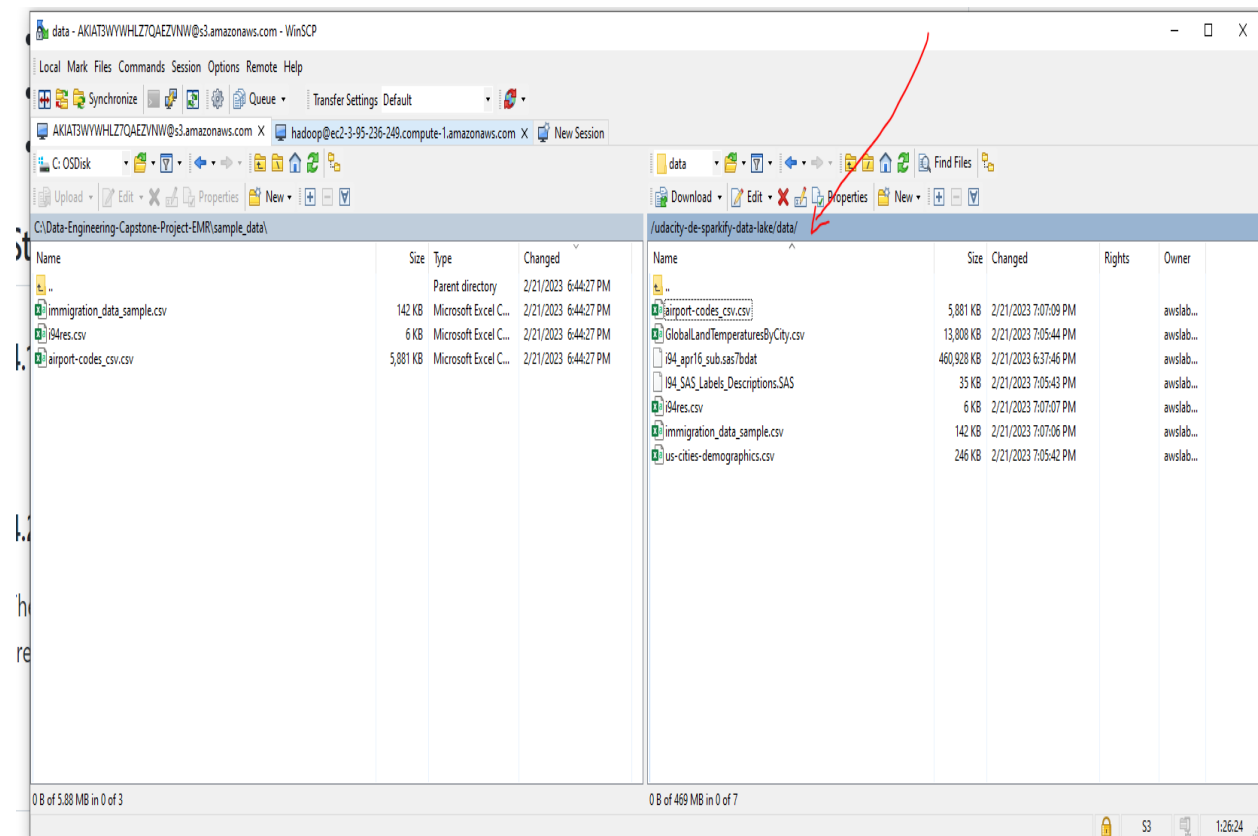


Additional Details:

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

Amazon S3 > Buckets > udacity-de-sparkify-data-lake

udacity-de-sparkify-data-lake [Info](#)

[Objects](#) | [Properties](#) | [Permissions](#) | [Metrics](#) | [Management](#) | [Access Points](#)

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](#)

[Refresh](#) [Copy S3 URI](#) [Copy URL](#) [Download](#) [Open](#) [Delete](#) [Actions](#) [Create folder](#) [Upload](#)

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	data/	Folder	-	-	-
<input type="checkbox"/>	output/	Folder	-	-	-

Amazon S3 > Buckets > udacity-de-sparkify-data-lake > output/ > immigration_fact/

immigration_fact/ [Copy S3 URI](#)

[Objects](#) | [Properties](#)

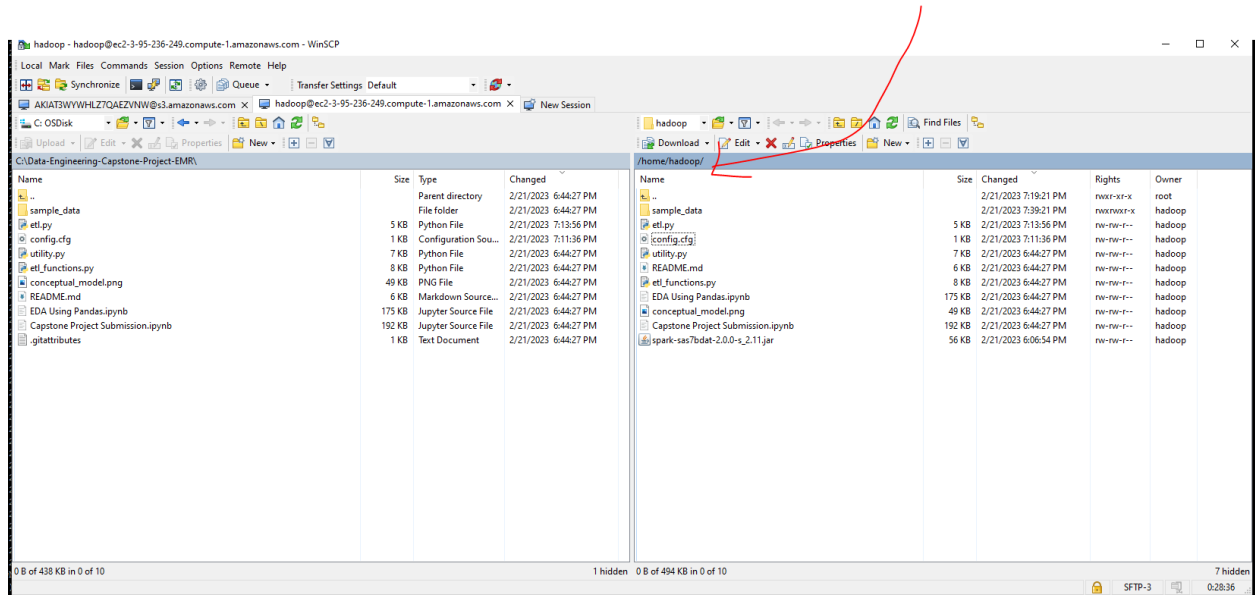
Objects (8)

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](#)

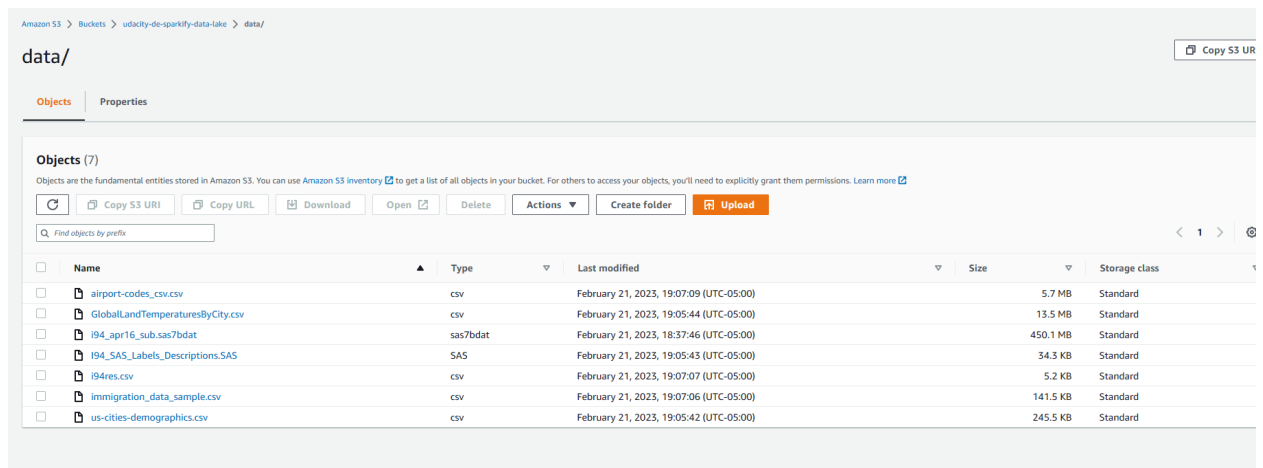
[Refresh](#) [Copy S3 URI](#) [Copy URL](#) [Download](#) [Open](#) [Delete](#) [Actions](#) [Create folder](#) [Upload](#)

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	_SUCCESS	-	February 21, 2023, 19:49:19 (UTC-05:00)	0 B	Standard
<input type="checkbox"/>	part-00000-700997f4-f792-427b-852b-c05093e7ab41-c000.snappy.parquet	parquet	February 21, 2023, 19:49:18 (UTC-05:00)	6.5 MB	Standard
<input type="checkbox"/>	part-00001-700997f4-f792-427b-852b-c05093e7ab41-c000.snappy.parquet	parquet	February 21, 2023, 19:49:14 (UTC-05:00)	6.5 MB	Standard
<input type="checkbox"/>	part-00002-700997f4-f792-427b-852b-c05093e7ab41-c000.snappy.parquet	parquet	February 21, 2023, 19:49:17 (UTC-05:00)	6.5 MB	Standard
<input type="checkbox"/>	part-00003-700997f4-f792-427b-852b-c05093e7ab41-c000.snappy.parquet	parquet	February 21, 2023, 19:49:18 (UTC-05:00)	6.4 MB	Standard
<input type="checkbox"/>	part-00004-700997f4-f792-427b-852b-c05093e7ab41-c000.snappy.parquet	parquet	February 21, 2023, 19:49:18 (UTC-05:00)	6.4 MB	Standard
<input type="checkbox"/>	part-00005-700997f4-f792-427b-852b-c05093e7ab41-c000.snappy.parquet	parquet	February 21, 2023, 19:49:14 (UTC-05:00)	6.5 MB	Standard
<input type="checkbox"/>	part-00006-700997f4-f792-427b-852b-c05093e7ab41-c000.snappy.parquet	parquet	February 21, 2023, 19:49:18 (UTC-05:00)	6.6 MB	Standard

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
```

```
23/02/22 00:49:25 INFO TaskSetManager: Finished task 25.0 in stage 60.0 (TID 327) in 390 ms on ip-172-31-42-129.ec2.internal (executor 1) (26/26)
23/02/22 00:49:25 INFO YarnScheduler: Removed TaskSet 60.0, whose tasks have all completed, from pool
23/02/22 00:49:25 INFO DAGScheduler: ResultStage 60 (parquet at NativeMethodAccessorImpl.java:0) finished in 1.793 s
23/02/22 00:49:25 INFO DAGScheduler: Job 40 finished: parquet at NativeMethodAccessorImpl.java:0, took 1.795490 s
23/02/22 00:49:25 INFO MultipartUploadOutputStream: close closed:false s3://udacity-de-sparkify-data-lake/output/demographics/_SUCCESS
23/02/22 00:49:26 INFO FileFormatWriter: Write Job b3b0f696-5f86-44e3-bf57-b06d6a097f39 committed.
23/02/22 00:49:26 INFO FileFormatWriter: Finished processing stats for write job b3b0f696-5f86-44e3-bf57-b06d6a097f39.
23/02/22 00:49:26 INFO SparkContext: Invoking stop() from shutdown hook
23/02/22 00:49:26 INFO SparkUI: Stopped Spark web UI at http://ip-172-31-39-146.ec2.internal:4040
23/02/22 00:49:26 INFO YarnClientSchedulerBackend: Interrupting monitor thread
23/02/22 00:49:26 INFO YarnClientSchedulerBackend: Shutting down all executors
23/02/22 00:49:26 INFO YarnSchedulerBackend$YarnDriverEndpoint: Asking each executor to shut down
23/02/22 00:49:26 INFO SchedulerExtensionServices: Stopping SchedulerExtensionServices
(serviceOption=None,
services=List(),
started=false)
23/02/22 00:49:26 INFO YarnClientSchedulerBackend: Stopped
23/02/22 00:49:26 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
23/02/22 00:49:26 INFO MemoryStore: MemoryStore cleared
23/02/22 00:49:26 INFO BlockManager: BlockManager stopped
23/02/22 00:49:26 INFO BlockManagerMaster: BlockManagerMaster stopped
23/02/22 00:49:26 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
23/02/22 00:49:26 INFO SparkContext: Successfully stopped SparkContext
23/02/22 00:49:26 INFO ShutdownHookManager: Shutdown hook called
23/02/22 00:49:26 INFO ShutdownHookManager: Deleting directory /mnt/tmp/spark-d9a2fad8-bed2-4e27-a053-2e4243c12b45
23/02/22 00:49:26 INFO ShutdownHookManager: Deleting directory /mnt/tmp/spark-4dabf64c-0cd0-436c-93ec-489da528f2e6
23/02/22 00:49:26 INFO ShutdownHookManager: Deleting directory /mnt/tmp/spark-d9a2fad8-bed2-4e27-a053-2e4243c12b45/pyspark-11904fd2-1fe6-4465-8392-956e0c7da411
[hadoop@ip-172-31-39-146 ~]$ spark-submit --packages saurfang:spark-sas7bdat:2.0.0-s_2.10 etl.py ✓
```

AWS EMR Cluster

The new EMR console will become the default console on Feb 28, 2023
[Switch to the new console](#) If you want, you can still switch back [Learn more](#)

EMR Serverless is now GA.
With EMR Serverless, get the benefits of Amazon EMR such as open source compatibility, latest versions and performance optimized runtime for popular frameworks along with easy provisioning, quick job startup, automatic capacity management, and simple cost controls. [Get Started with EMR Serverless](#)

Clone Terminate AWS CLI export

Cluster: **ajaydatalake** Terminated Terminated according to the attached auto-termination policy after 3600 idle seconds

Summary Application user interfaces Monitoring Hardware Configurations Events Steps Bootstrap actions

Summary

ID: j-3A9XIXELPX9CC
Creation date: 2023-02-21 19:15 (UTC-5)
End date: 2023-02-21 20:52 (UTC-5)
Elapsed time: 1 hour, 36 minutes
After last step completes: Cluster waits
Termination protection: Off
Tags: -- [View All](#) / [Edit](#)
Master public DNS: ec2-3-95-236-249.compute-1.amazonaws.com
[Connect to the Master Node Using SSH](#)

Network and hardware

Availability zone: us-east-1a
Subnet ID: [subnet-07149d4e32d77800b](#)
Master: Terminated 1 m5.xlarge
Core: Terminated 2 m5.xlarge
Task: --
Cluster scaling: Not enabled
Auto-termination: Terminate if idle for 1 hour

Configuration details

Release label: emr-5.36.0
Hadoop distribution: Amazon 2.10.1
Applications: Hive 2.3.9, Pig 0.17.0, Hue 4.10.0, JupyterHub 1.4.1, Ganglia 3.7.2, JupyterEnterpriseGateway 2.1.0, Spark 2.4.8, Zeppelin 0.10.0, Livy 0.7.1
Log URI: s3://aws-logs-265668672243-us-east-1/elasticmapreduce/
EMRFS consistent view: Disabled
Custom AMI ID: --
Amazon Linux Release: 2.0.20221210.1 [Learn more](#)

Security and access

Key name: sjaysparkify/keys/private
EC2 instance profile: EMR_EC2_DefaultRole
EMR role: EMR_DefaultRole
Auto Scaling role: EMR_AutoScaling_DefaultRole
Visible to all users: All [Change](#)
Security groups for Master: [sg-06869eb94514e35f2](#) (ElasticMapReduce-master)
Security groups for Core & [sg-01f9a1e33ad1d905](#) (ElasticMapReduce-slave)
Task:

Application user interfaces

Persistent user interfaces [Spark history server](#), [YARN timeline server](#), [Tez UI](#)
On-cluster user interfaces [Not Enabled](#) [Enable an SSH Connection](#)

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# 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#
```