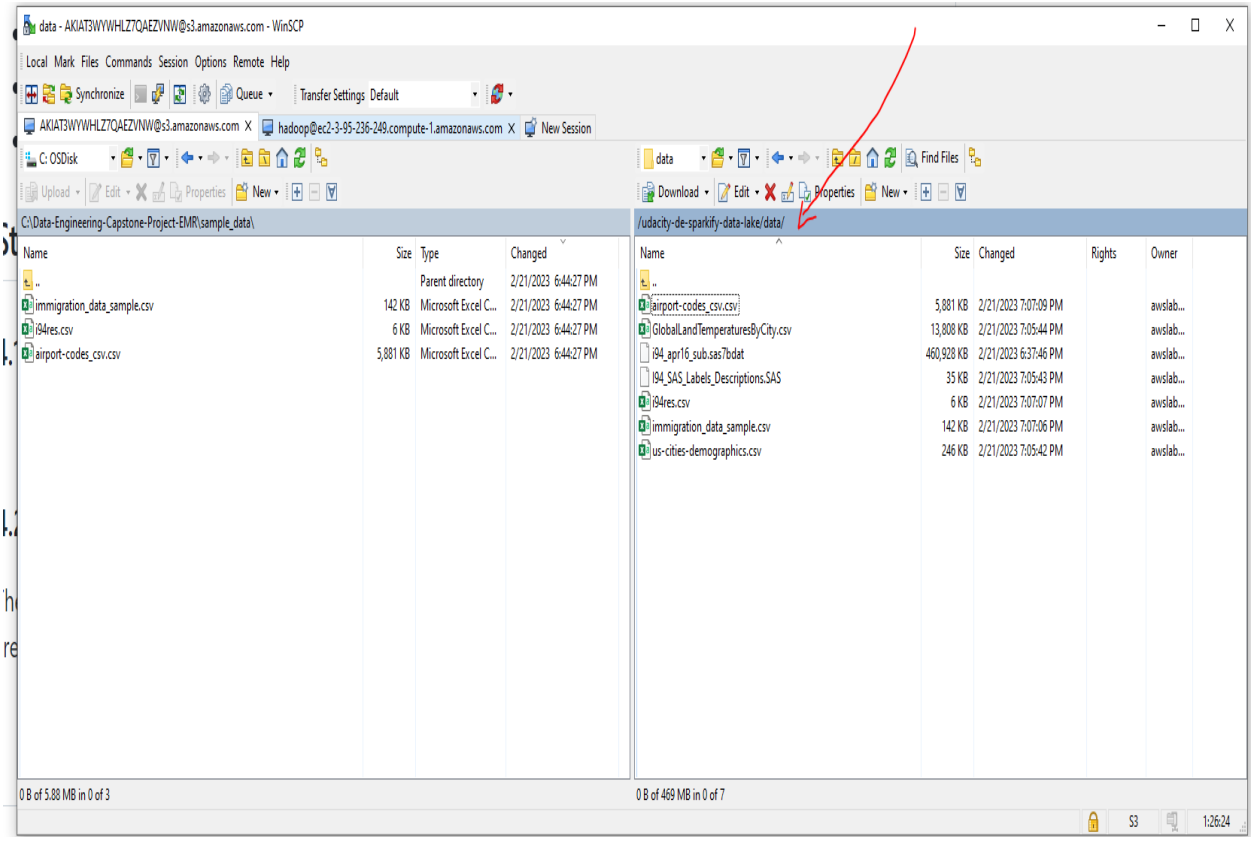
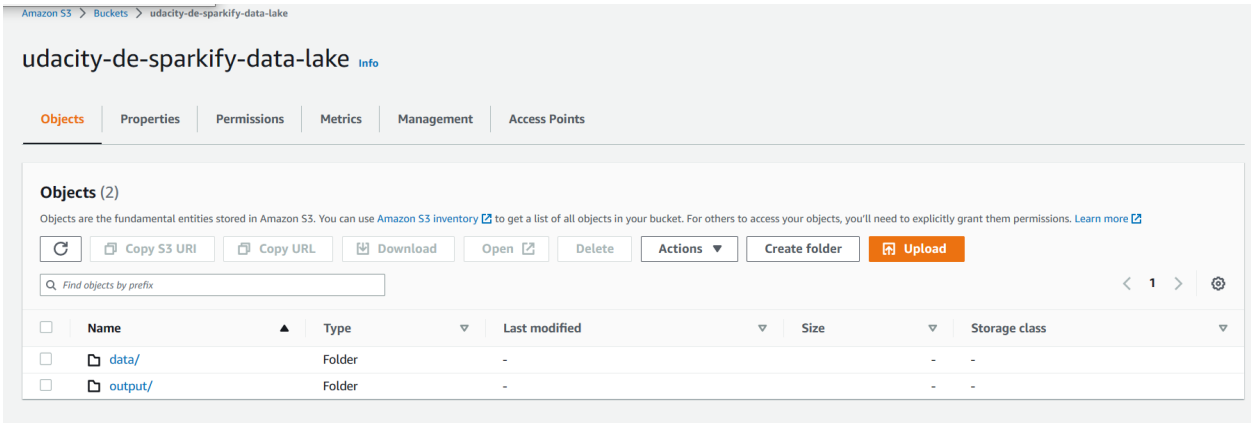
Additional Details:

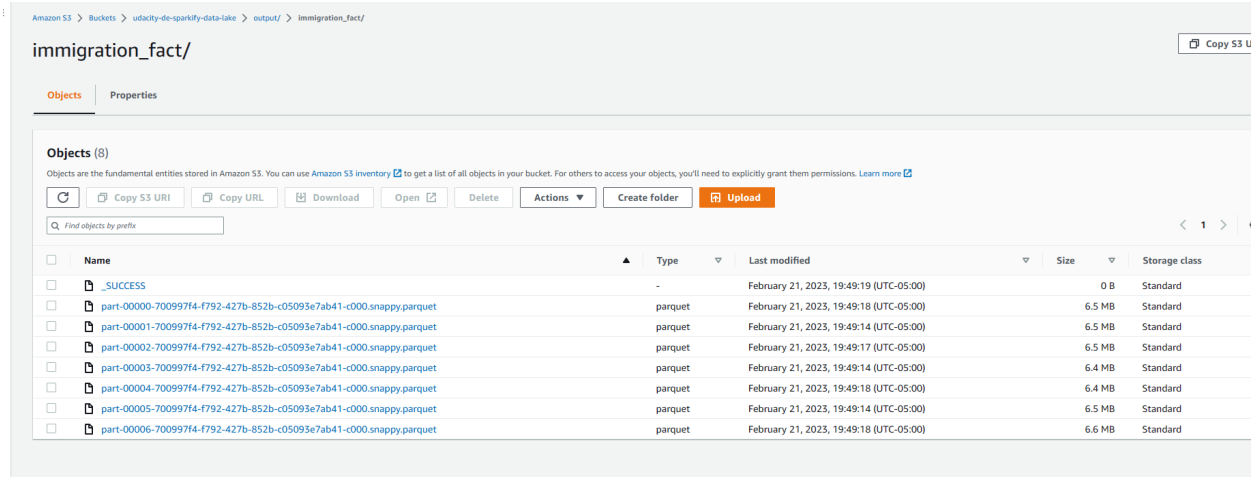
Copy the Data file to S3 bucket

1) Create a bucket using AWS S3 portal and folders inside the bucket to organize the data and output
2) Use winscp tool to connect to S3 using private key
3) Drag and drop the files to the Data folder

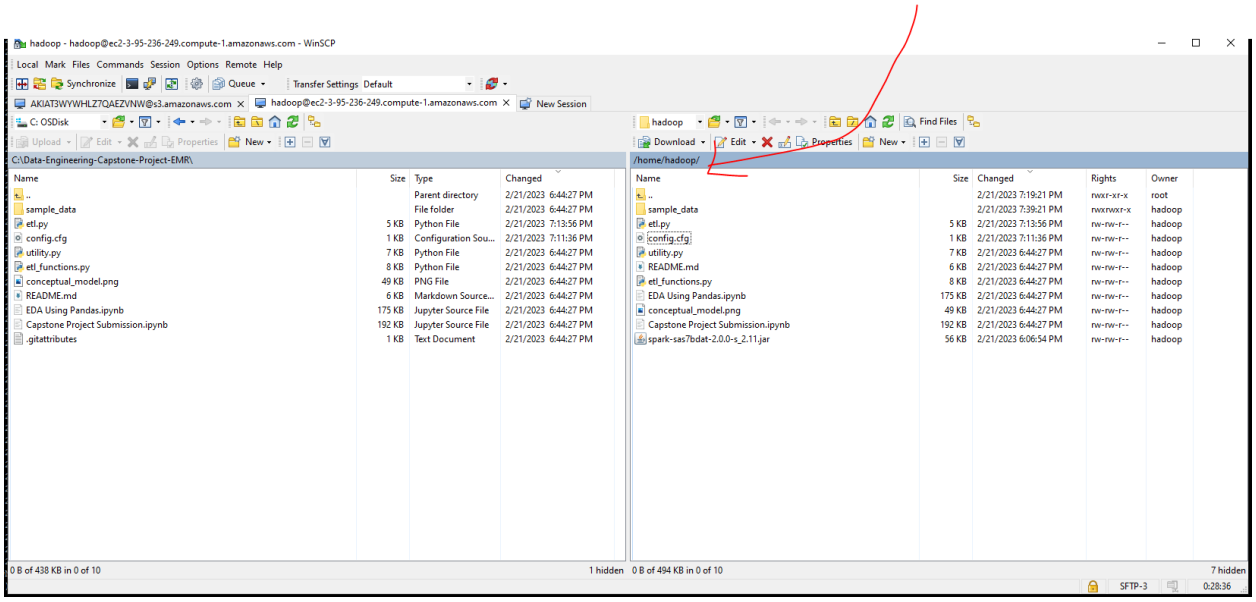


S3 Bucket

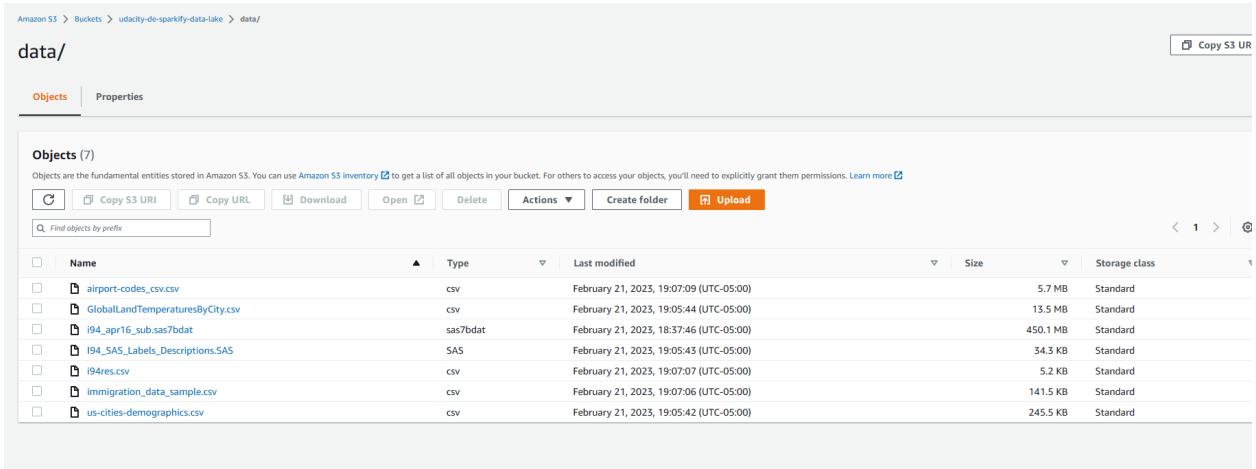




Copy the project files (etl.py and supporting files) to Hadoop home directory



Data Files in S3



How to run:

Run on Master node of AWS EMR cluster

1) Putty to the Master node

2) Run the following command , some of the libraries are not in EMR so you may need to do pip install library name

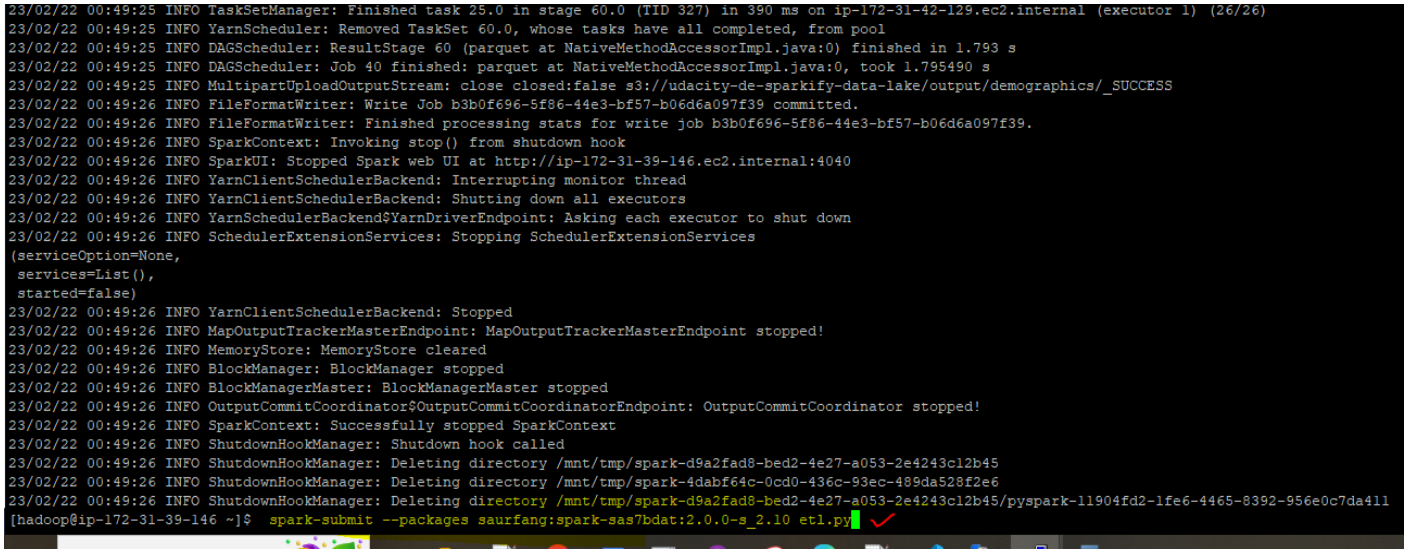
pip install pandas

pip install seaborn

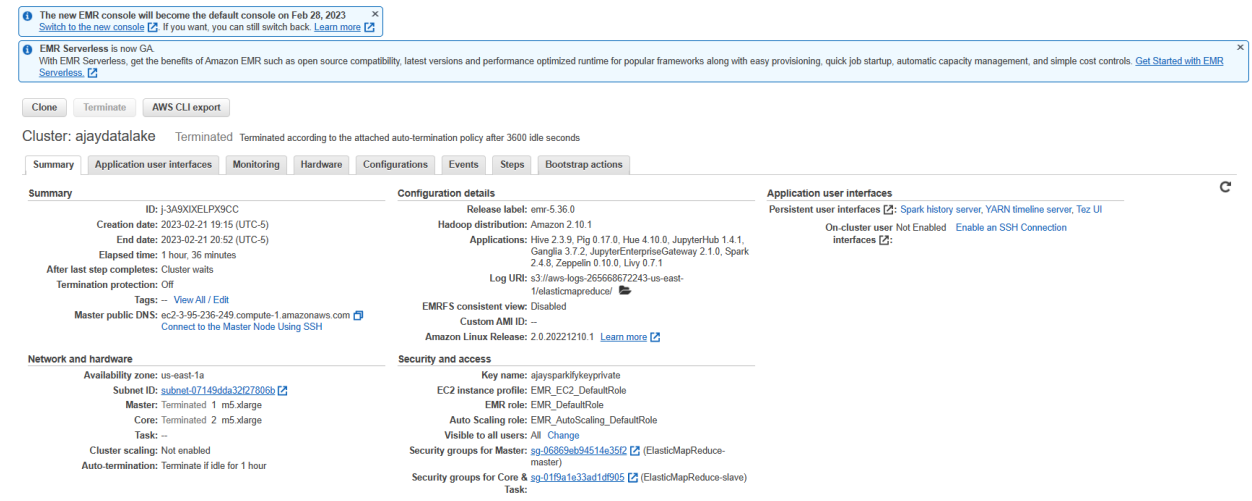
pip install requests

3) I was getting some issue with sas7bdat so I download the jar file also and kept in the hadoop home directory

[hadoop@ip-172-31-39-146 ~]$ spark-submit --packages saurfang:spark-sas7bdat:2.0.0-s_2.10 etl.py

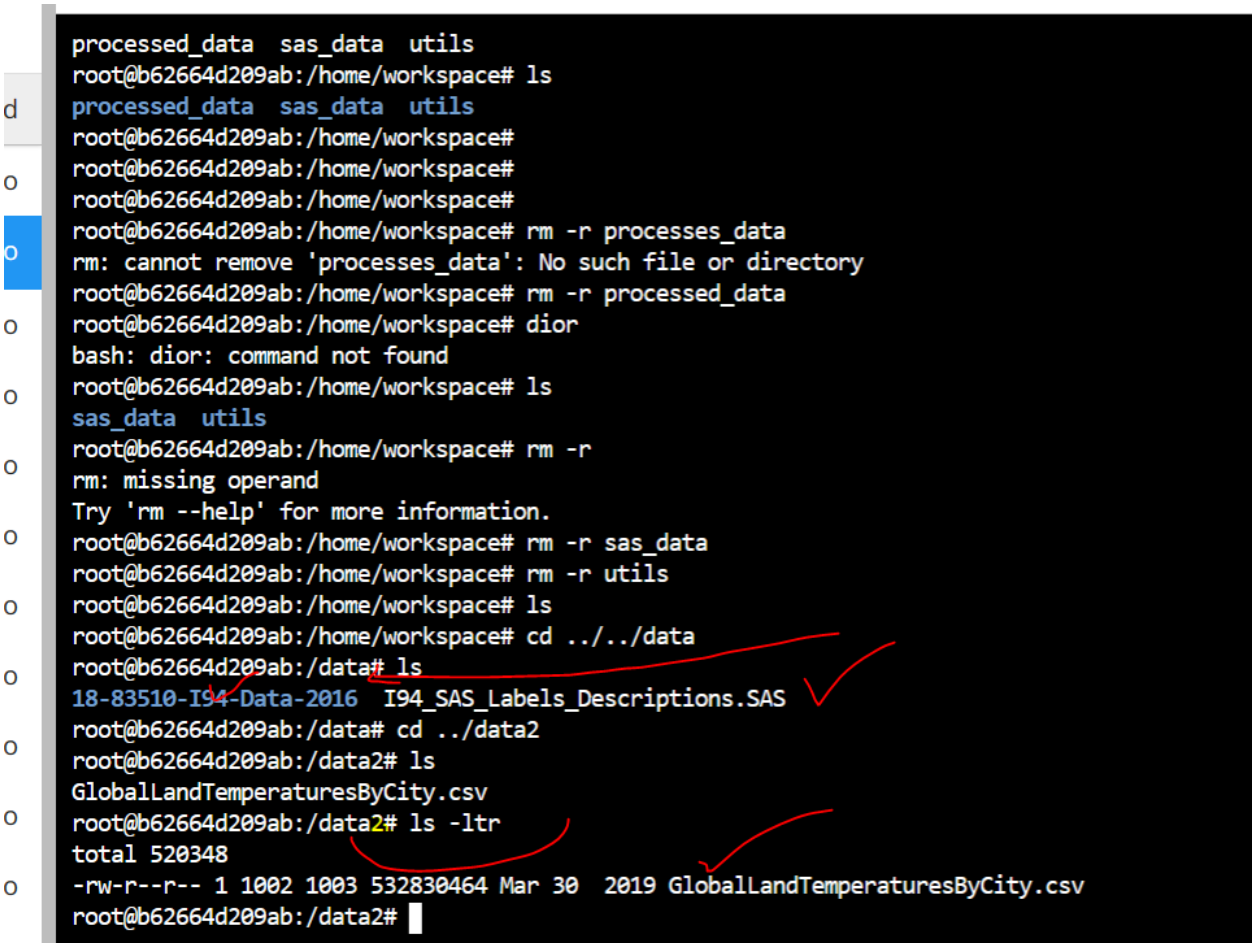


AWS EMR Cluster



Other files directories:

Open terminal
Go to data and data2 directories to see the files


processed_data sas_data utils
root@b62664d209ab:/home/workspace# ls
processed_data sas_data utils
root@b62664d209ab:/home/workspace#
root@b62664d209ab:/home/workspace#
root@b62664d209ab:/home/workspace#
root@b62664d209ab:/home/workspace# rm -r processes_data
rm: cannot remove 'processes_data': No such file or directory
root@b62664d209ab:/home/workspace# rm -r processed_data
root@b62664d209ab:/home/workspace# dior
bash: dior: command not found
root@b62664d209ab:/home/workspace# ls
sas_data utils
root@b62664d209ab:/home/workspace# rm -r
rm: missing operand
Try 'rm --help' for more information.
root@b62664d209ab:/home/workspace# rm -r sas_data
root@b62664d209ab:/home/workspace# rm -r utils
root@b62664d209ab:/home/workspace# ls
root@b62664d209ab:/home/workspace# cd ../../data
root@b62664d209ab:/data# ls
18-83510-I94-Data-2016 I94_SAS_Labels_Descriptions.SAS
root@b62664d209ab:/data# cd ../data2
root@b62664d209ab:/data2# ls
GlobalLandTemperaturesByCity.csv
root@b62664d209ab:/data2# ls -ltr
total 520348
-rw-r--r-- 1 1002 1003 532830464 Mar 30 2019 GlobalLandTemperaturesByCity.csv
root@b62664d209ab:/data2#