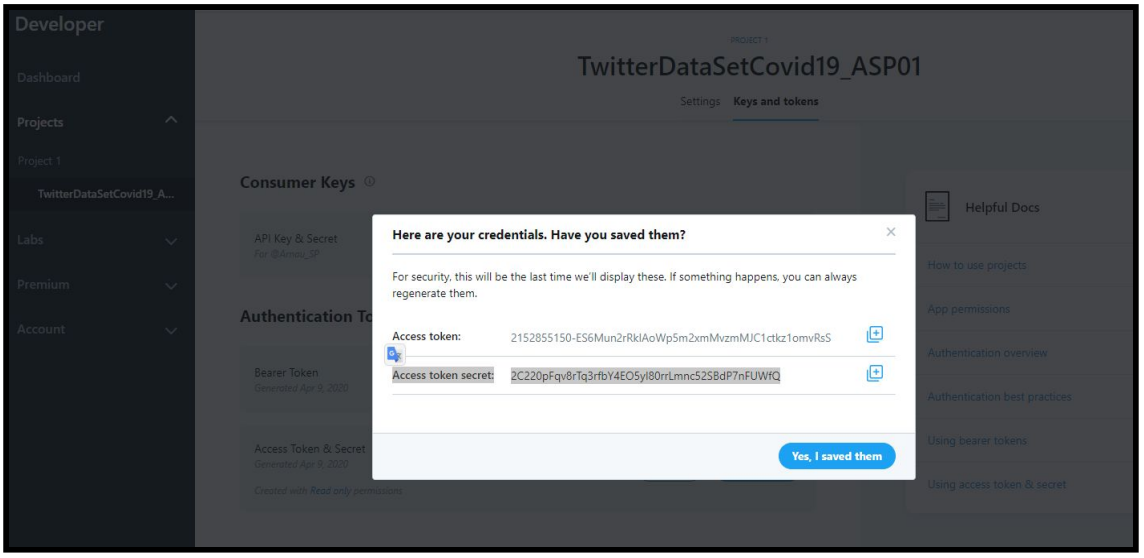
```
curl -X POST -H 'Content-Type: application/json; charset=UTF-8' -d '[{"username":"xrqwrqwryzas","password":"12124sfsfsfas123"}]' http://yourdomain.com:81/
```

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

• *Flume Sources : Twitter*

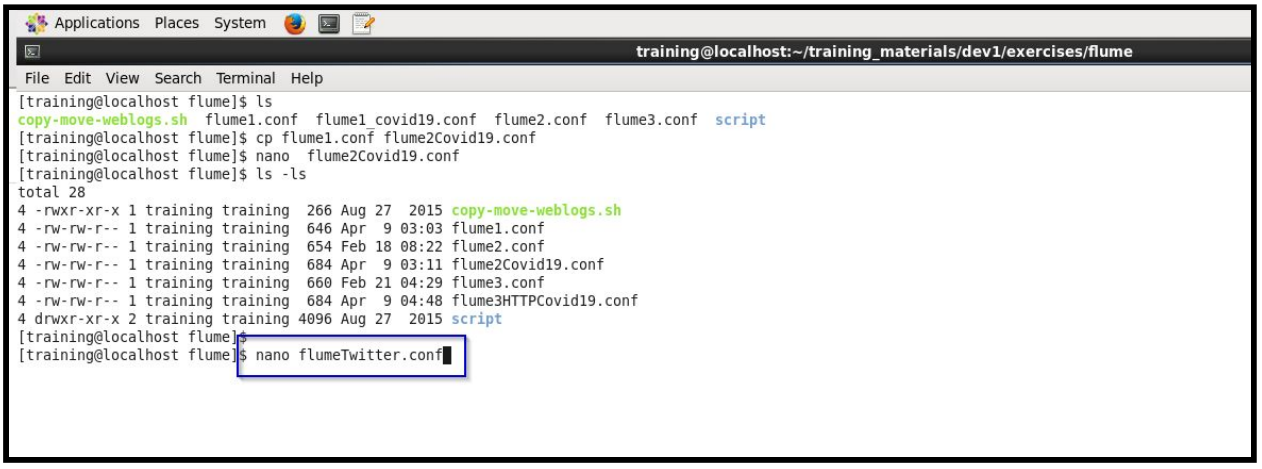
- *Abans de tot hem d'anar a https://developer.twitter.com/ i registrant-nos. Ens preguntaran el motiu*

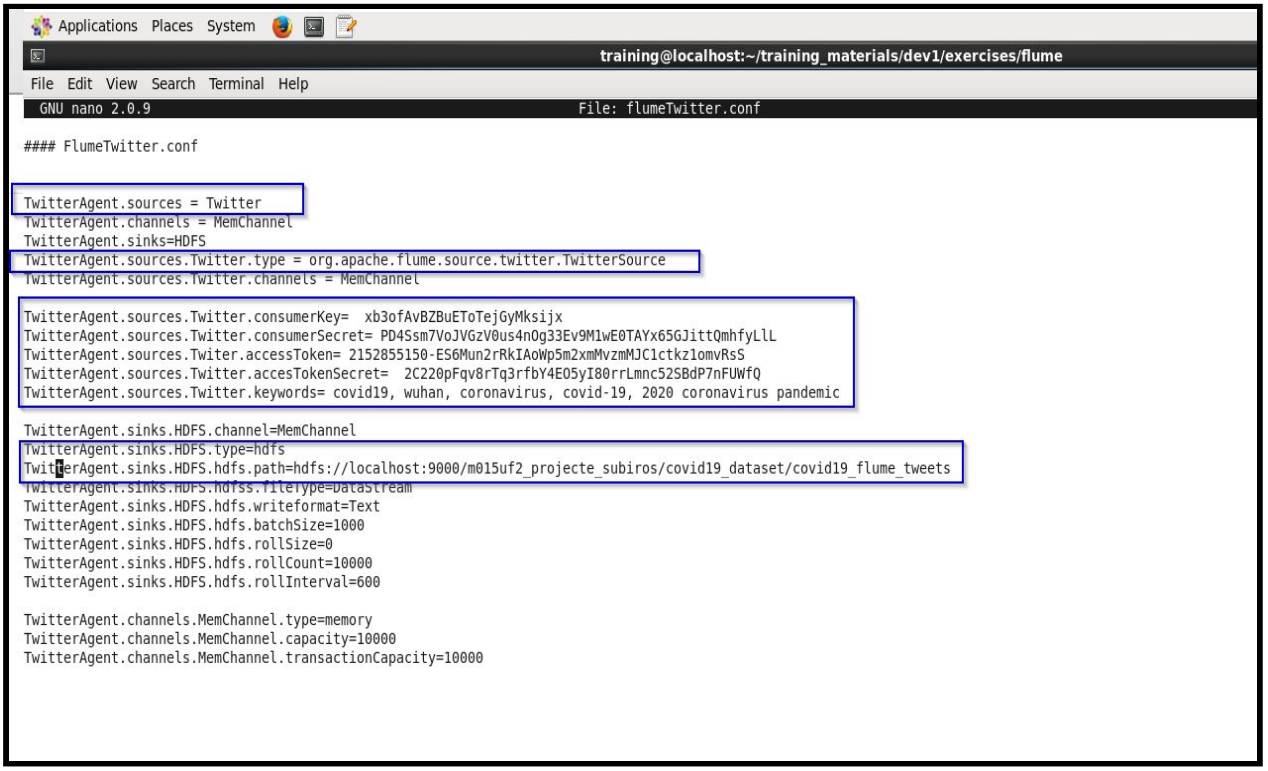




Nom i Cognoms	Data
Arnau Subirós Puigarnau	21-04-2020

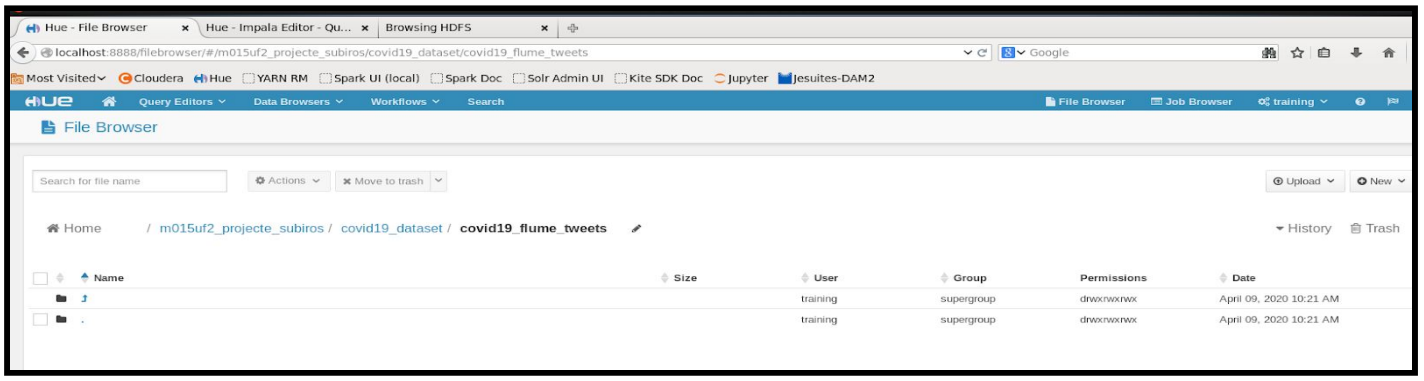
- ***Configurem l'arxiu de configuracio***





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

- ***Creacio d'un nou directori HDFS anomenat covid19_flume_tweets***

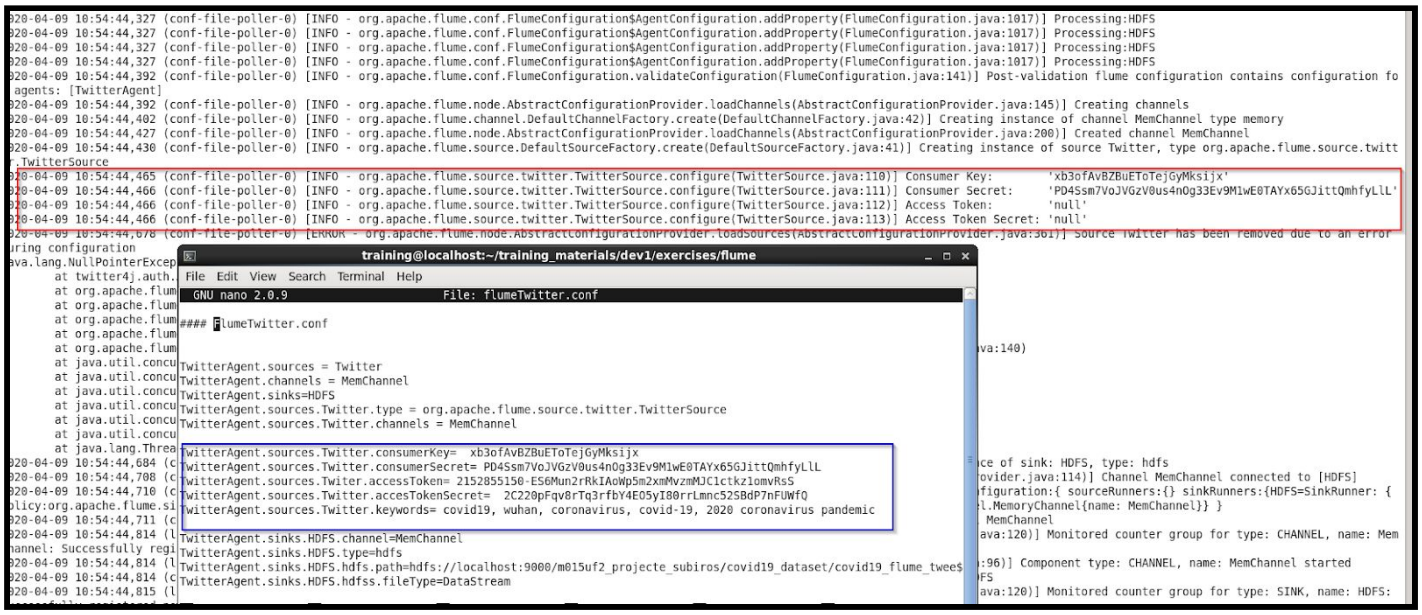


- ***Iniciem l'agent flume***

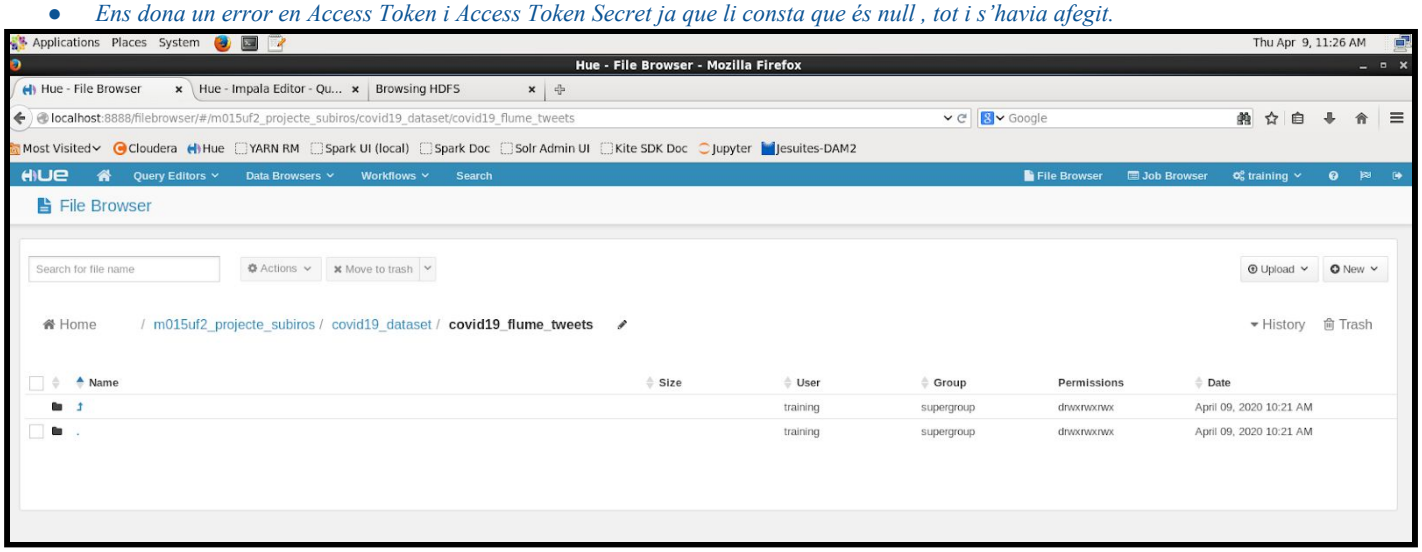
[training@localhost flume]$ *flume-ng agent --conf /etc/flume-ng/conf --conf-file ./flumeTwitter.conf --name TwitterAgent -Dflume.root.logger=INFO,console*

Nom i Cognoms **Data**

Arnau Subirós Puigarnau 21-04-2020



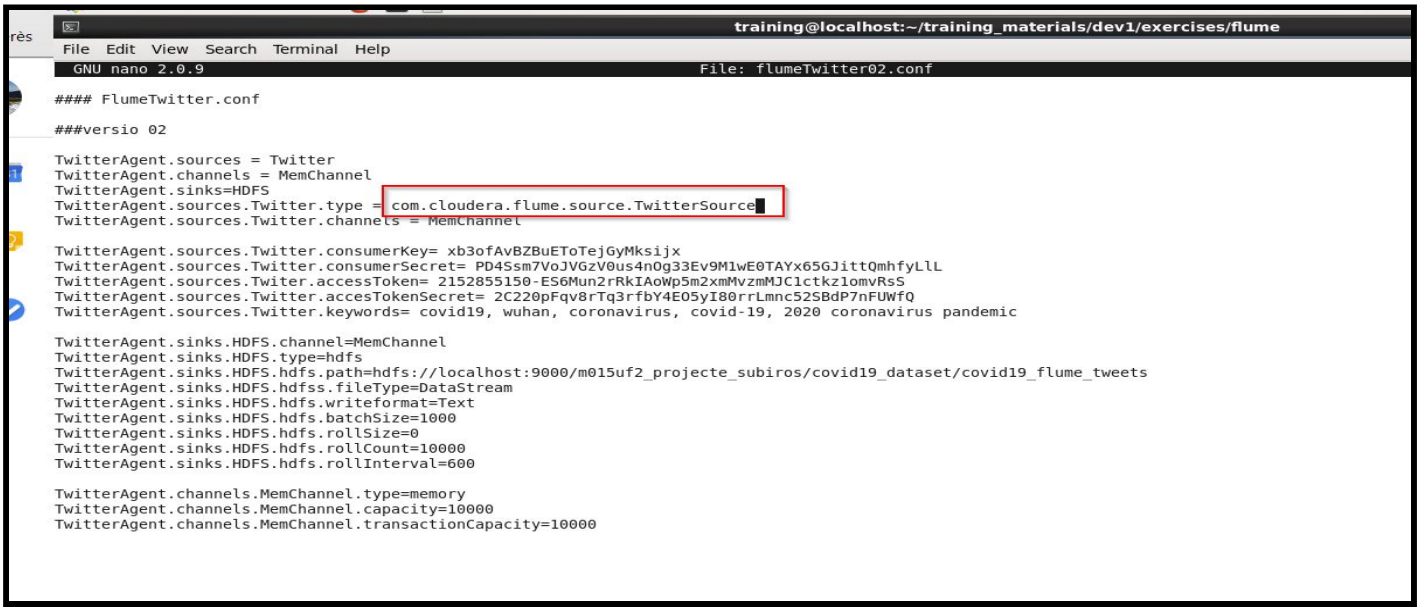
- *Ens dona un error en Access Token i Access Token Secret ja que li consta que és null , tot i s'havia afegit.*



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

• *Flume Sources : Twitter (versio 2)*

arxiu : flumeTwitter02.conf



```
  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= PD4Ssm7VoJVGzV0us4n0g33Ev9M1wE0TAYx65GJittQmhfyLlL
TwitterAgent.sources.Twiter.accessToken= 2152855150-ES6Mun2rRkIAoWp5m2xmMvzmMJC1ctkz1omvRsS
TwitterAgent.sources.Twitter.accesTokenSecret= 2C220pFqv8rTq3rfbY4E05yI80rrLmnc52SBdP7nFUWfQ
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=hdfs://localhost:9000/m015uf2_projecte_subiros/covid19_dataset/covid19_flume_tweets
TwitterAgent.sinks.HDFS.hdfss.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 st
arting
2020-04-10 03:32:42,495 (conf-file-poller-0) [INFO - org.apache.flume.node.PollingPropertiesFileConfigurationProvider$FileWatcherRunnable.run(PollingPropertiesFileConfigurationProvider.java:133)] Reloading c
onfiguration file:./flumeTwitter02.conf
2020-04-10 03:32:42,499 (conf-file-poller-0) [INFO - org.apache.flume.conf.FlumeConfiguration$AgentConfiguration.addProperty(FlumeConfiguration.java:1017)] Processing:HDFS
2020-04-10 03:32:42,500 (conf-file-poller-0) [INFO - org.apache.flume.conf.FlumeConfiguration$AgentConfiguration.addProperty(FlumeConfiguration.java:1017)] Processing:HDFS
2020-04-10 03:32:42,500 (conf-file-poller-0) [INFO - org.apache.flume.conf.FlumeConfiguration$AgentConfiguration.addProperty(FlumeConfiguration.java:1017)] Processing:HDFS
2020-04-10 03:32:42,500 (conf-file-poller-0) [INFO - org.apache.flume.conf.FlumeConfiguration$AgentConfiguration.addProperty(FlumeConfiguration.java:1017)] Processing:HDFS
2020-04-10 03:32:42,500 (conf-file-poller-0) [INFO - org.apache.flume.conf.FlumeConfiguration$AgentConfiguration.addProperty(FlumeConfiguration.java:1017)] Processing:HDFS
2020-04-10 03:32:42,500 (conf-file-poller-0) [INFO - org.apache.flume.conf.FlumeConfiguration$AgentConfiguration.addProperty(FlumeConfiguration.java:931)] Added sinks: HDFS Agent: TwitterAgent
2020-04-10 03:32:42,500 (conf-file-poller-0) [INFO - org.apache.flume.conf.FlumeConfiguration$AgentConfiguration.addProperty(FlumeConfiguration.java:1017)] Processing:HDFS
2020-04-10 03:32:42,500 (conf-file-poller-0) [INFO - org.apache.flume.conf.FlumeConfiguration$AgentConfiguration.addProperty(FlumeConfiguration.java:1017)] Processing:HDFS
2020-04-10 03:32:42,500 (conf-file-poller-0) [INFO - org.apache.flume.conf.FlumeConfiguration$AgentConfiguration.addProperty(FlumeConfiguration.java:1017)] Processing:HDFS
2020-04-10 03:32:42,500 (conf-file-poller-0) [INFO - org.apache.flume.conf.FlumeConfiguration$AgentConfiguration.addProperty(FlumeConfiguration.java:1017)] Processing:HDFS
2020-04-10 03:32:42,634 (conf-file-poller-0) [INFO - org.apache.flume.conf.FlumeConfiguration.validateConfiguration(FlumeConfiguration.java:141)] Post-validation flume configuration contains configuration fo
r agents: [TwitterAgent]
2020-04-10 03:32:42,634 (conf-file-poller-0) [INFO - org.apache.flume.node.AbstractConfigurationProvider.loadChannels(AbstractConfigurationProvider.java:145)] Creating channels
2020-04-10 03:32:42,639 (conf-file-poller-0) [INFO - org.apache.flume.channel.DefaultChannelFactory.create(DefaultChannelFactory.java:42)] Creating instance of channel MemChannel type memory
2020-04-10 03:32:42,642 (conf-file-poller-0) [INFO - org.apache.flume.node.AbstractConfigurationProvider.loadChannels(AbstractConfigurationProvider.java:200)] Created channel MemChannel
2020-04-10 03:32:42,643 (conf-file-poller-0) [INFO - org.apache.flume.source.DefaultSourceFactory.create(DefaultSourceFactory.java:41)] Creating instance of source Twitter, type com.cloudera.flume.source.Twi
tterSource
2020-04-10 03:32:42,644 (conf-file-poller-0) [ERROR - org.apache.flume.node.PollingPropertiesFileConfigurationProvider$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.getClass(DefaultSourceFactory.java:69)
        at org.apache.flume.source.DefaultSourceFactory.create(DefaultSourceFactory.java:42)
        at org.apache.flume.node.AbstractConfigurationProvider.loadSources(AbstractConfigurationProvider.java:322)
        at org.apache.flume.node.AbstractConfigurationProvider.getConfiguration(AbstractConfigurationProvider.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:366)
        at java.net.URLClassLoader$1.run(URLClassLoader.java:355)
        at java.security.AccessController.doPrivileged(Native Method)
        at java.net.URLClassLoader.findClass(URLClassLoader.java:354)
        at java.lang.ClassLoader.loadClass(ClassLoader.java:425)
        at sun.misc.Launcher$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.getClass(DefaultSourceFactory.java:67)
        ... 11 more
```

Nom i Cognoms

Arnau Subirós Puigarnau

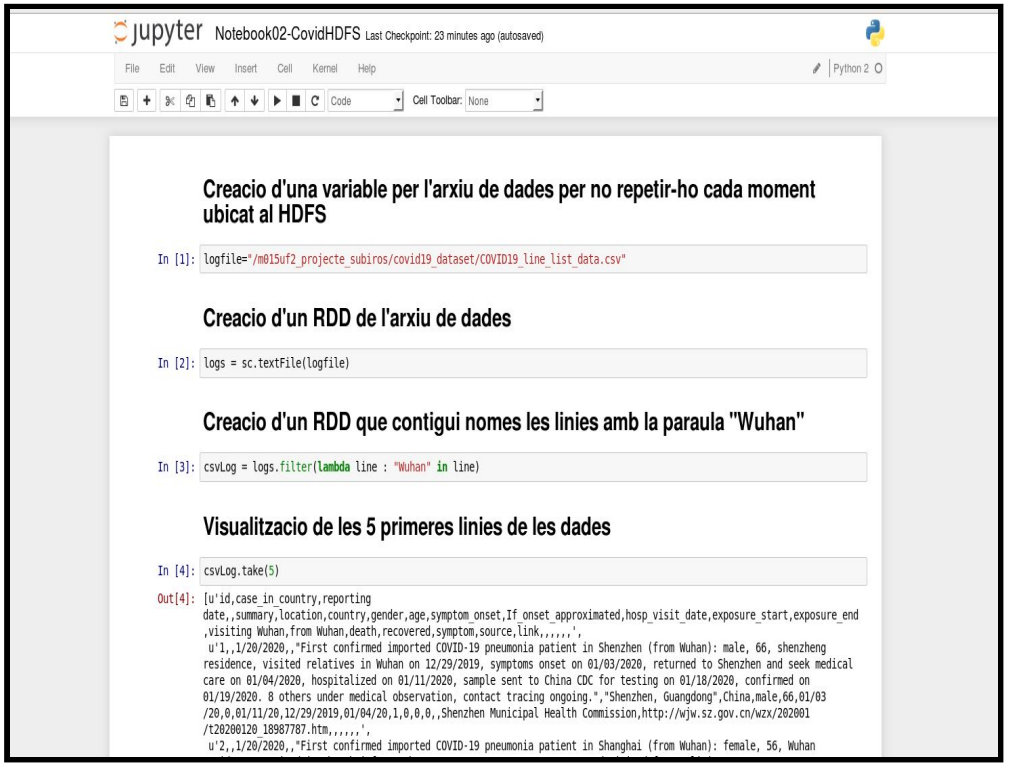
Data

21-04-2020

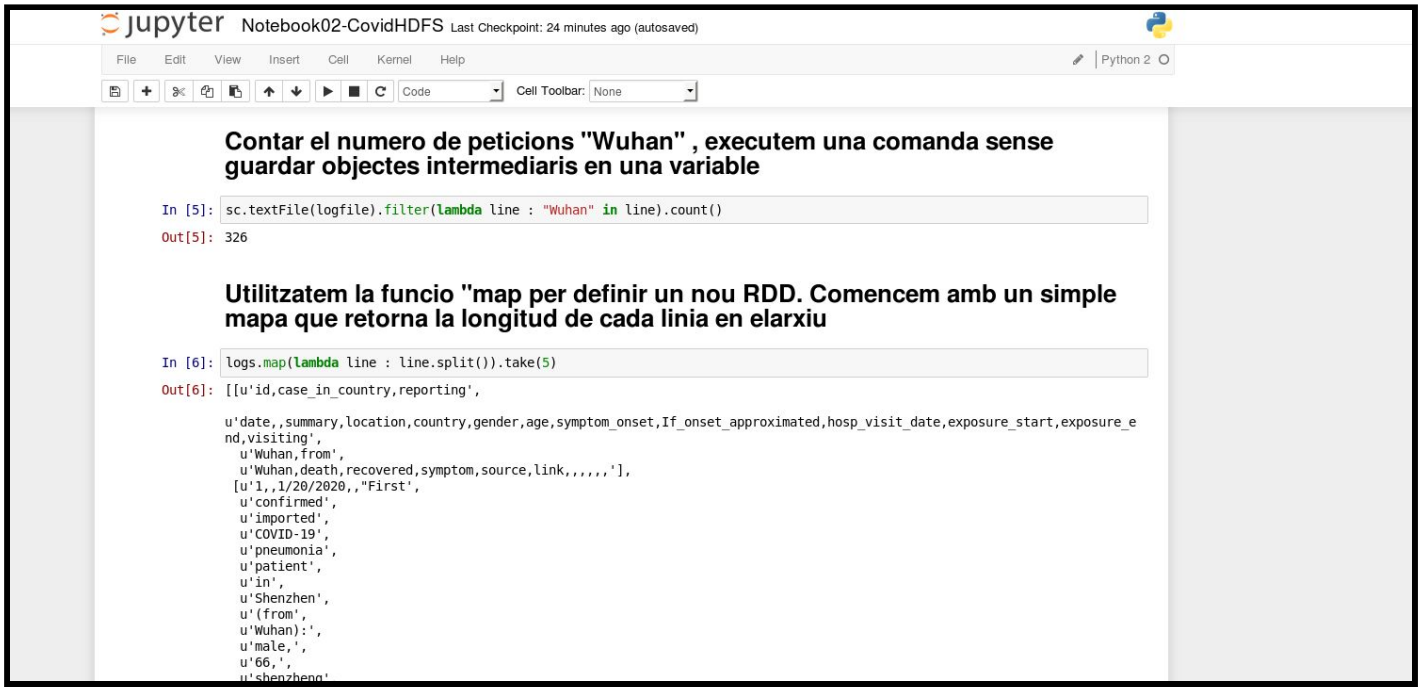
3. Transformació i persistència d'un RDD amb spark shell

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

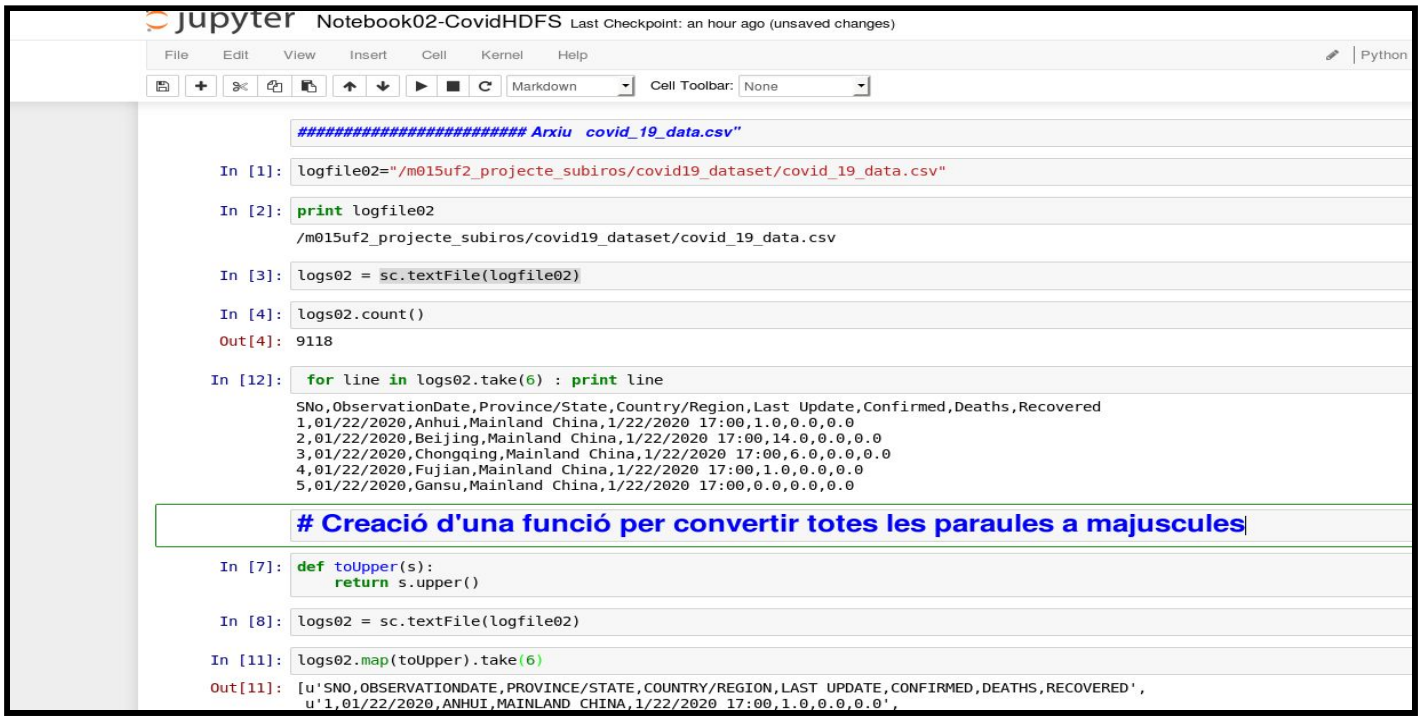
```
[training@localhost data]$ pyspark
[I 02:14:29.189 NotebookApp] Using MathJax:
```



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



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



Nom i Cognoms | **Data**

Arnau Subirós Puigarnau | 21-04-2020

jupyter Notebook02-CovidHDFS Last Checkpoint: Yesterday at 4:12 AM (unsaved changes)

File Edit View Insert Cell Kernel Help | Python 2

Creacio d'una variable per l'arxiu de dades per no repetir-ho cada moment ubicat al HDFS

```
In [20]: logfile03="/m015uf2_projecte_subiros/covid19_dataset/time_series_covid_19_confirmed.csv"
```

Creacio de multiples RDDS

```
In [21]: logfile03_RDD1 = sc.textFile(logfile03)

In [22]: Angola_RDD1 = logfile03_RDD1.filter(lambda line : "Angola" in line)

In [23]: Angola_RDD1.take(3)

Out[23]: [u',Angola,-11.2027,17.8739,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
         ,0,0,0,0,0,0,0,0,0,0,0,1,2,2,3,3,3,4']

In [24]: logfile03_RDD2 =  logfile03_RDD1

In [25]: Algeria_RDD2 = logfile03_RDD2.filter(lambda line : "Algeria" in line)

In [26]: Algeria_RDD2.take(3)

Out[26]: [u',Algeria,28.0339,1.6596,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,1,3,5,12,12,17,1
         7,19,20,20,20,24,26,37,48,54,60,74,87,90,139,201,230,264,302,367']

In [27]: logfile03_RDD3 = logfile03_RDD2

In [28]: Armenia_RDD3 = logfile03_RDD3.filter(lambda line : "Armenia" in line)

In [29]: Armenia_RDD3.take(3)
```

Després de crear diversos RDD els visualitzem unint-los

```
In [35]: Algeria_RDD2.union(Armenia_RDD3).union(Angola_RDD1).take(3)

Out[35]: [u',Algeria,28.0339,1.6596,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,1,3,5,12,12,17,1
         7,19,20,20,20,24,26,37,48,54,60,74,87,90,139,201,230,264,302,367',
          u',Armenia,40.0691,45.0382,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,1
         ,1,1,1,4,8,18,26,52,78,84,115,136,160,194,235,249,265,290',
          u',Angola,-11.2027,17.8739,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
         ,0,0,0,0,0,0,0,0,0,0,0,1,2,2,3,3,3,4']

In [ ]:
```

Nom i Cognoms

Arnau Subirós Puigarnau

Data

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

jupyter Covidversio2 Last Checkpoint: an hour ago (unsaved changes)

File Edit View Insert Cell Kernel Help Python 2

Markdown Cell Toolbar: None

JESUÏTES educació

APACHE HADOOP I APACHE SPARK (i ECOSISTEMA)

M015 uf2 Big Data

Curs: DAM2 2019-20

Alumne: Arnau Subirós

APACHE Spark™

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_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_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)
```

In [5]:
```
data.count()
```

Out[5]:
```
9118
```

In [11]:
```
#filra 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 intermitjes
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 ( número)
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']]
```

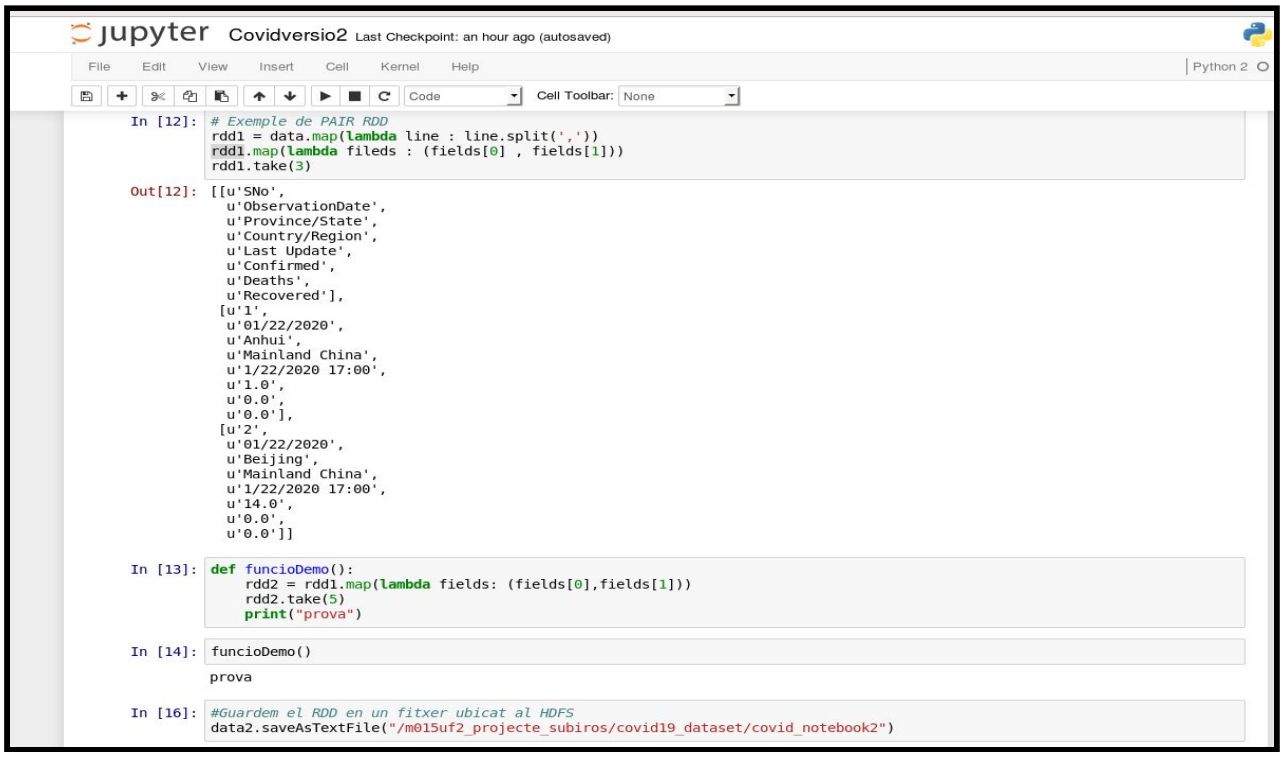
In [10]:
```
#Aga nomes data 2 ,ls qual son  la primera paraula e cada entrada. Es crea un RDD i nomes es mostres els 5 primers .
data2 = data.map(lambda line: line.split()[0])
data2.take(5)
```

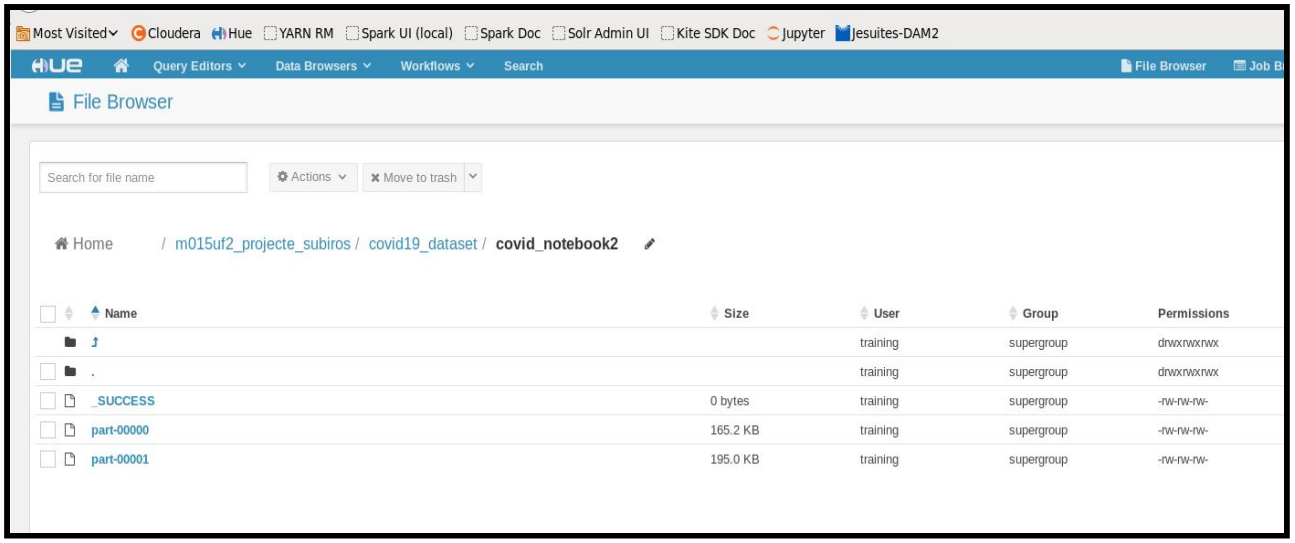
Out[10]:
```
[u'SNo,ObservationDate,Province/State,Country/Region,Last',
 u'1,01/22/2020,Anhui,Mainland',
 u'2,01/22/2020,Beijing,Mainland',
 u'3,01/22/2020,Chongqing,Mainland',
 u'4,01/22/2020,Fujian,Mainland']
```

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
```

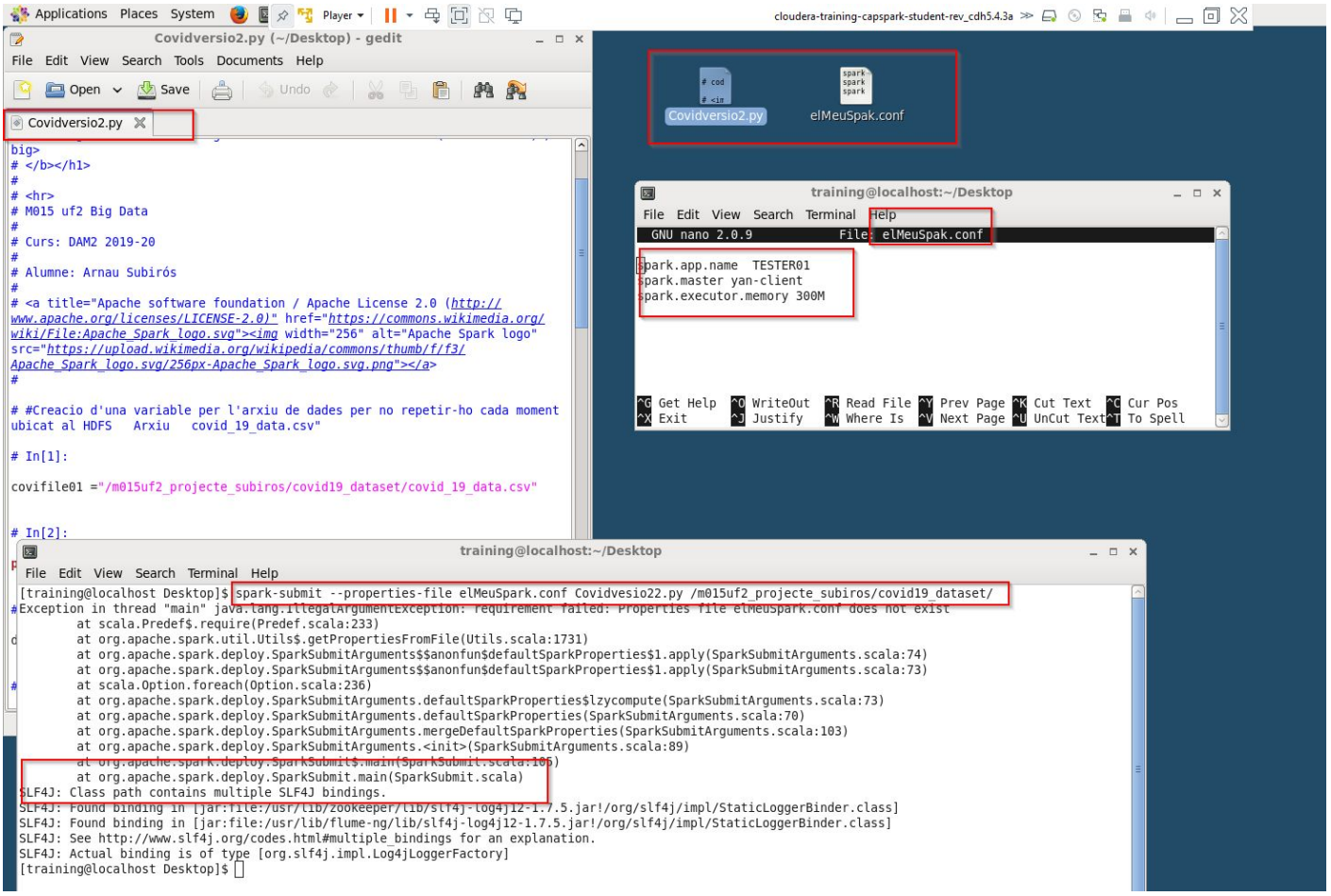
Nom i Cognoms	Data
Arnau Subirós Puigarnau	21-04-2020





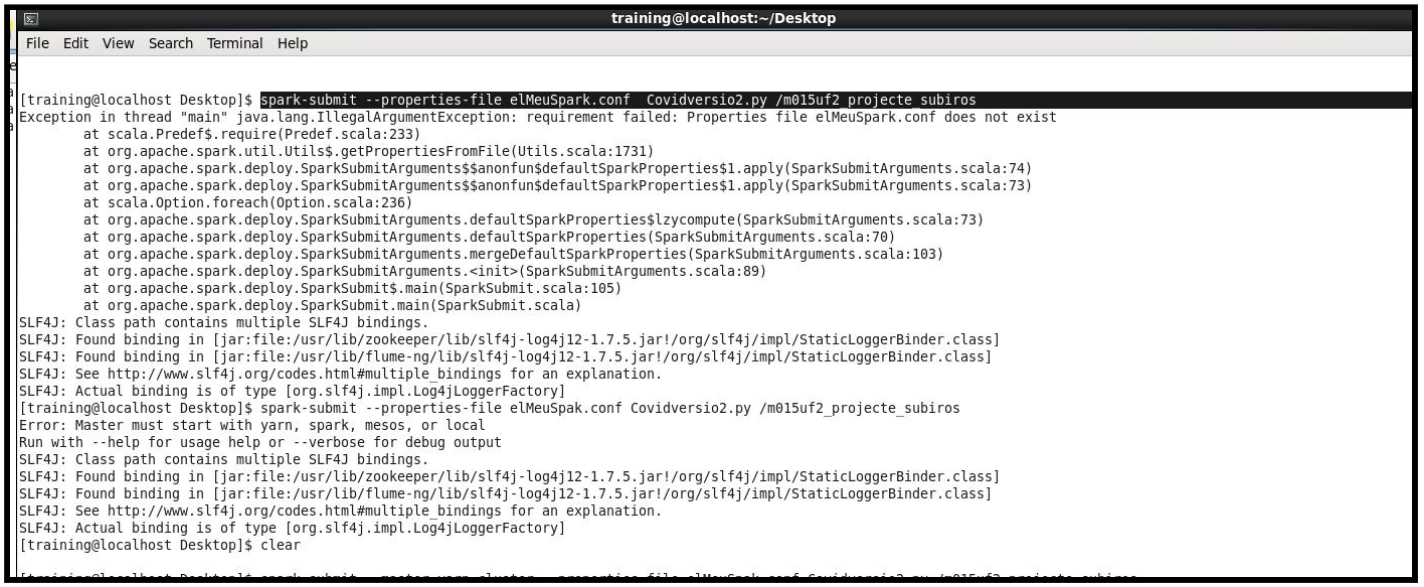
Nom i Cognoms	Data
Arnau Subirós Puigarnau	21-04-2020

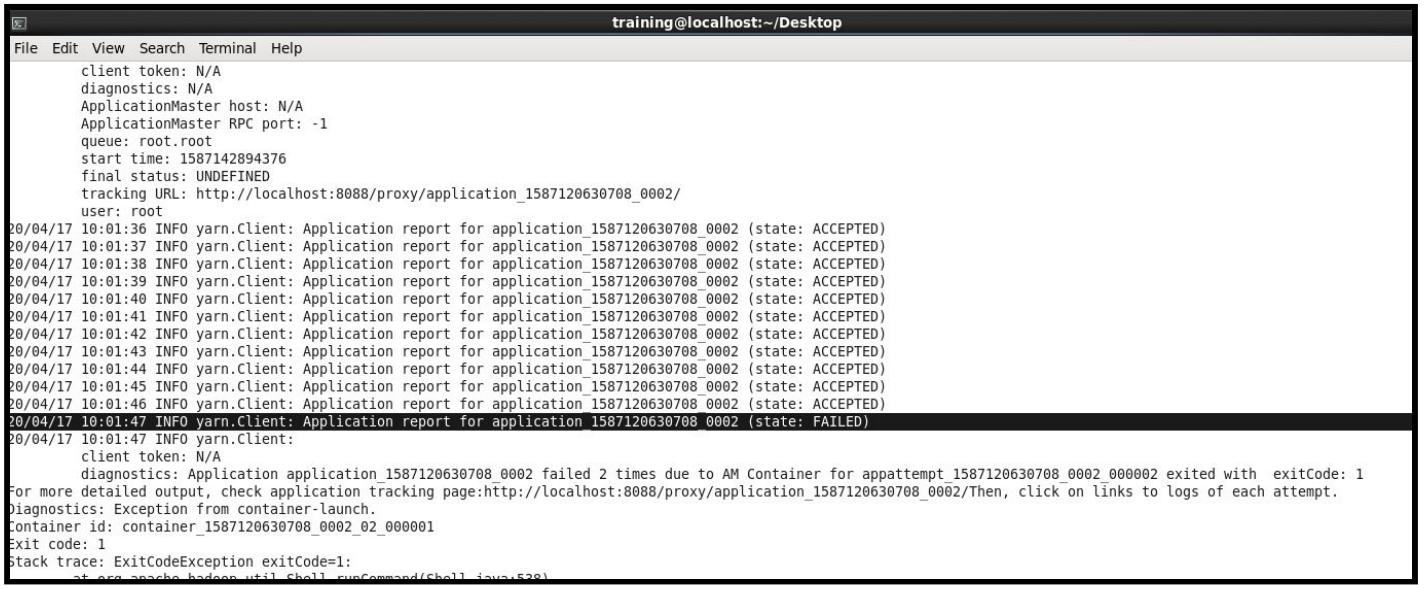
- ***Intentem crear una aplicacio amb SPARK***



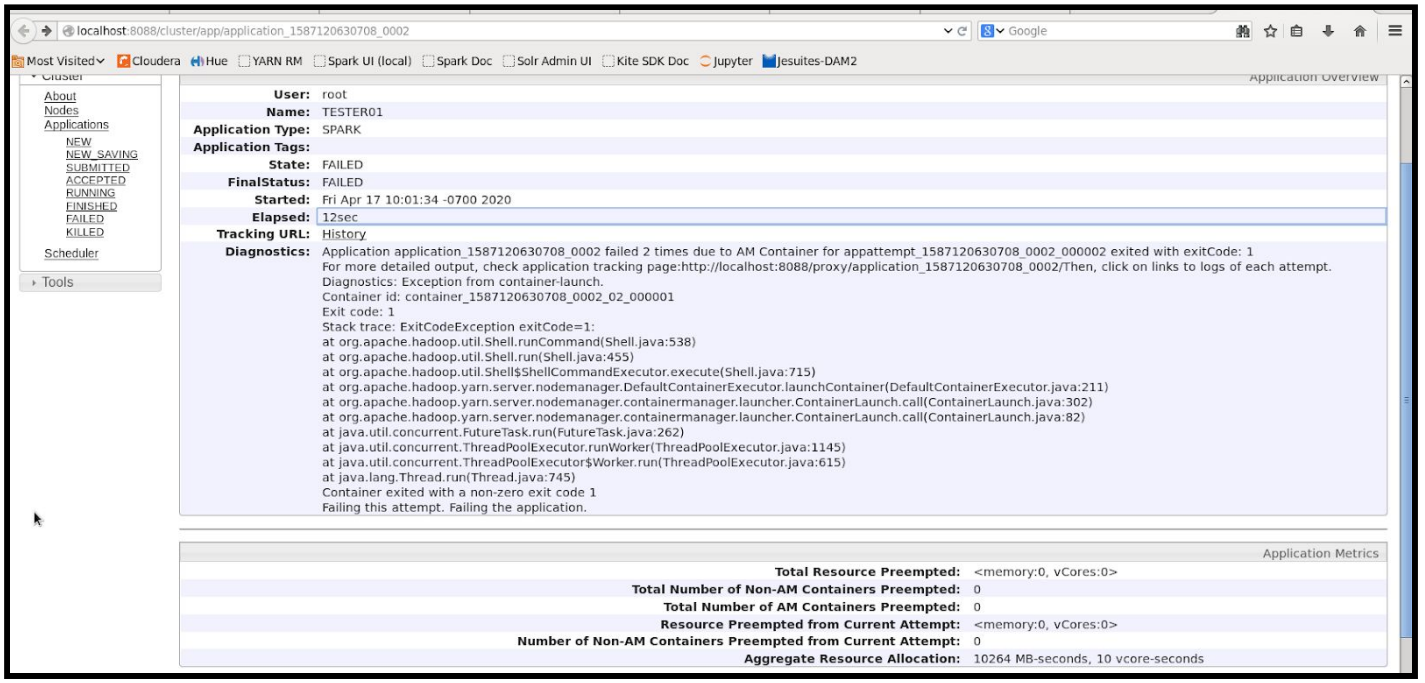
Nom i Cognoms **Data**

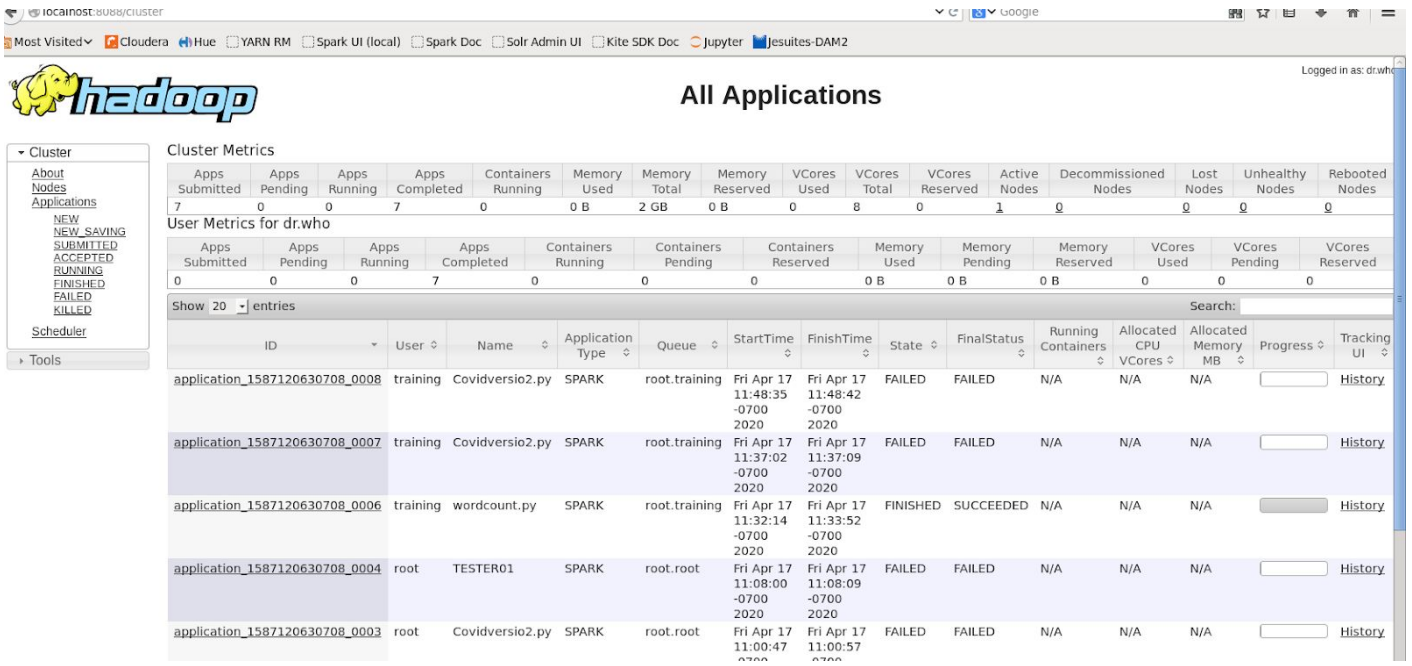
Arnau Subirós Puigarnau 21-04-2020



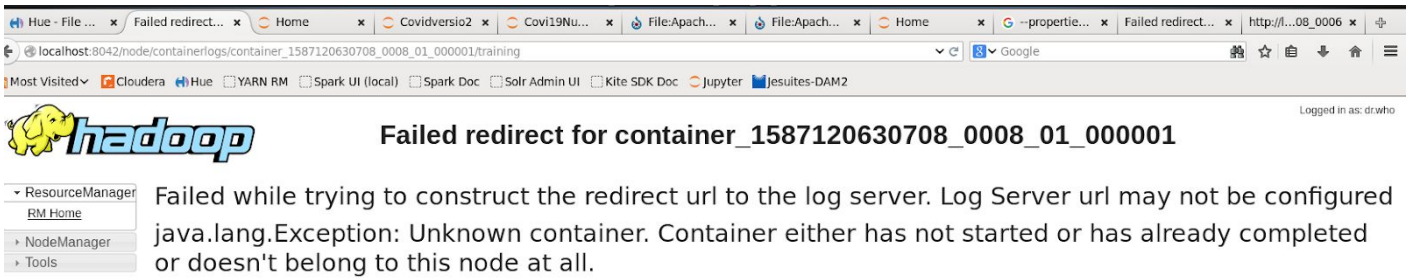


Nom i Cognoms	Data
Arnau Subirós Puigarnau	21-04-2020

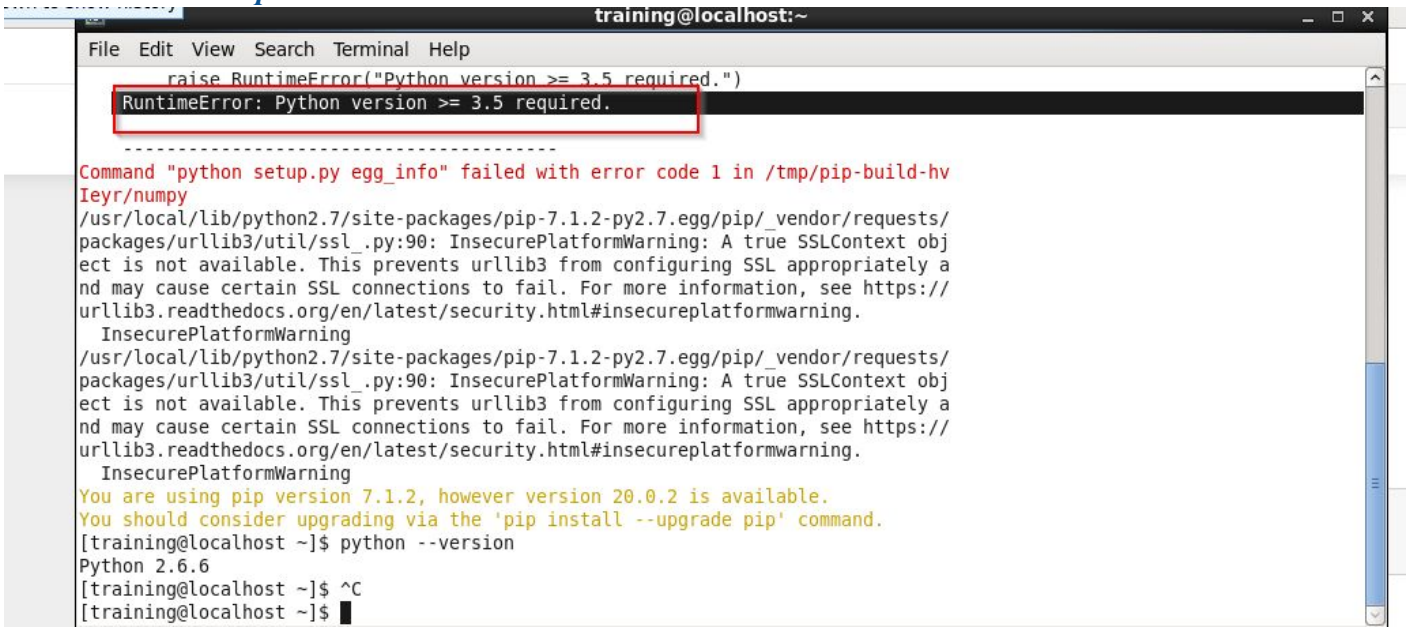




Nom i Cognoms	Data
Arnau Subirós Puigarnau	21-04-2020

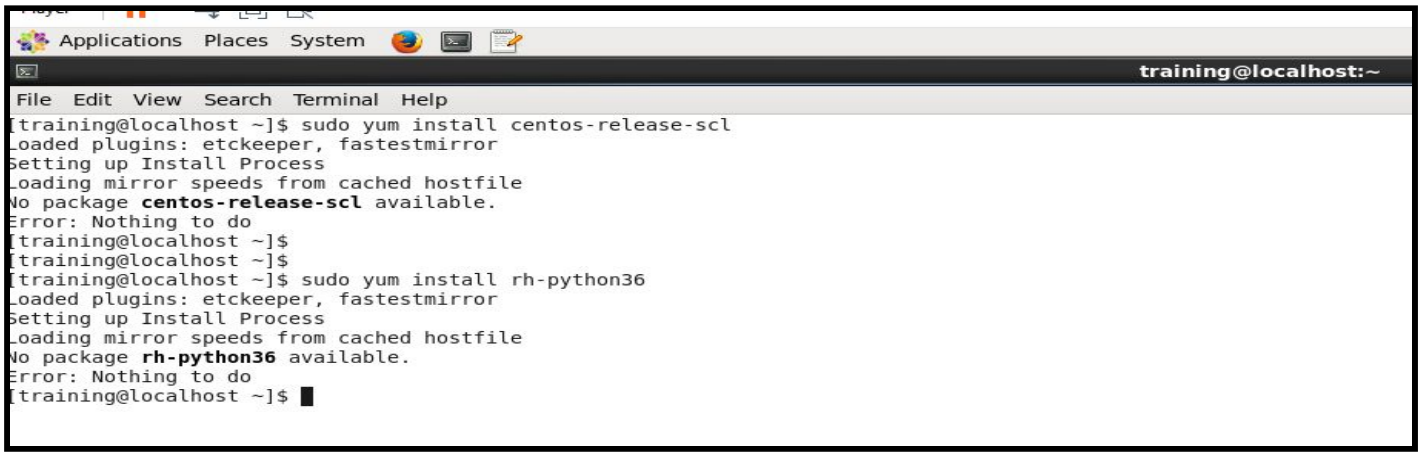


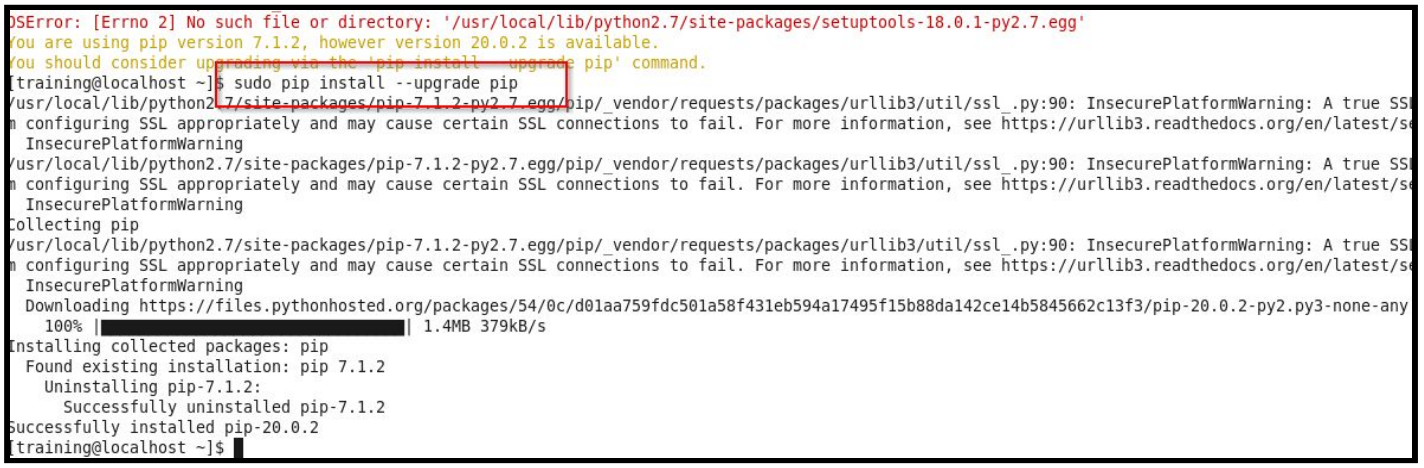
intentem instal.lar pandas



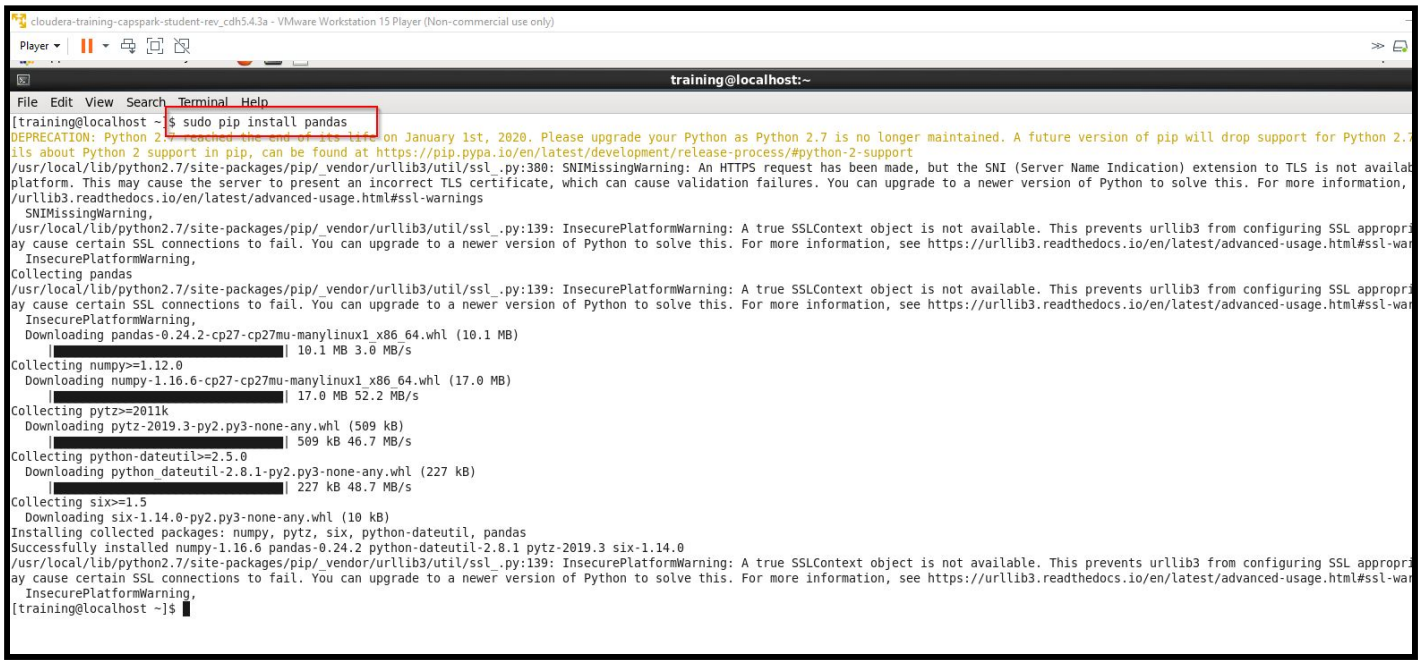
- ***intentem instal.lar sense èxit Python 3***

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





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



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

Bibiografia :

- *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/*