EAFIT

Special topics in telematics

Laboratory 3

Luisa María Álvarez García – Computer Science

(lmalvarez8@eafit.edu.co)

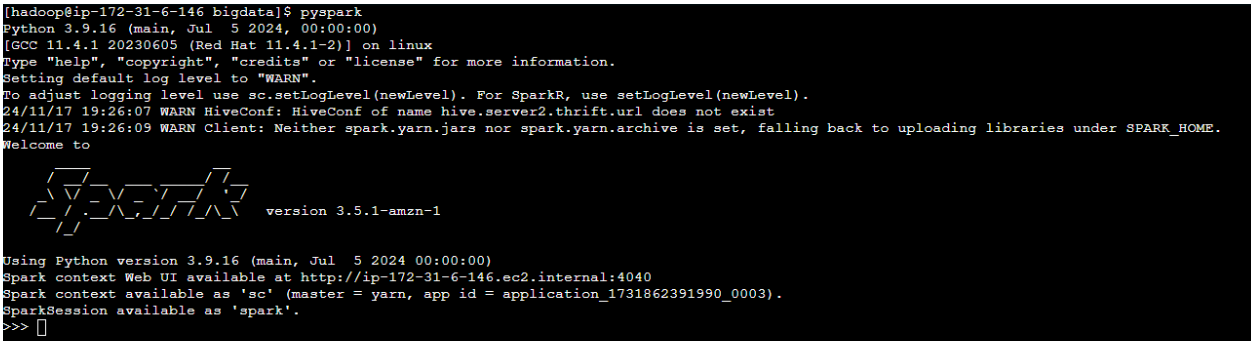
Teacher: Edwin Nelson Montoya

Medellin, November 11, 2024

1. To connect into the spark console, first into, into the main node console in ec2 and inside the console write the command:

```
pyspark
```

This Will open the spark consoles, and it should look like:



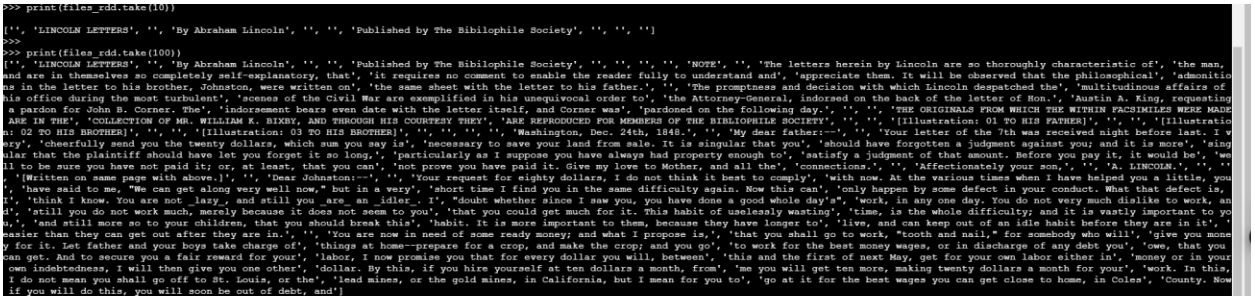
```
[hadoop@ip-172-31-6-146 bigdata]$ pyspark
Python 3.9.16 (main, Jul  5 2024, 00:00:00)
[GCC 11.4.1 20230605 (Red Hat 11.4.1-2)] on linux
Type "help", "copyright", "credits" or "license" for more information.
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/11/17 19:26:07 WARN HiveConf: HiveConf of name hive.server2.thrift.url does not exist
24/11/17 19:26:09 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries under SPARK_HOME.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 3.5.1-amzn-1
      /_/

Using Python version 3.9.16 (main, Jul  5 2024 00:00:00)
Spark context Web UI available at http://ip-172-31-6-146.ec2.internal:4040
Spark context available as 'sc' (master = yarn, app id = application_1731862391990_0003).
SparkSession available as 'spark'.
>>> 
```

2. Inside the console run:

```
files_rdd = sc.textFile("hdfs:///datasets/gutenberg-small/*.txt")
>>> files_rdd = sc.textFile("s3://emontoyadatasets/gutenberg-small/*.txt")
```

And it should look like:



```
>>> print(files_rdd.take(10))

['', 'LINCOLN LETTERS', '', 'By Abraham Lincoln', '', '', 'Published by The Bibilophile Society', '', '', '']
>>>
>>> print(files_rdd.take(100))
['', 'LINCOLN LETTERS', '', 'By Abraham Lincoln', '', '', 'Published by The Bibilophile Society', '', '', '', '', 'NOTE', '', 'The letters herein by Lincoln are so thoroughly characteristic of', 'the man,
and are in themselves so completely self-explanatory, that', 'it requires no comment to enable the reader fully to understand and', 'appreciate them. It will be observed that the philosophical', 'admonitio
ns in the letter to his brother, Johnston, were written on', 'the same sheet with the letter to his father.', '', 'The promptness and decision with which Lincoln despatched the', 'multitudinous affairs of
his office during the most turbulent', 'scenes of the Civil War are exemplified in his unequivocal order to', 'the Attorney-General, indorsed on the back of the letter of Hon.', 'Austin A. King, requesting
 a pardon for John B. Corner. The', 'indorsement bears even date with the letter itself, and Corner was', 'pardoned on the following day.', '', '', 'THE ORIGINALS FROM WHICH THE WITHIN FACSIMILES WERE MADE
 ARE IN THE', 'COLLECTION OF MR. WILLIAM K. BIXBY, AND THROUGH HIS COURTESY THEY', 'ARE REPRODUCED FOR MEMBERS OF THE BIBLIOPHILE SOCIETY', '', '', '[Illustration: 01 TO HIS FATHER]', '', '', '[Illustratio
n: 02 TO HIS BROTHER]', '', '', '[Illustration: 03 TO HIS BROTHER]', '', '', '', '', 'Washington, Dec. 24th, 1848.', '', 'My dear father:--', '', 'Your letter of the 7th was received night before last. I v
ery', 'cheerfully send you the twenty dollars, which sum you say is', 'necessary to save your land from sale. It is singular that you', 'should have forgotten a judgment against you; and it is more', 'sing
ular that the plaintiff should have let you forget it so long,', 'particularly as I suppose you have always had property enough to', 'satisfy a judgment of that amount. Before you pay it, it would be', 'we
ll to be sure you have not paid it; or, at least, that you can', 'not prove you have paid it. Give my love to Mother, and all the', 'connections.', '', 'Affectionately your son,', '', 'A. LINCOLN.', '', ''
, '[Written on same page with above.]', '', 'Dear Johnston:--', '', 'Your request for eighty dollars, I do not think it best to comply', 'with now. At the various times when I have helped you a little, you
', 'have said to me, "We can get along very well now," but in a very', 'short time I find you in the same difficulty again. Now this can', 'only happen by some defect in your conduct. What that defect is,
I', 'think I know. You are not _lazy_, and still you _are_ an _idler_. I', "doubt whether since I saw you, you have done a good whole day's", 'work, in any one day. You do not very much dislike to work, an
d', 'still you do not work much, merely because it does not seem to you', 'that you could get much for it. This habit of uselessly wasting', 'time, is the whole difficulty; and it is vastly important to yo
u,', 'and still more so to your children, that you should break this', 'habit. It is more important to them, because they have longer to', 'live, and can keep out of an idle habit before they are in it', '
easier than they can get out after they are in.', '', 'You are now in need of some ready money; and what I propose is,', 'that you shall go to work, "tooth and nail," for somebody who will', 'give you mone
y for it. Let father and your boys take charge of', 'things at home--prepare for a crop, and make the crop; and you go', 'to work for the best money wages, or in discharge of any debt you', 'owe, that you
can get. And to secure you a fair reward for your', 'labor, I now promise you that for every dollar you will, between', 'this and the first of next May, get for your own labor either in', 'money or in your
 own indebtedness, I will then give you one other', 'dollar. By this, if you hire yourself at ten dollars a month, from', 'me you will get ten more, making twenty dollars a month for your', 'work. In this,
 I do not mean you shall go off to St. Louis, or the', 'lead mines, or the gold mines, in California, but I mean for you to', 'go at it for the best wages you can get close to home, in Coles', 'County. Now
 if you will do this, you will soon be out of debt, and']
```

3. After that, run in the console:

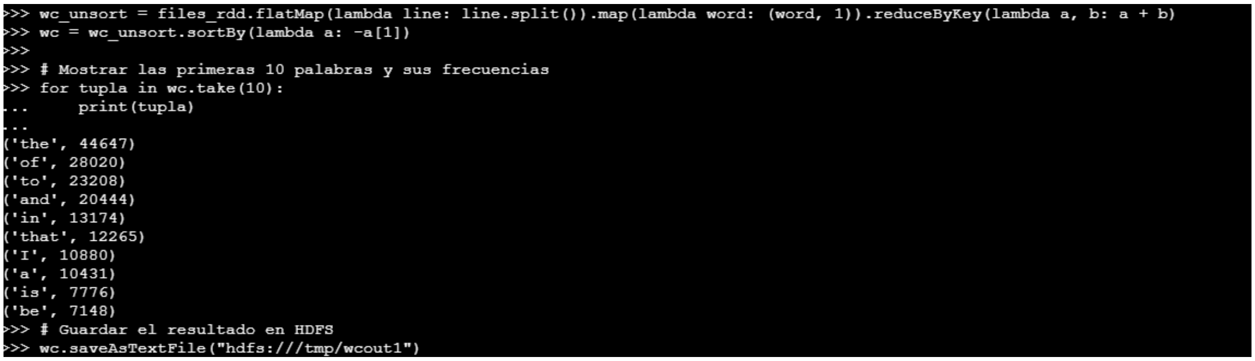
```
wc_unsort = files_rdd.flatMap(lambda line: line.split()).map(lambda word: (word, 1)).reduceByKey(lambda a, b: a + b)
wc = wc_unsort.sortBy(lambda a: -a[1])

# Mostrar las primeras 10 palabras y sus frecuencias

for tupla in wc.take(10):
  print(tupla)
```

```
# Guardar el resultado en HDFS
wc.saveAsTextFile("hdfs:///tmp/wcout1")
```

And show:



```
>>> wc_unsort = files_rdd.flatMap(lambda line: line.split()).map(lambda word: (word, 1)).reduceByKey(lambda a, b: a + b)
>>> wc = wc_unsort.sortBy(lambda a: -a[1])
>>>
>>> # Mostrar las primeras 10 palabras y sus frecuencias
>>> for tupla in wc.take(10):
...     print(tupla)
...
('the', 44647)
('of', 28020)
('to', 23208)
('and', 20444)
('in', 13174)
('that', 12265)
('I', 10880)
('a', 10431)
('is', 7776)
('be', 7148)
>>> # Guardar el resultado en HDFS
>>> wc.saveAsTextFile("hdfs:///tmp/wcout1")
```

4. To consolidate only one output file, run:

```
wc.coalesce(1).saveAsTextFile("hdfs:///tmp/wcout2")
```

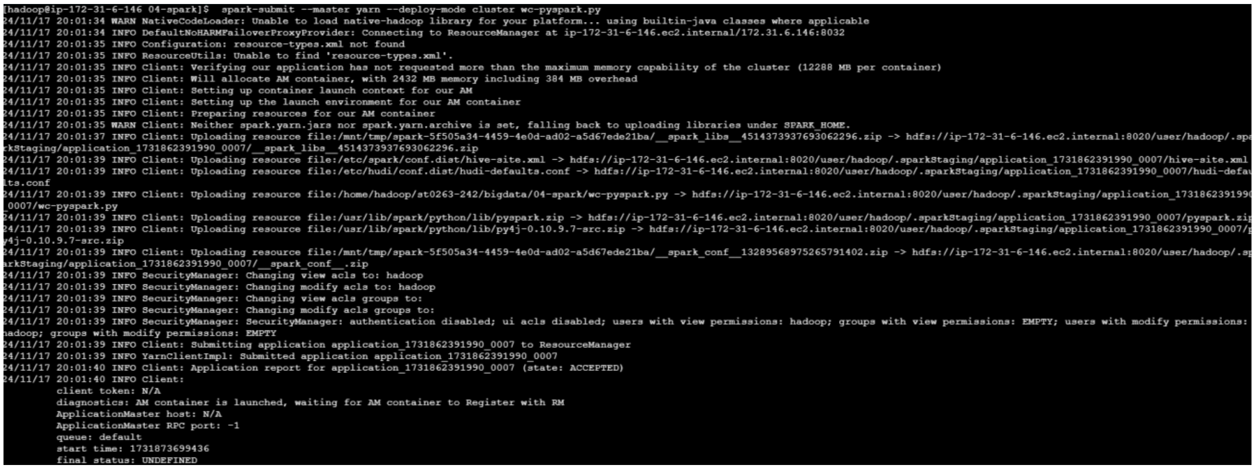
5. Now to do it with a python file, in this case called *wc-pyspark.py* located into the 04-spark:

```
cd 04-spark
```

6. Run the command:

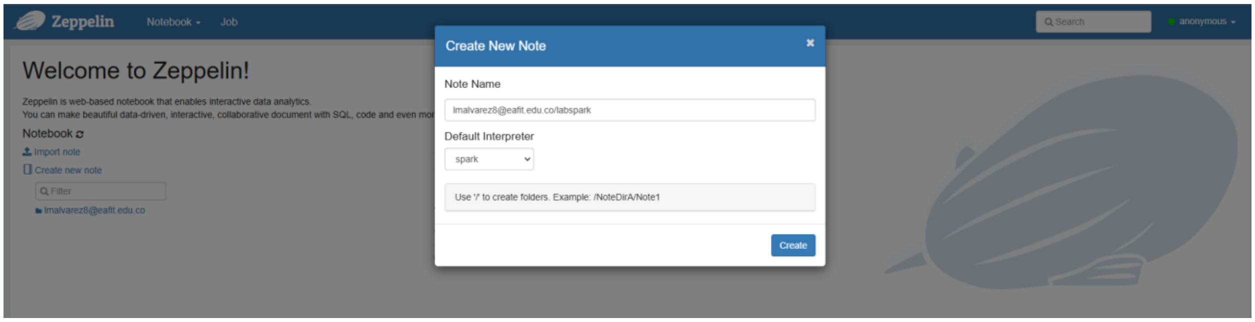
```
spark-submit --master yarn --deploy-mode cluster wc-pyspark.py
```

And it should show:



```
[hadoop@ip-172-31-6-146 04-spark]$  spark-submit --master yarn --deploy-mode cluster wc-pyspark.py
24/11/17 20:01:34 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
24/11/17 20:01:34 INFO DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at ip-172-31-6-146.ec2.internal/172.31.6.146:8032
24/11/17 20:01:35 INFO Configuration: resource-types.xml not found
24/11/17 20:01:35 INFO ResourceUtils: Unable to find 'resource-types.xml'.
24/11/17 20:01:35 INFO Client: Verifying our application has not requested more than the maximum memory capability of the cluster (12288 MB per container)
24/11/17 20:01:35 INFO Client: Will allocate AM container, with 2432 MB memory including 384 MB overhead
24/11/17 20:01:35 INFO Client: Setting up container launch context for our AM
24/11/17 20:01:35 INFO Client: Setting up the launch environment for our AM container
24/11/17 20:01:35 INFO Client: Preparing resources for our AM container
24/11/17 20:01:35 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries under SPARK_HOME.
24/11/17 20:01:37 INFO Client: Uploading resource file:/mnt/tmp/spark-5f505a34-4459-4e0d-ad02-a5d67ede21ba/__spark_libs__4514373937693062296.zip -> hdfs://ip-172-31-6-146.ec2.internal:8020/user/hadoop/.spa
rkStaging/application_1731862391990_0007/__spark_libs__4514373937693062296.zip
24/11/17 20:01:39 INFO Client: Uploading resource file:/etc/spark/conf.dist/hive-site.xml -> hdfs://ip-172-31-6-146.ec2.internal:8020/user/hadoop/.sparkStaging/application_1731862391990_0007/hive-site.xml
24/11/17 20:01:39 INFO Client: Uploading resource file:/etc/hudi/conf.dist/hudi-defaults.conf -> hdfs://ip-172-31-6-146.ec2.internal:8020/user/hadoop/.sparkStaging/application_1731862391990_0007/hudi-defau
lts.conf
24/11/17 20:01:39 INFO Client: Uploading resource file:/home/hadoop/st0263-242/bigdata/04-spark/wc-pyspark.py -> hdfs://ip-172-31-6-146.ec2.internal:8020/user/hadoop/.sparkStaging/application_1731862391990
_0007/wc-pyspark.py
24/11/17 20:01:39 INFO Client: Uploading resource file:/usr/lib/spark/python/lib/pyspark.zip -> hdfs://ip-172-31-6-146.ec2.internal:8020/user/hadoop/.sparkStaging/application_1731862391990_0007/pyspark.zip
24/11/17 20:01:39 INFO Client: Uploading resource file:/usr/lib/spark/python/lib/py4j-0.10.9.7-src.zip -> hdfs://ip-172-31-6-146.ec2.internal:8020/user/hadoop/.sparkStaging/application_1731862391990_0007/p
y4j-0.10.9.7-src.zip
24/11/17 20:01:39 INFO Client: Uploading resource file:/mnt/tmp/spark-5f505a34-4459-4e0d-ad02-a5d67ede21ba/__spark_conf__1328956897526579l402.zip -> hdfs://ip-172-31-6-146.ec2.internal:8020/user/hadoop/.sp
arkStaging/application_1731862391990_0007/__spark_conf__.zip
24/11/17 20:01:39 INFO SecurityManager: Changing view acls to: hadoop
24/11/17 20:01:39 INFO SecurityManager: Changing modify acls to: hadoop
24/11/17 20:01:39 INFO SecurityManager: Changing view acls groups to:
24/11/17 20:01:39 INFO SecurityManager: Changing modify acls groups to:
24/11/17 20:01:39 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: hadoop; groups with view permissions: EMPTY; users with modify permissions:
hadoop; groups with modify permissions: EMPTY
24/11/17 20:01:39 INFO Client: Submitting application application_1731862391990_0007 to ResourceManager
24/11/17 20:01:39 INFO YarnClientImpl: Submitted application application_1731862391990_0007
24/11/17 20:01:40 INFO Client: Application report for application_1731862391990_0007 (state: ACCEPTED)
24/11/17 20:01:40 INFO Client:
         client token: N/A
         diagnostics: AM container is launched, waiting for AM container to Register with RM
         ApplicationMaster host: N/A
         ApplicationMaster RPC port: -1
         queue: default
         start time: 1731873699436
         final status: UNDEFINED
```

7. Now into zeppelin into the EMR cluster, enter and open the custom TCP port 8890 for 0.0.0.0/0 and créate a note like this:



Welcome to Zeppelin!

Zeppelin is web-based notebook that enables interactive data analytics.
You can make beautiful data-driven, interactive, collaborative document with SQL, code and even mor

Notebook

lmalvarez8@eafit.edu.co

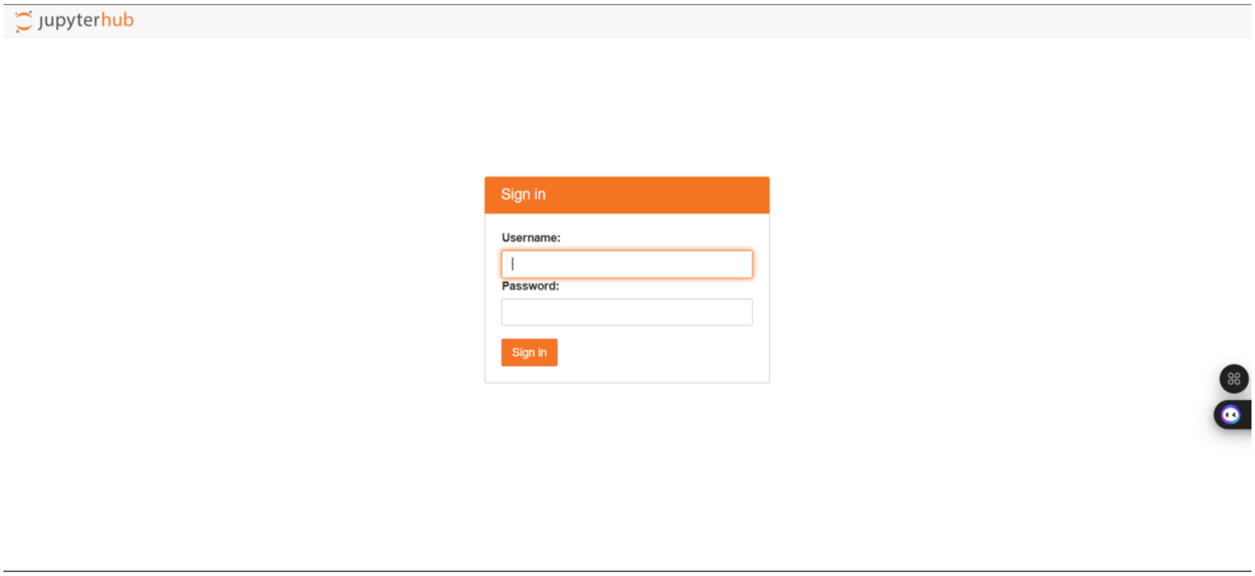
Create New Note

Note Name: lmalvarez8@eafit.edu.co/labspark

Default Interpreter: spark

Use '/' to create folders. Example: /NoteDirA/Note1

8. Now for make the covid analysis, first enter into the jupyterhub application located into the EMR listed applications, and open the port; after that login using the user: *jovyan* and the password: *jupyter*.



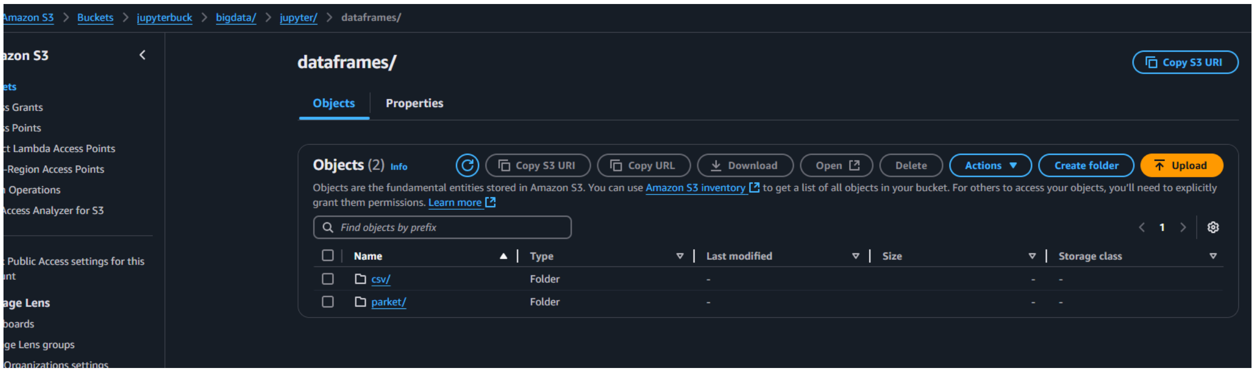
Sign in

Username:

Password:

Now inside s3 create a bucket called bigdata/datasets and upload the covid.csv file, you can have it in this url https://www.datos.gov.co/api/views/gt2j-8ykr/rows.csv?accessType=DOWNLOAD
And inside of it upload the pyspark_save and the sparksql_covid and select for the pyspark file the pyspark kernel and the spark kernel for the other file, and after that run all cells.

Now, after run all cells, s3 will look like:

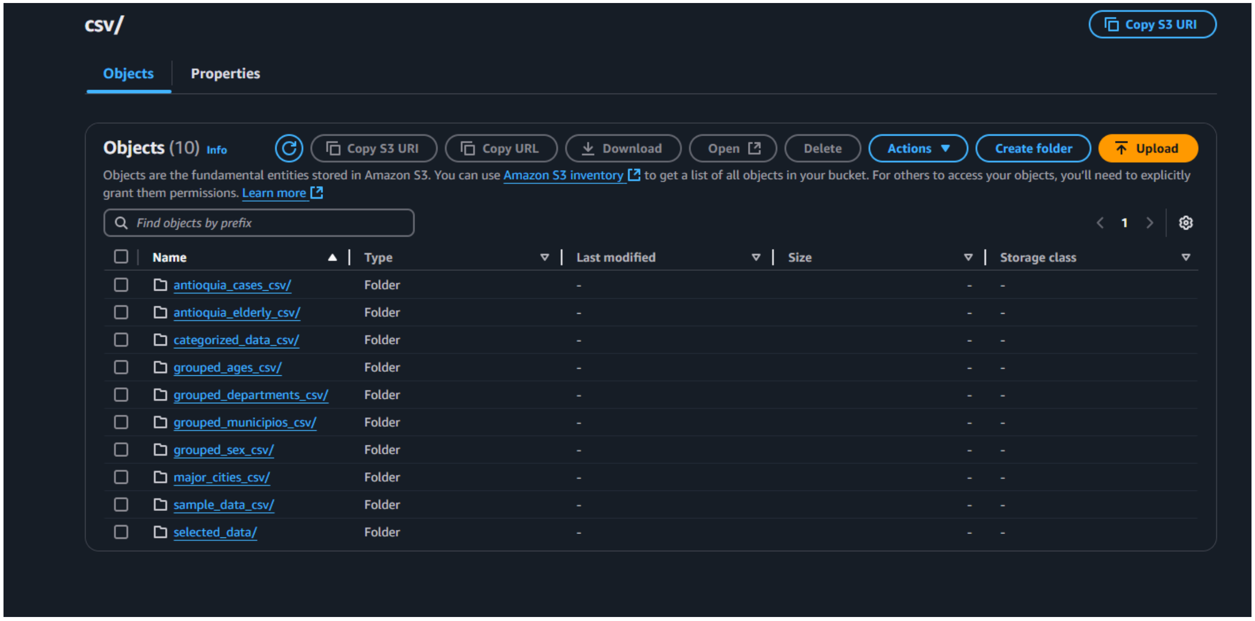


dataframes/

Objects (2)

Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

	Name	Type	Last modified	Size	Storage class
□	csv/	Folder	-	-	-
□	parket/	Folder	-	-	-

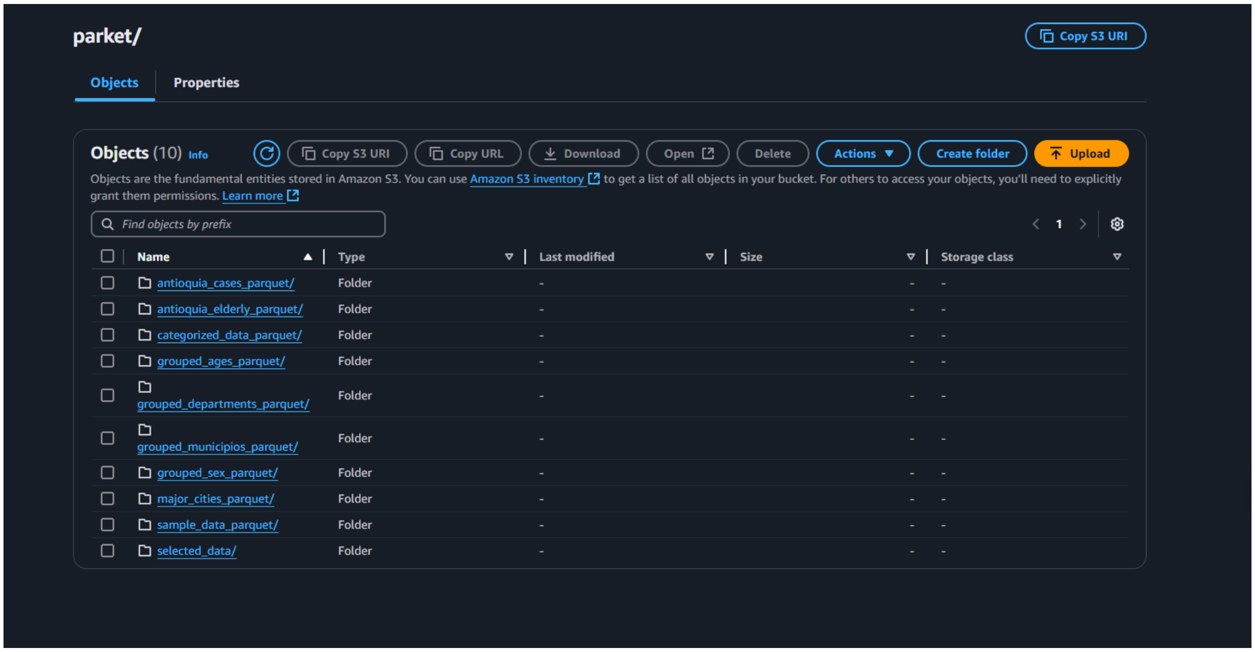


csv/

Objects (10)

Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

	Name	Type	Last modified	Size	Storage class
□	antioquia_cases_csv/	Folder	-	-	-
□	antioquia_elderly_csv/	Folder	-	-	-
□	categorized_data_csv/	Folder	-	-	-
□	grouped_ages_csv/	Folder	-	-	-
□	grouped_departments_csv/	Folder	-	-	-
□	grouped_municipios_csv/	Folder	-	-	-
□	grouped_sex_csv/	Folder	-	-	-
□	major_cities_csv/	Folder	-	-	-
□	sample_data_csv/	Folder	-	-	-
□	selected_data/	Folder	-	-	-



parket/

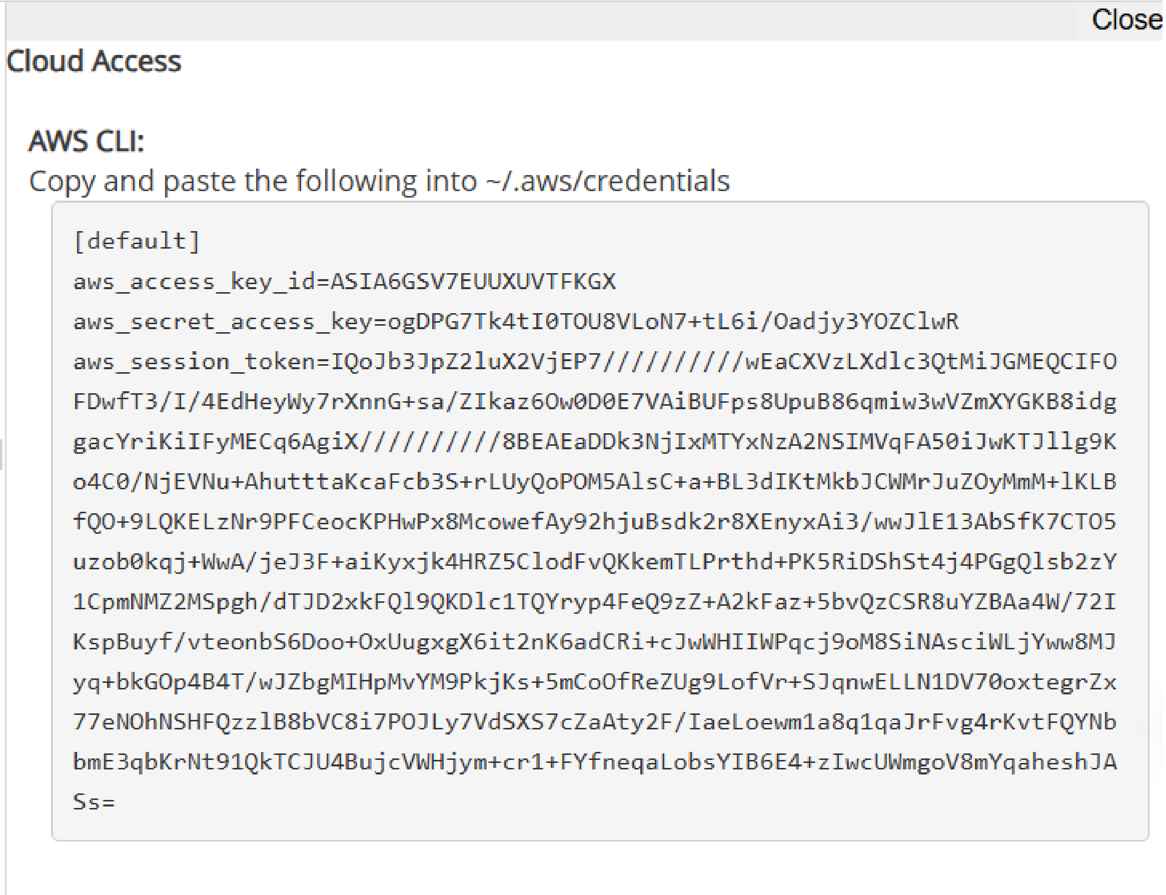
Objects (10)

Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

□	Name ▲	Type	Last modified	Size	Storage class
□	antioquia_cases_parquet/	Folder	-	-	-
□	antioquia_elderly_parquet/	Folder	-	-	-
□	categorized_data_parquet/	Folder	-	-	-
□	grouped_ages_parquet/	Folder	-	-	-
□	grouped_departments_parquet/	Folder	-	-	-
□	grouped_municipios_parquet/	Folder	-	-	-
□	grouped_sex_parquet/	Folder	-	-	-
□	major_cities_parquet/	Folder	-	-	-
□	sample_data_parquet/	Folder	-	-	-
□	selected_data/	Folder	-	-	-

S3 uri: s3://jupyterbuck/bigdata/jupyter/

9. In other case, if you are using colab, you just need to insert the notebook on colab, change the secret, access key you can find into the aws readme.



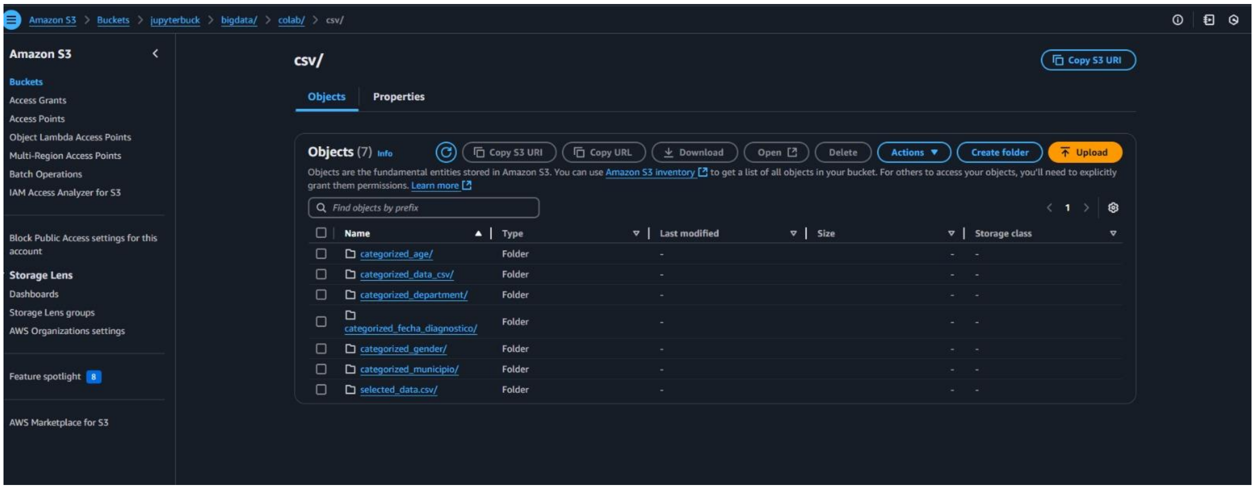
Cloud Access

AWS CLI:

Copy and paste the following into ~/.aws/credentials

```
[default]
aws_access_key_id=ASIA6GSV7EUUXUVTFKGX
aws_secret_access_key=ogDPG7Tk4tI0TOU8VLoN7+tL6i/Oadjy3YOZClwR
aws_session_token=IQoJb3JpZ2luX2VjEP7//////////wEaCXVzLXdlc3QtMiJGMEQCIFO
FDwfT3/I/4EdHeyWy7rXnnG+sa/ZIkaz6Ow0D0E7VAiBUFps8UpuB86qmiw3wVZmXYGKB8idg
gacYriKiIFyMECq6AgiX//////////8BEAEaDDk3NjIxMTYxNzA2NSIMVqFA50iJwKTJllg9K
o4C0/NjEVNu+AhutttaKcaFcb3S+rLUyQoPOM5AlsC+a+BL3dIKtMkbJCWMrJuZOyMmM+lKLB
fQO+9LQKELzNr9PFCeocKPHwPx8McowefAy92hjuBsdk2r8XEnyxAi3/wwJlE13AbSfK7CTO5
uzob0kqj+WwA/jeJ3F+aiKyxjk4HRZ5ClodFvQKkemTLPrthd+PK5RiDShSt4j4PGgQlsb2zY
1CpmNMZ2MSpgh/dTJD2xkFQl9QKDlc1TQYryp4FeQ9zZ+A2kFaz+5bvQzCSR8uYZBAa4W/72I
KspBuyf/vteonbS6Doo+OxUugxgX6it2nK6adCRi+cJwWHIIWPqcj9oM8SiNAsciWLjYww8MJ
yq+bkGOp4B4T/wJZbgMIHpMvYM9PkjKs+5mCoOfReZUg9LofVr+SJqnwELLN1DV70oxtegrZx
77eNOhNSHFQzzlB8bVC8i7POJLy7VdSXS7cZaAty2F/IaeLoewm1a8q1qaJrFvg4rKvtFQYNb
bmE3qbKrNt91QkTCJU4BujcVWHjym+cr1+FYfneqaLobsYIB6E4+zIwcUWmgoV8mYqaheshJA
Ss=
```

10. Select the access key, the secret key and the session, after that, just run the colab notebook and note that now your s3 contains:

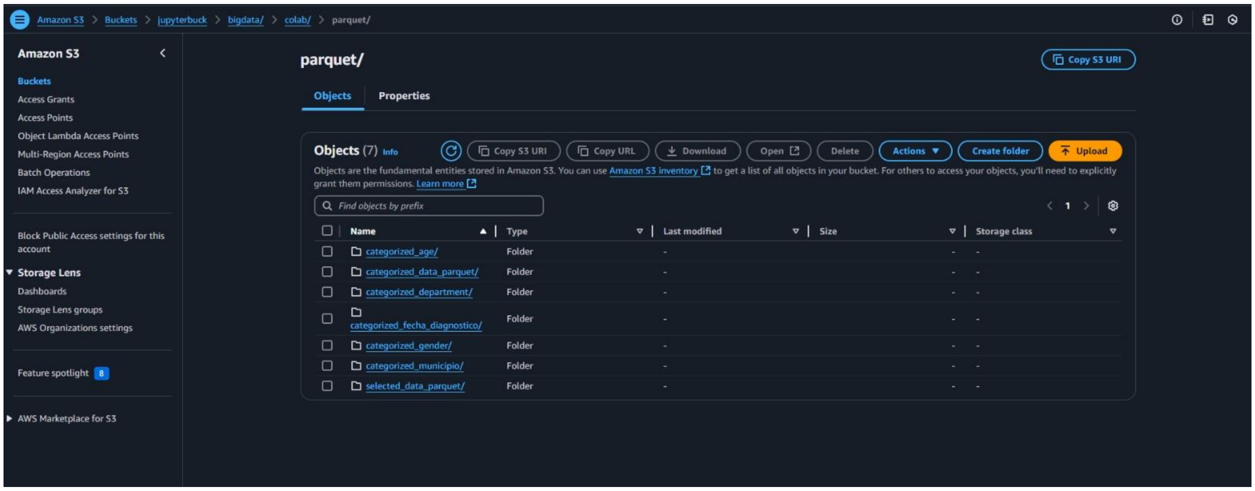


csv/

Objects (7) Info

Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

□	Name	Type	Last modified	Size	Storage class
□	categorized_age/	Folder	-	-	-
□	categorized_data_csv/	Folder	-	-	-
□	categorized_department/	Folder	-	-	-
□	categorized_fecha_diagnostico/	Folder	-	-	-
□	categorized_gender/	Folder	-	-	-
□	categorized_municipio/	Folder	-	-	-
□	selected_data.csv/	Folder	-	-	-



parquet/

Objects (7) Info

Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

□	Name	Type	Last modified	Size	Storage class
□	categorized_age/	Folder	-	-	-
□	categorized_data_parquet/	Folder	-	-	-
□	categorized_department/	Folder	-	-	-
□	categorized_fecha_diagnostico/	Folder	-	-	-
□	categorized_gender/	Folder	-	-	-
□	categorized_municipio/	Folder	-	-	-
□	selected_data_parquet/	Folder	-	-	-

S3 uri: s3://jupyterbuck/bigdata/colab/