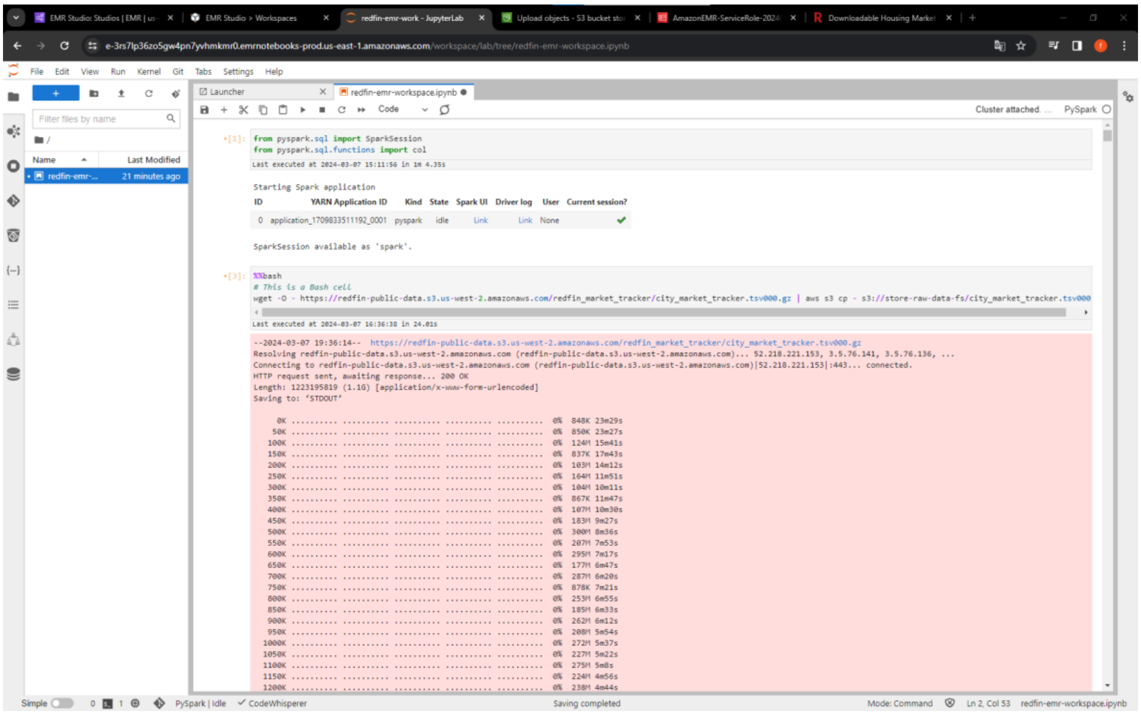

EMR Studio: Studios | EMR | us
EMR Studio > Workspaces
redfin-emr-work - JupyterLab
Upload objects - S3 bucket sto
AmazonEMR-ServiceRole-2024
Downloadable Housing Market
e-3rs7lp36zo5gw4pn7yvhmkmr0.emrnotebooks-prod.us-east-1.amazonaws.com/workspace/lab/tree/redfin-emr-workspace.ipynb
File Edit View Run Kernel Git Tabs Settings Help
Filter files by name
Name
Last Modified
redfin-emr-...
21 minutes ago
Launcher
redfin-emr-workspace.ipynb
Code
Cluster attached. ...
PySpark
Simple
0
1
PySpark | Idle
✓ CodeWhisperer
Saving completed
Mode: Command
Ln 2, Col 53
redfin-emr-workspace.ipynb

```
•[1]: from pyspark.sql import SparkSession
      from pyspark.sql.functions import col
      Last executed at 2024-03-07 15:11:56 in 1m 4.35s
```

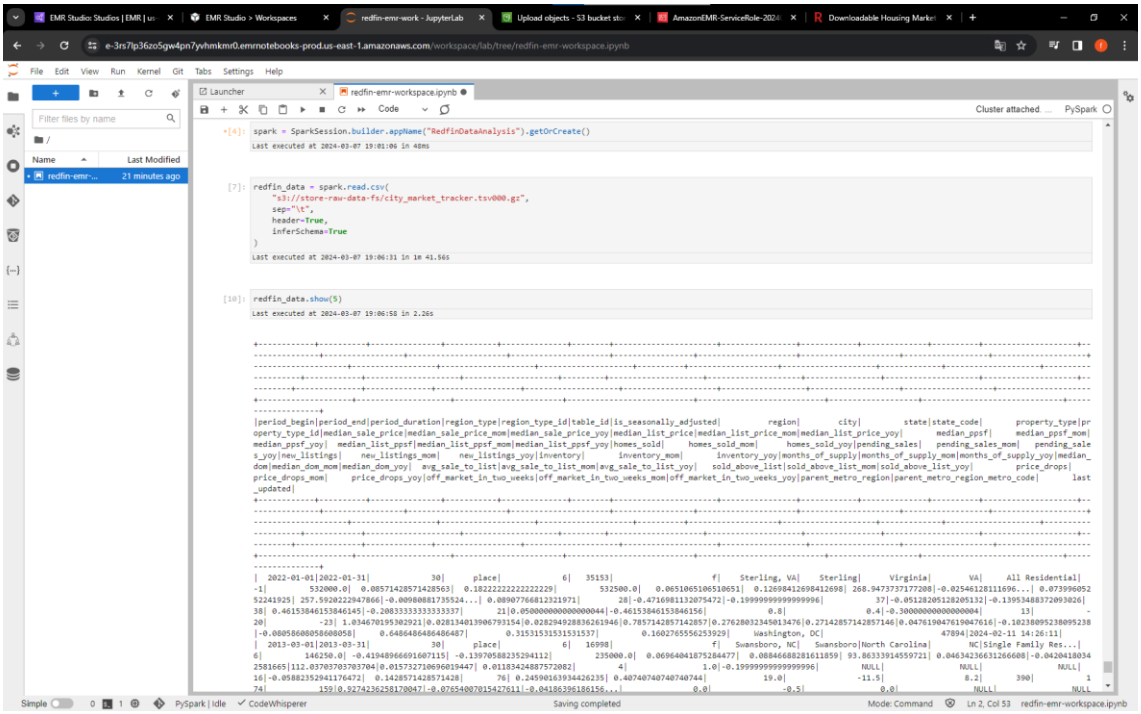
Starting Spark application

ID	YARN Application ID	Kind	State	Spark UI	Driver log	User	Current session?
0	application_1709833511192_0001	pyspark	idle	Link	Link	None	✓

SparkSession available as 'spark'.

```
•[3]: %%bash
      # This is a Bash cell
      wget -O - https://redfin-public-data.s3.us-west-2.amazonaws.com/redfin_market_tracker/city_market_tracker.tsv000.gz | aws s3 cp - s3://store-raw-data-fs/city_market_tracker.tsv000
      Last executed at 2024-03-07 16:36:38 in 24.01s
```

```
--2024-03-07 19:36:14--  https://redfin-public-data.s3.us-west-2.amazonaws.com/redfin_market_tracker/city_market_tracker.tsv000.gz
Resolving redfin-public-data.s3.us-west-2.amazonaws.com (redfin-public-data.s3.us-west-2.amazonaws.com)... 52.218.221.153, 3.5.76.141, 3.5.76.136, ...
Connecting to redfin-public-data.s3.us-west-2.amazonaws.com (redfin-public-data.s3.us-west-2.amazonaws.com)|52.218.221.153|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 1223195819 (1.1G) [application/x-www-form-urlencoded]
Saving to: 'STDOUT'

     0K .......... .......... .......... .......... ..........  0%  848K 23m29s
    50K .......... .......... .......... .......... ..........  0%  850K 23m27s
   100K .......... .......... .......... .......... ..........  0%  124M 15m41s
   150K .......... .......... .......... .......... ..........  0%  837K 17m43s
   200K .......... .......... .......... .......... ..........  0%  103M 14m12s
   250K .......... .......... .......... .......... ..........  0%  164M 11m51s
   300K .......... .......... .......... .......... ..........  0%  104M 10m11s
   350K .......... .......... .......... .......... ..........  0%  867K 11m47s
   400K .......... .......... .......... .......... ..........  0%  107M 10m30s
   450K .......... .......... .......... .......... ..........  0%  183M 9m27s
   500K .......... .......... .......... .......... ..........  0%  300M 8m36s
   550K .......... .......... .......... .......... ..........  0%  207M 7m53s
   600K .......... .......... .......... .......... ..........  0%  295M 7m17s
   650K .......... .......... .......... .......... ..........  0%  177M 6m47s
   700K .......... .......... .......... .......... ..........  0%  287M 6m20s
   750K .......... .......... .......... .......... ..........  0%  878K 7m21s
   800K .......... .......... .......... .......... ..........  0%  253M 6m55s
   850K .......... .......... .......... .......... ..........  0%  185M 6m33s
   900K .......... .......... .......... .......... ..........  0%  262M 6m12s
   950K .......... .......... .......... .......... ..........  0%  208M 5m54s
  1000K .......... .......... .......... .......... ..........  0%  272M 5m37s
  1050K .......... .......... .......... .......... ..........  0%  227M 5m22s
  1100K .......... .......... .......... .......... ..........  0%  275M 5m8s
  1150K .......... .......... .......... .......... ..........  0%  224M 4m56s
  1200K .......... .......... .......... .......... ..........  0%  238M 4m44s
```

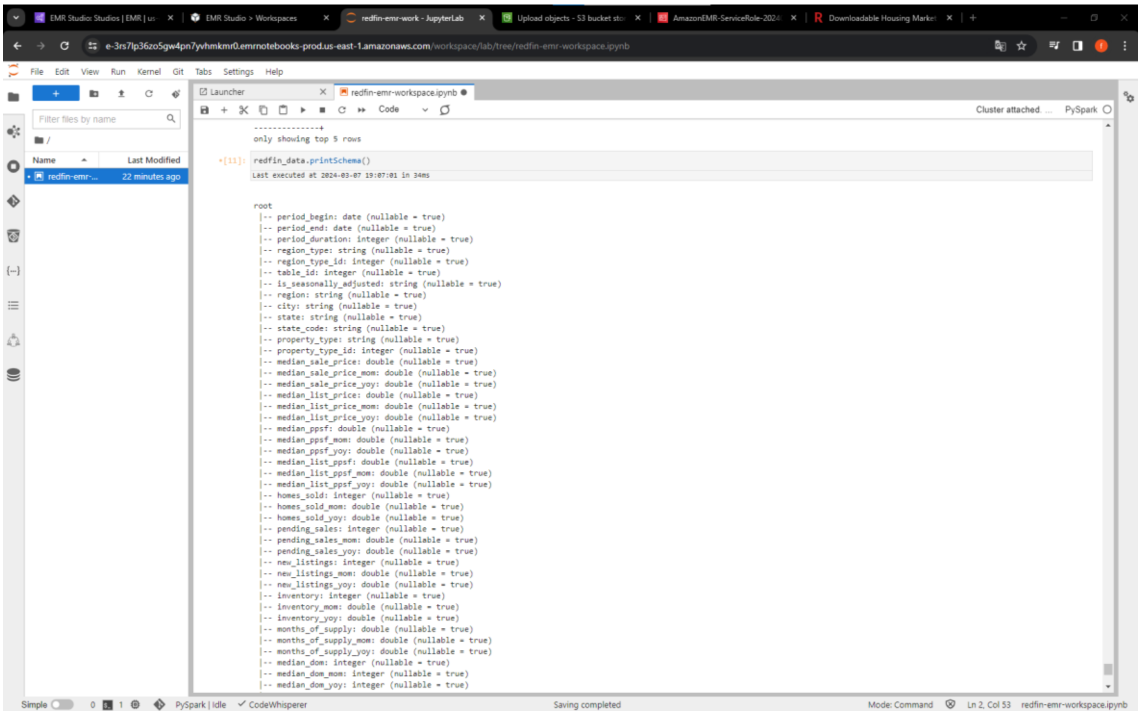

EMR Studio: Studios | EMR | us
EMR Studio > Workspaces
redfin-emr-work - JupyterLab
Upload objects - S3 bucket sto
AmazonEMR-ServiceRole-2024
Downloadable Housing Market
e-3rs7lp36zo5gw4pn7yvhmkmr0.emrnotebooks-prod.us-east-1.amazonaws.com/workspace/lab/tree/redfin-emr-workspace.ipynb
File Edit View Run Kernel Git Tabs Settings Help
Filter files by name
Name
Last Modified
redfin-emr-...
21 minutes ago
Launcher
redfin-emr-workspace.ipynb
Code
Cluster attached. ...
PySpark
Simple
0
1
PySpark | Idle
✓ CodeWhisperer
Saving completed
Mode: Command
Ln 2, Col 53
redfin-emr-workspace.ipynb

```
•[4]: spark = SparkSession.builder.appName("RedfinDataAnalysis").getOrCreate()
      Last executed at 2024-03-07 19:01:06 in 48ms
```

```
 [7]: redfin_data = spark.read.csv(
          "s3://store-raw-data-fs/city_market_tracker.tsv000.gz",
          sep="\t",
          header=True,
          inferSchema=True
      )
      Last executed at 2024-03-07 19:06:31 in 1m 41.56s
```

```
[10]: redfin_data.show(5)
      Last executed at 2024-03-07 19:06:58 in 2.26s
```

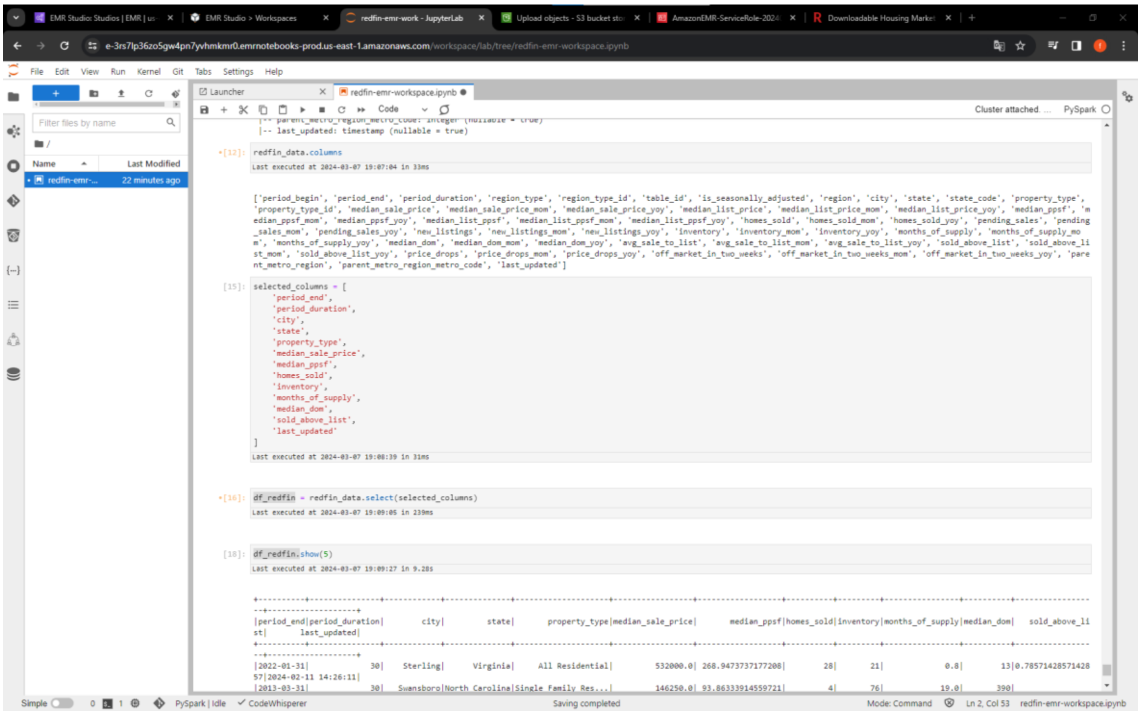
```
+------------+----------+---------------+-----------+--------------+--------+----------------------+----------------+------------+--------------+----------+--------------------+--
-------------+-----------------+---------------------+---------------------+-----------------+---------------------+---------------------+--------------------+--------------------+
--------------------+--------------------+---------------------+---------------------+----------+--------------------+--------------------+-------------+--------------------+----------
----------+--------------------+---------------------+---------+--------------------+--------------------+----------------+--------------------+--------------------+---------------+--
--------+--------------+--------------+-------------------+--------------------+--------------------+--------------------+-------------------+-------------------+--------------------+
+--------------------+--------------------+-----------------------+---------------------------+---------------------------+-------------------+------------------------------+-------
-------------+
|period_begin|period_end|period_duration|region_type|region_type_id|table_id|is_seasonally_adjusted|          region|        city|         state|state_code|       property_type|pr
operty_type_id|median_sale_price|median_sale_price_mom|median_sale_price_yoy|median_list_price|median_list_price_mom|median_list_price_yoy|         median_ppsf|     median_ppsf_mom|
median_ppsf_yoy|  median_list_ppsf|median_list_ppsf_mom|median_list_ppsf_yoy|homes_sold|      homes_sold_mom|      homes_sold_yoy|pending_sales|   pending_sales_mom|   pending_sale
s_yoy|new_listings|     new_listings_mom|     new_listings_yoy|inventory|       inventory_mom|       inventory_yoy|months_of_supply|months_of_supply_mom|months_of_supply_yoy|median_
dom|median_dom_mom|median_dom_yoy|   avg_sale_to_list|avg_sale_to_list_mom|avg_sale_to_list_yoy|   sold_above_list|sold_above_list_mom|sold_above_list_yoy|         price_drops|
price_drops_mom|      price_drops_yoy|off_market_in_two_weeks|off_market_in_two_weeks_mom|off_market_in_two_weeks_yoy|parent_metro_region|parent_metro_region_metro_code|        last
_updated|
+------------+----------+---------------+-----------+--------------+--------+----------------------+----------------+------------+--------------+----------+--------------------+--
-------------+-----------------+---------------------+---------------------+-----------------+---------------------+---------------------+--------------------+--------------------+
--------------------+--------------------+---------------------+---------------------+----------+--------------------+--------------------+-------------+--------------------+----------
----------+--------------------+---------------------+---------+--------------------+--------------------+----------------+--------------------+--------------------+---------------+--
--------+--------------+--------------+-------------------+--------------------+--------------------+--------------------+-------------------+-------------------+--------------------+
+--------------------+--------------------+-----------------------+---------------------------+---------------------------+-------------------+------------------------------+-------
-------------+
|  2022-01-01|2022-01-31|             30|      place|             6|   35153|                     f|    Sterling, VA|    Sterling|      Virginia|        VA|     All Residential|
-1|         532000.0|  0.08571428571428563|  0.18222222222222229|         532500.0|   0.0651065106510651|  0.12698412698412698| 268.9473737177208|-0.02546128111696...| 0.073996052
52241925| 257.5920222947866|-0.00980881735524...| 0.08907766812321971|        28|-0.4716981132075472|-0.19999999999999996|           37|-0.05128205128205132|-0.13953488372093026|
38| 0.46153846153846145|-0.20833333333333337|        21|0.050000000000000044|-0.46153846153846156|             0.8|                 0.4|-0.30000000000000004|         13|              -
20|           -23| 1.034670195302921|0.028134013906793154|0.028294928836261946|0.7857142857142857|0.27628032345013476|0.27142857142857146|0.047619047619047616|-0.10238095238095238
|-0.08058608058608058|     0.6486486486486487|      0.31531531531531537|        0.1602765556253929|     Washington, DC|                         47894|2024-02-11 14:26:11|
|  2013-03-01|2013-03-31|             30|      place|             6|   16998|                     f|   Swansboro, NC|   Swansboro|North Carolina|        NC|Single Family Res...|
6|         146250.0| -0.41948966691607115| -0.13970588235294112|         235000.0|  0.06964041875284477|  0.08846688281611859| 93.86333914559721| 0.04634236631266608|-0.0420418034
2581665|112.03703703703704|0.015732710696019447| 0.01183424887572082|         4|                 1.0|-0.19999999999999996|         NULL|                NULL|                NULL|
16|-0.05882352941176472|  0.1428571428571428|        76| 0.24590163934426235| 0.40740740740740744|            19.0|               -11.5|                 8.2|        390|              1
74|           159|0.9274236258170047|-0.07654007015427611|-0.04186396186156...|               0.0|               -0.5|                0.0|                NULL|                NULL
```


EMR Studio: Studios | EMR | us
EMR Studio > Workspaces
redfin-emr-work - JupyterLab
Upload objects - S3 bucket sto
AmazonEMR-ServiceRole-2024
Downloadable Housing Market
e-3rs7lp36zo5gw4pn7yvhmkmr0.emrnotebooks-prod.us-east-1.amazonaws.com/workspace/lab/tree/redfin-emr-workspace.ipynb
File
Edit
View
Run
Kernel
Git
Tabs
Settings
Help
Filter files by name
Name
Last Modified
redfin-emr-...
22 minutes ago
Launcher
redfin-emr-workspace.ipynb
Code
Cluster attached. ...
PySpark
Simple
PySpark | Idle
CodeWhisperer
Saving completed
Mode: Command
Ln 2, Col 53
redfin-emr-workspace.ipynb

```
--------------+
only showing top 5 rows
```

```
[11]: redfin_data.printSchema()
Last executed at 2024-03-07 19:07:01 in 34ms
```

```
root
 |-- period_begin: date (nullable = true)
 |-- period_end: date (nullable = true)
 |-- period_duration: integer (nullable = true)
 |-- region_type: string (nullable = true)
 |-- region_type_id: integer (nullable = true)
 |-- table_id: integer (nullable = true)
 |-- is_seasonally_adjusted: string (nullable = true)
 |-- region: string (nullable = true)
 |-- city: string (nullable = true)
 |-- state: string (nullable = true)
 |-- state_code: string (nullable = true)
 |-- property_type: string (nullable = true)
 |-- property_type_id: integer (nullable = true)
 |-- median_sale_price: double (nullable = true)
 |-- median_sale_price_mom: double (nullable = true)
 |-- median_sale_price_yoy: double (nullable = true)
 |-- median_list_price: double (nullable = true)
 |-- median_list_price_mom: double (nullable = true)
 |-- median_list_price_yoy: double (nullable = true)
 |-- median_ppsf: double (nullable = true)
 |-- median_ppsf_mom: double (nullable = true)
 |-- median_ppsf_yoy: double (nullable = true)
 |-- median_list_ppsf: double (nullable = true)
 |-- median_list_ppsf_mom: double (nullable = true)
 |-- median_list_ppsf_yoy: double (nullable = true)
 |-- homes_sold: integer (nullable = true)
 |-- homes_sold_mom: double (nullable = true)
 |-- homes_sold_yoy: double (nullable = true)
 |-- pending_sales: integer (nullable = true)
 |-- pending_sales_mom: double (nullable = true)
 |-- pending_sales_yoy: double (nullable = true)
 |-- new_listings: integer (nullable = true)
 |-- new_listings_mom: double (nullable = true)
 |-- new_listings_yoy: double (nullable = true)
 |-- inventory: integer (nullable = true)
 |-- inventory_mom: double (nullable = true)
 |-- inventory_yoy: double (nullable = true)
 |-- months_of_supply: double (nullable = true)
 |-- months_of_supply_mom: double (nullable = true)
 |-- months_of_supply_yoy: double (nullable = true)
 |-- median_dom: integer (nullable = true)
 |-- median_dom_mom: integer (nullable = true)
 |-- median_dom_yoy: integer (nullable = true)
```


EMR Studio: Studios | EMR | us
EMR Studio > Workspaces
redfin-emr-work - JupyterLab
Upload objects - S3 bucket sto
AmazonEMR-ServiceRole-2024
Downloadable Housing Market
e-3rs7lp36zo5gw4pn7yvhmkmr0.emrnotebooks-prod.us-east-1.amazonaws.com/workspace/lab/tree/redfin-emr-workspace.ipynb
File
Edit
View
Run
Kernel
Git
Tabs
Settings
Help
Filter files by name
Name
Last Modified
redfin-emr-...
22 minutes ago
Launcher
redfin-emr-workspace.ipynb
Code
Cluster attached. ...
PySpark
Simple
PySpark | Idle
CodeWhisperer
Saving completed
Mode: Command
Ln 2, Col 53
redfin-emr-workspace.ipynb

```
 |-- parent_metro_region_metro_code: integer (nullable = true)
 |-- last_updated: timestamp (nullable = true)
```

```
[12]: redfin_data.columns
Last executed at 2024-03-07 19:07:04 in 33ms
```

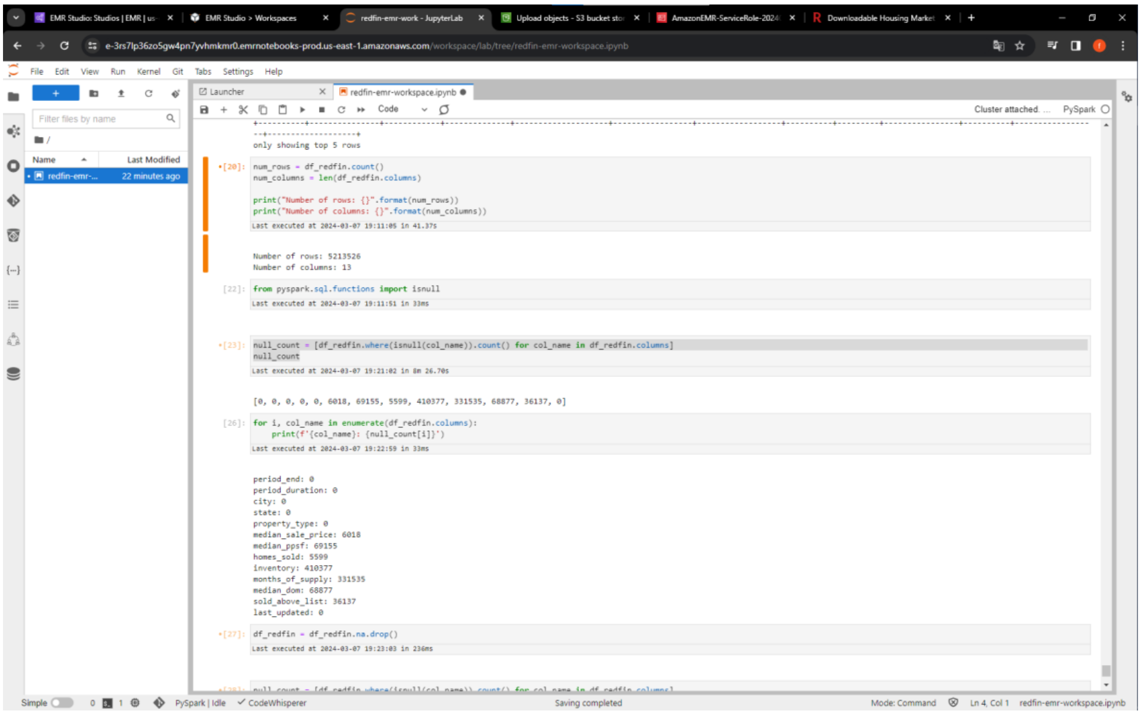
```
['period_begin', 'period_end', 'period_duration', 'region_type', 'region_type_id', 'table_id', 'is_seasonally_adjusted', 'region', 'city', 'state', 'state_code', 'property_type',
'property_type_id', 'median_sale_price', 'median_sale_price_mom', 'median_sale_price_yoy', 'median_list_price', 'median_list_price_mom', 'median_list_price_yoy', 'median_ppsf', 'm
edian_ppsf_mom', 'median_ppsf_yoy', 'median_list_ppsf', 'median_list_ppsf_mom', 'median_list_ppsf_yoy', 'homes_sold', 'homes_sold_mom', 'homes_sold_yoy', 'pending_sales', 'pending
_sales_mom', 'pending_sales_yoy', 'new_listings', 'new_listings_mom', 'new_listings_yoy', 'inventory', 'inventory_mom', 'inventory_yoy', 'months_of_supply', 'months_of_supply_mo
m', 'months_of_supply_yoy', 'median_dom', 'median_dom_mom', 'median_dom_yoy', 'avg_sale_to_list', 'avg_sale_to_list_mom', 'avg_sale_to_list_yoy', 'sold_above_list', 'sold_above_li
st_mom', 'sold_above_list_yoy', 'price_drops', 'price_drops_mom', 'price_drops_yoy', 'off_market_in_two_weeks', 'off_market_in_two_weeks_mom', 'off_market_in_two_weeks_yoy', 'pare
nt_metro_region', 'parent_metro_region_metro_code', 'last_updated']
```

```
[15]: selected_columns = [
          'period_end',
          'period_duration',
          'city',
          'state',
          'property_type',
          'median_sale_price',
          'median_ppsf',
          'homes_sold',
          'inventory',
          'months_of_supply',
          'median_dom',
          'sold_above_list',
          'last_updated'
      ]
Last executed at 2024-03-07 19:08:39 in 31ms
```

```
[16]: df_redfin = redfin_data.select(selected_columns)
Last executed at 2024-03-07 19:09:05 in 239ms
```

```
[18]: df_redfin.show(5)
Last executed at 2024-03-07 19:09:27 in 9.28s
```

```
+----------+---------------+------------+--------------+--------------------+-----------------+-----------------+----------+---------+----------------+----------+---------------
--+-------------------+
|period_end|period_duration|        city|         state|       property_type|median_sale_price|      median_ppsf|homes_sold|inventory|months_of_supply|median_dom|   sold_above_li
st|       last_updated|
+----------+---------------+------------+--------------+--------------------+-----------------+-----------------+----------+---------+----------------+----------+---------------
--+-------------------+
|2022-01-31|             30|    Sterling|      Virginia|     All Residential|         532000.0| 268.947373717208|        28|       21|             0.8|        13|0.78571428571428
57|2024-02-11 14:26:11|
|2013-03-31|             30|   Swansboro|North Carolina|Single Family Res...|         146250.0| 93.86333914559721|         4|       76|            19.0|       390|
```


EMR Studio: Studios | EMR | us-
EMR Studio > Workspaces
redfin-emr-work - JupyterLab
Upload objects - S3 bucket stor
AmazonEMR-ServiceRole-2024
Downloadable Housing Market
e-3rs7lp36zo5gw4pn7yvhmkmr0.emrnotebooks-prod.us-east-1.amazonaws.com/workspace/lab/tree/redfin-emr-workspace.ipynb
File Edit View Run Kernel Git Tabs Settings Help
Launcher
redfin-emr-workspace.ipynb
Code
Cluster attached. ...
PySpark
Filter files by name
Name
Last Modified
redfin-emr-...
22 minutes ago
Simple
PySpark | Idle
CodeWhisperer
Saving completed
Mode: Command
Ln 4, Col 1
redfin-emr-workspace.ipynb

```
--+------------------+
only showing top 5 rows
```

```
[20]: num_rows = df_redfin.count()
      num_columns = len(df_redfin.columns)

      print("Number of rows: {}".format(num_rows))
      print("Number of columns: {}".format(num_columns))
      Last executed at 2024-03-07 19:11:05 in 41.37s

      Number of rows: 5213526
      Number of columns: 13
```

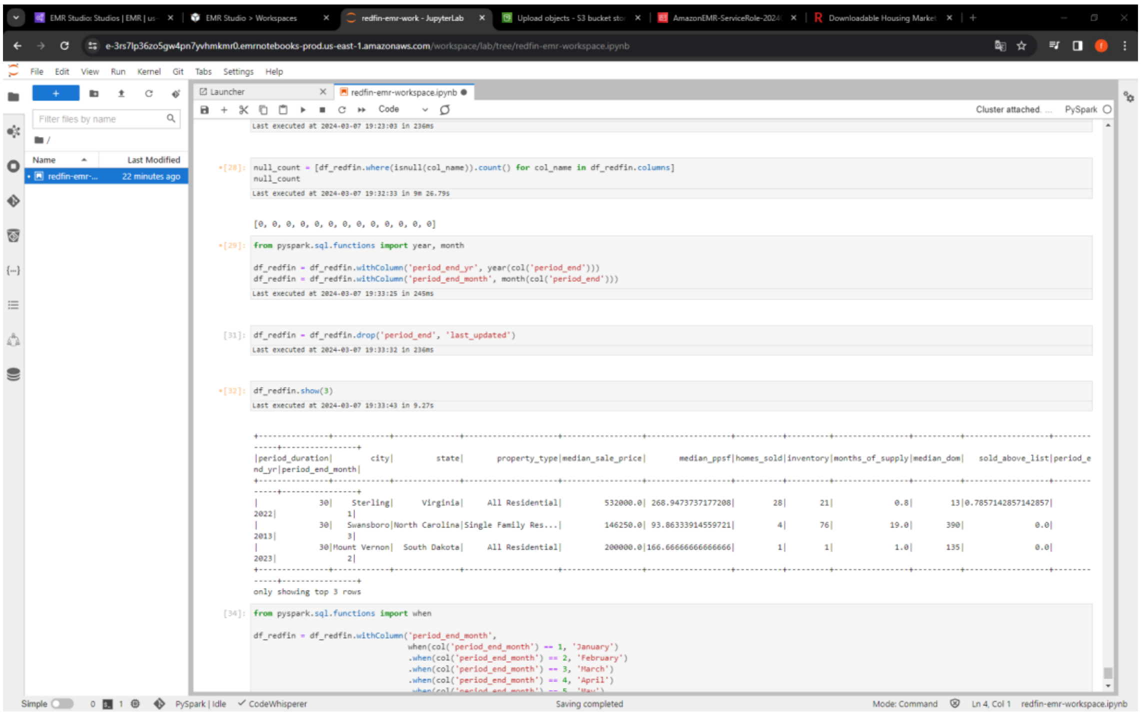
```
[22]: from pyspark.sql.functions import isnull
      Last executed at 2024-03-07 19:11:51 in 33ms
```

```
[23]: null_count = [df_redfin.where(isnull(col_name)).count() for col_name in df_redfin.columns]
      null_count
      Last executed at 2024-03-07 19:21:02 in 8m 26.70s

      [0, 0, 0, 0, 0, 6018, 69155, 5599, 410377, 331535, 68877, 36137, 0]
```

```
[26]: for i, col_name in enumerate(df_redfin.columns):
          print(f'{col_name}: {null_count[i]}')
      Last executed at 2024-03-07 19:22:59 in 33ms

      period_end: 0
      period_duration: 0
      city: 0
      state: 0
      property_type: 0
      median_sale_price: 6018
      median_ppsf: 69155
      homes_sold: 5599
      inventory: 410377
      months_of_supply: 331535
      median_dom: 68877
      sold_above_list: 36137
      last_updated: 0
```

```
[27]: df_redfin = df_redfin.na.drop()
      Last executed at 2024-03-07 19:23:03 in 236ms
```

```
[28]: null_count = [df_redfin.where(isnull(col_name)).count() for col_name in df_redfin.columns]
```


EMR Studio: Studios | EMR | us-
EMR Studio > Workspaces
redfin-emr-work - JupyterLab
Upload objects - S3 bucket stor
AmazonEMR-ServiceRole-2024
Downloadable Housing Market
e-3rs7lp36zo5gw4pn7yvhmkmr0.emrnotebooks-prod.us-east-1.amazonaws.com/workspace/lab/tree/redfin-emr-workspace.ipynb
File Edit View Run Kernel Git Tabs Settings Help
Launcher
redfin-emr-workspace.ipynb
Code
Cluster attached. ...
PySpark
Filter files by name
Name
Last Modified
redfin-emr-...
22 minutes ago
Simple
PySpark | Idle
CodeWhisperer
Saving completed
Mode: Command
Ln 4, Col 1
redfin-emr-workspace.ipynb

```
      Last executed at 2024-03-07 19:23:03 in 236ms
```

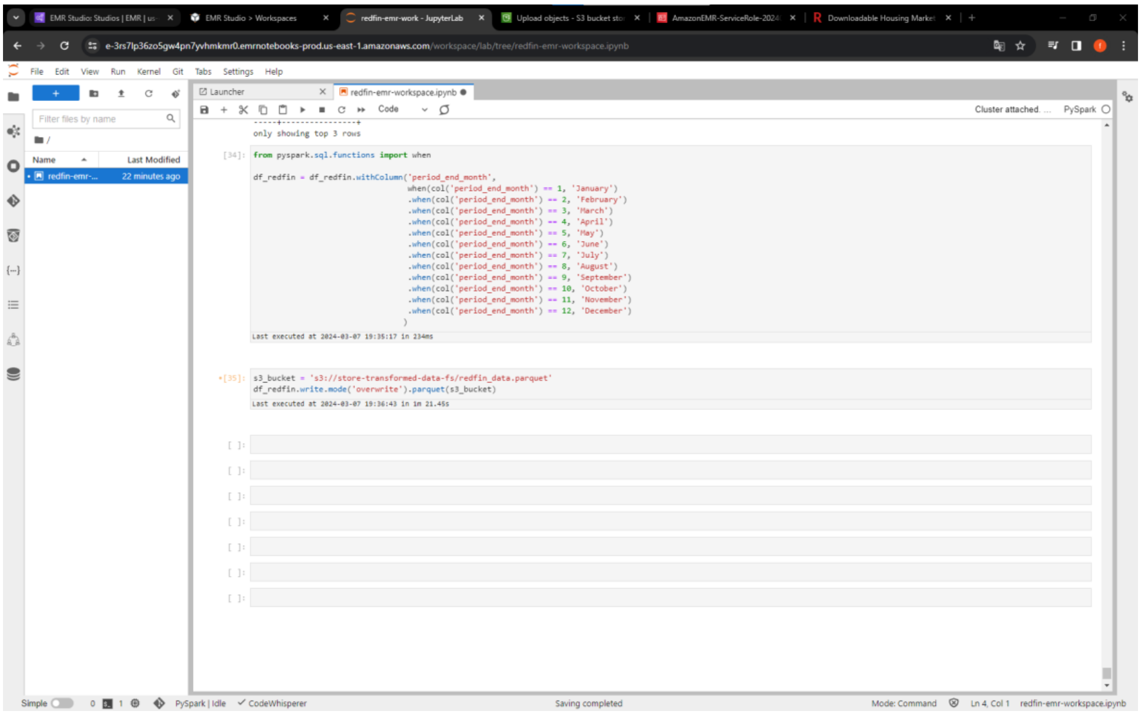
```
[28]: null_count = [df_redfin.where(isnull(col_name)).count() for col_name in df_redfin.columns]
      null_count
      Last executed at 2024-03-07 19:32:33 in 9m 26.79s

      [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
```

```
[29]: from pyspark.sql.functions import year, month

      df_redfin = df_redfin.withColumn('period_end_yr', year(col('period_end')))
      df_redfin = df_redfin.withColumn('period_end_month', month(col('period_end')))
      Last executed at 2024-03-07 19:33:25 in 245ms
```

```
[31]: df_redfin = df_redfin.drop('period_end', 'last_updated')
      Last executed at 2024-03-07 19:33:32 in 236ms
```

```
[32]: df_redfin.show(3)
      Last executed at 2024-03-07 19:33:43 in 9.27s

+---------------+------------+--------------+--------------------+-----------------+------------------+----------+---------+----------------+----------+------------------+--------
-----+----------------+
|period_duration|        city|         state|       property_type|median_sale_price|       median_ppsf|homes_sold|inventory|months_of_supply|median_dom|   sold_above_list|period_e
nd_yr|period_end_month|
+---------------+------------+--------------+--------------------+-----------------+------------------+----------+---------+----------------+----------+------------------+--------
-----+----------------+
|             30|    Sterling|      Virginia|     All Residential|         532000.0| 268.9473737177208|        28|       21|             0.8|        13|0.7857142857142857|
2022|               1|
|             30|   Swansboro|North Carolina|Single Family Res...|         146250.0| 93.86333914559721|         4|       76|            19.0|       390|               0.0|
2013|               3|
|             30|Mount Vernon|  South Dakota|     All Residential|         200000.0|166.66666666666666|         1|        1|             1.0|       135|               0.0|
2023|               2|
+---------------+------------+--------------+--------------------+-----------------+------------------+----------+---------+----------------+----------+------------------+--------
-----+----------------+
only showing top 3 rows
```

```
[34]: from pyspark.sql.functions import when

      df_redfin = df_redfin.withColumn('period_end_month',
                                       when(col('period_end_month') == 1, 'January')
                                       .when(col('period_end_month') == 2, 'February')
                                       .when(col('period_end_month') == 3, 'March')
                                       .when(col('period_end_month') == 4, 'April')
                                       .when(col('period_end_month') == 5, 'May')
```

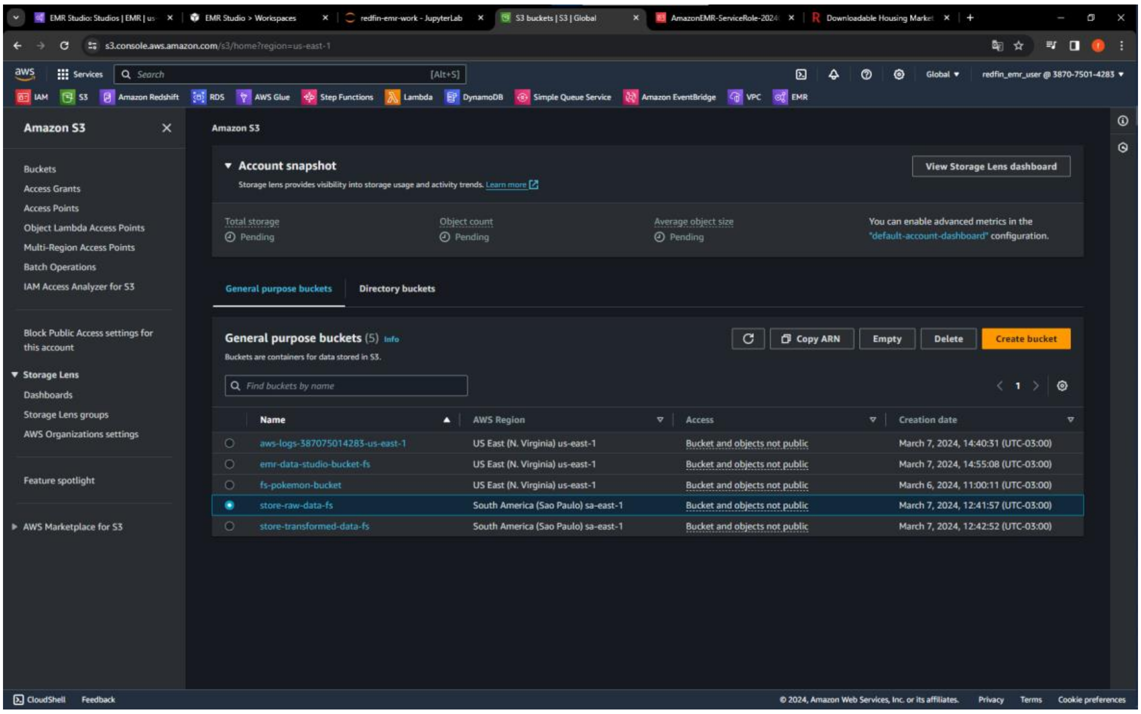


```
only showing top 3 rows
```

```
[34]: from pyspark.sql.functions import when

df_redfin = df_redfin.withColumn('period_end_month',
                                 when(col('period_end_month') == 1, 'January')
                                 .when(col('period_end_month') == 2, 'February')
                                 .when(col('period_end_month') == 3, 'March')
                                 .when(col('period_end_month') == 4, 'April')
                                 .when(col('period_end_month') == 5, 'May')
                                 .when(col('period_end_month') == 6, 'June')
                                 .when(col('period_end_month') == 7, 'July')
                                 .when(col('period_end_month') == 8, 'August')
                                 .when(col('period_end_month') == 9, 'September')
                                 .when(col('period_end_month') == 10, 'October')
                                 .when(col('period_end_month') == 11, 'November')
                                 .when(col('period_end_month') == 12, 'December')
                                )
Last executed at 2024-03-07 19:35:17 in 234ms
```

```
[35]: s3_bucket = 's3://store-transformed-data-fs/redfin_data.parquet'
df_redfin.write.mode('overwrite').parquet(s3_bucket)
Last executed at 2024-03-07 19:36:43 in 1m 21.45s
```

e-3rs7lp36zo5gw4pn7yvhmkmr0.emrnotebooks-prod.us-east-1.amazonaws.com/workspace/lab/tree/redfin-emr-workspace.ipynb
File Edit View Run Kernel Git Tabs Settings Help
Launcher
redfin-emr-workspace.ipynb
Code
Cluster attached. PySpark
Name Last Modified
redfin-emr-... 22 minutes ago
Simple 0 1 PySpark | Idle CodeWhisperer
Saving completed
Mode: Command Ln 4, Col 1 redfin-emr-workspace.ipynb



Amazon S3

Account snapshot

Storage lens provides visibility into storage usage and activity trends. Learn more

Total storage: Pending
Object count: Pending
Average object size: Pending

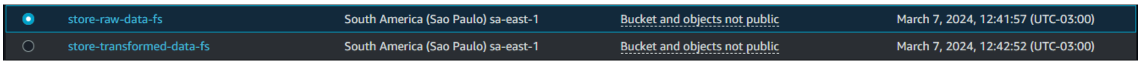
You can enable advanced metrics in the "default-account-dashboard" configuration.

General purpose buckets (5) Info

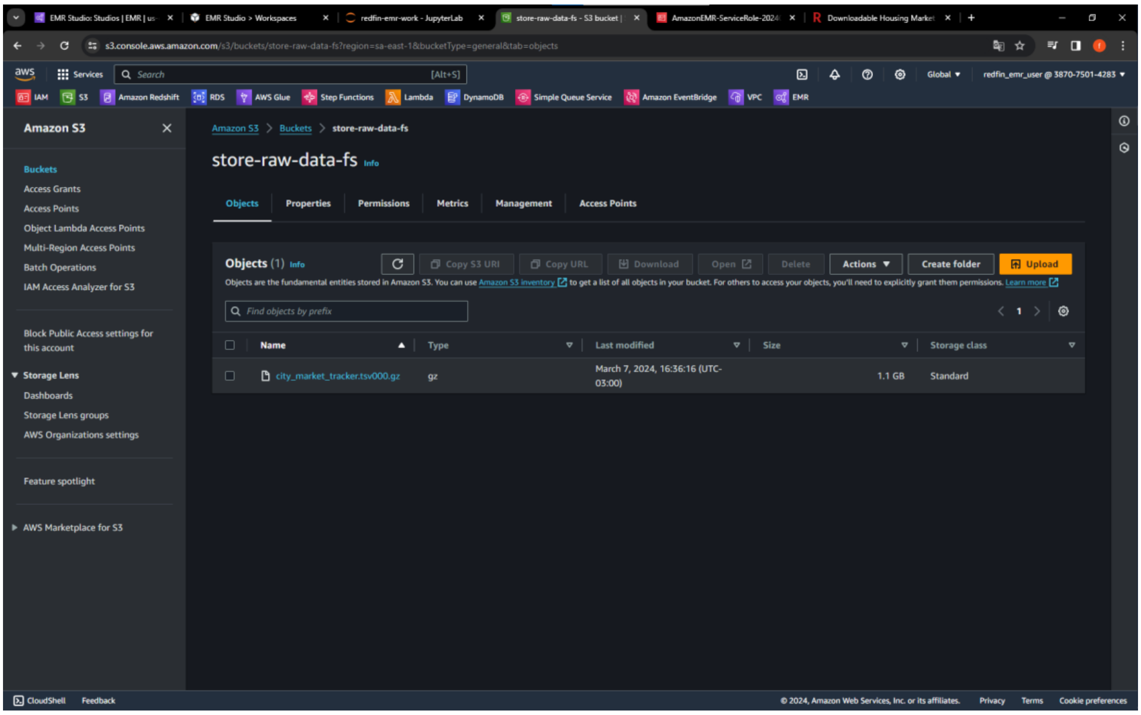
Buckets are containers for data stored in S3.

	Name	AWS Region	Access	Creation date
○	aws-logs-387075014283-us-east-1	US East (N. Virginia) us-east-1	Bucket and objects not public	March 7, 2024, 14:40:31 (UTC-03:00)
○	emr-data-studio-bucket-fs	US East (N. Virginia) us-east-1	Bucket and objects not public	March 7, 2024, 14:55:08 (UTC-03:00)
○	fs-pokemon-bucket	US East (N. Virginia) us-east-1	Bucket and objects not public	March 6, 2024, 11:00:11 (UTC-03:00)
●	store-raw-data-fs	South America (Sao Paulo) sa-east-1	Bucket and objects not public	March 7, 2024, 12:41:57 (UTC-03:00)
○	store-transformed-data-fs	South America (Sao Paulo) sa-east-1	Bucket and objects not public	March 7, 2024, 12:42:52 (UTC-03:00)

s3.console.aws.amazon.com/s3/home?region=us-east-1
Services
Search [Alt+S]
Global
redfin_emr_user @ 3870-7501-4283
IAM S3 Amazon Redshift RDS AWS Glue Step Functions Lambda DynamoDB Simple Queue Service Amazon EventBridge VPC EMR
Amazon S3
Buckets
Access Grants
Access Points
Object Lambda Access Points
Multi-Region Access Points
Batch Operations
IAM Access Analyzer for S3
Block Public Access settings for this account
Storage Lens
Dashboards
Storage Lens groups
AWS Organizations settings
Feature spotlight
AWS Marketplace for S3
View Storage Lens dashboard
General purpose buckets
Directory buckets
Copy ARN
Empty
Delete
Create bucket
Find buckets by name
CloudShell Feedback
© 2024, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences



●	store-raw-data-fs	South America (Sao Paulo) sa-east-1	Bucket and objects not public	March 7, 2024, 12:41:57 (UTC-03:00)
○	store-transformed-data-fs	South America (Sao Paulo) sa-east-1	Bucket and objects not public	March 7, 2024, 12:42:52 (UTC-03:00)



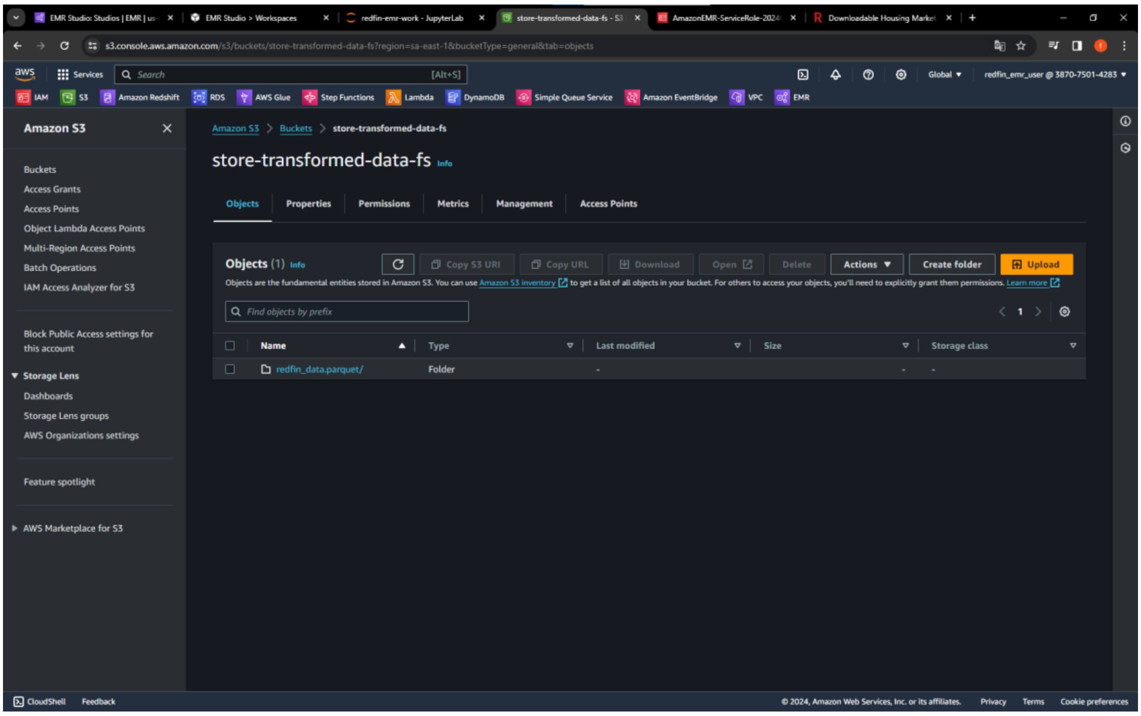
Amazon S3 > Buckets > store-raw-data-fs

store-raw-data-fs Info

Objects (1) 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
□	city_market_tracker.tsv000.gz	gz	March 7, 2024, 16:36:16 (UTC-03:00)	1.1 GB	Standard


EMR Studio: Studios | EMR | us-
EMR Studio > Workspaces
redfin-emr-work - JupyterLab
store-transformed-data-fs - S3
AmazonEMR-ServiceRole-2024
Downloadable Housing Market
s3.console.aws.amazon.com/s3/buckets/store-transformed-data-fs?region=sa-east-1&bucketType=general&tab=objects
aws
Services
Search
[Alt+S]
Global
redfin_emr_user @ 3870-7501-4283
IAM
S3
Amazon Redshift
RDS
AWS Glue
Step Functions
Lambda
DynamoDB
Simple Queue Service
Amazon EventBridge
VPC
EMR
Amazon S3
Buckets
Access Grants
Access Points
Object Lambda Access Points
Multi-Region Access Points
Batch Operations
IAM Access Analyzer for S3
Block Public Access settings for this account
Storage Lens
Dashboards
Storage Lens groups
AWS Organizations settings
Feature spotlight
AWS Marketplace for S3
Objects
Properties
Permissions
Metrics
Management
Access Points
Copy S3 URI
Copy URL
Download
Open
Delete
Actions
Create folder
Upload
Find objects by prefix
1
CloudShell
Feedback
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

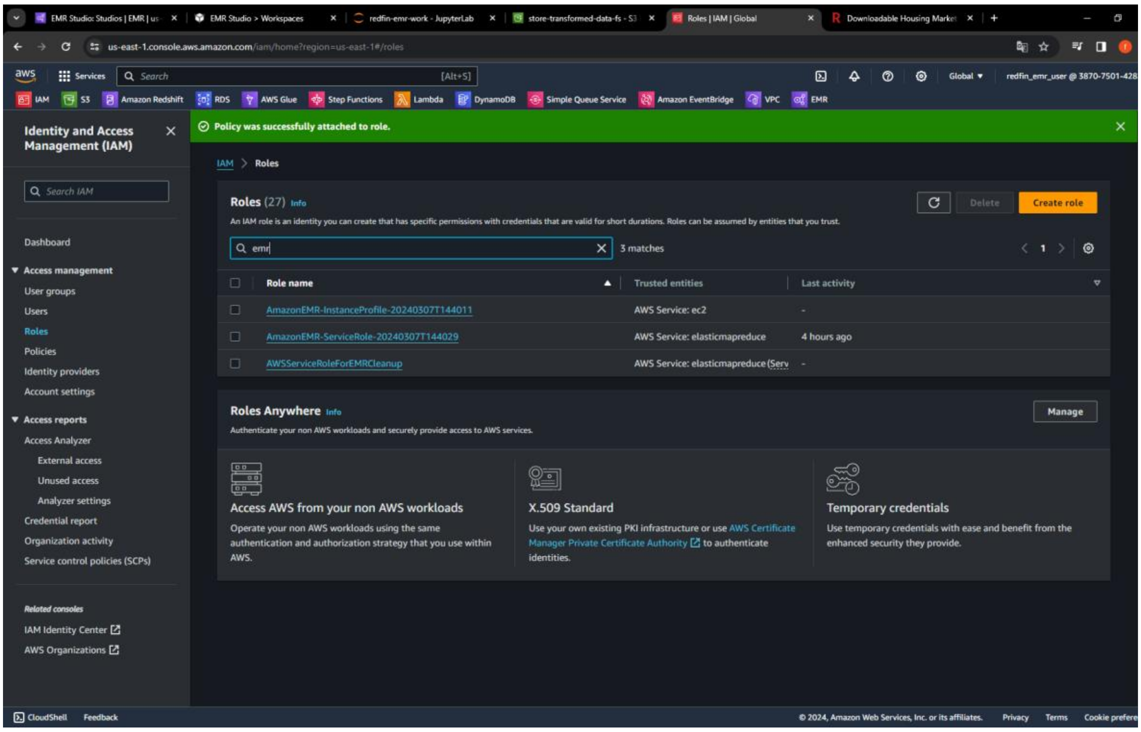
Amazon S3 > Buckets > store-transformed-data-fs

store-transformed-data-fs Info

Objects (1) 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
□	redfin_data.parquet/	Folder	-	-	-



Policy was successfully attached to role.

IAM > Roles

Roles (27) Info

An IAM role is an identity you can create that has specific permissions with credentials that are valid for short durations. Roles can be assumed by entities that you trust.

emr

3 matches

□	Role name	Trusted entities	Last activity
□	AmazonEMR-InstanceProfile-20240307T144011	AWS Service: ec2	-
□	AmazonEMR-ServiceRole-20240307T144029	AWS Service: elasticmapreduce	4 hours ago
□	AWSServiceRoleForEMRCleanup	AWS Service: elasticmapreduce (Serv	-

Roles Anywhere Info

Authenticate your non AWS workloads and securely provide access to AWS services.

Access AWS from your non AWS workloads

Operate your non AWS workloads using the same authentication and authorization strategy that you use within AWS.

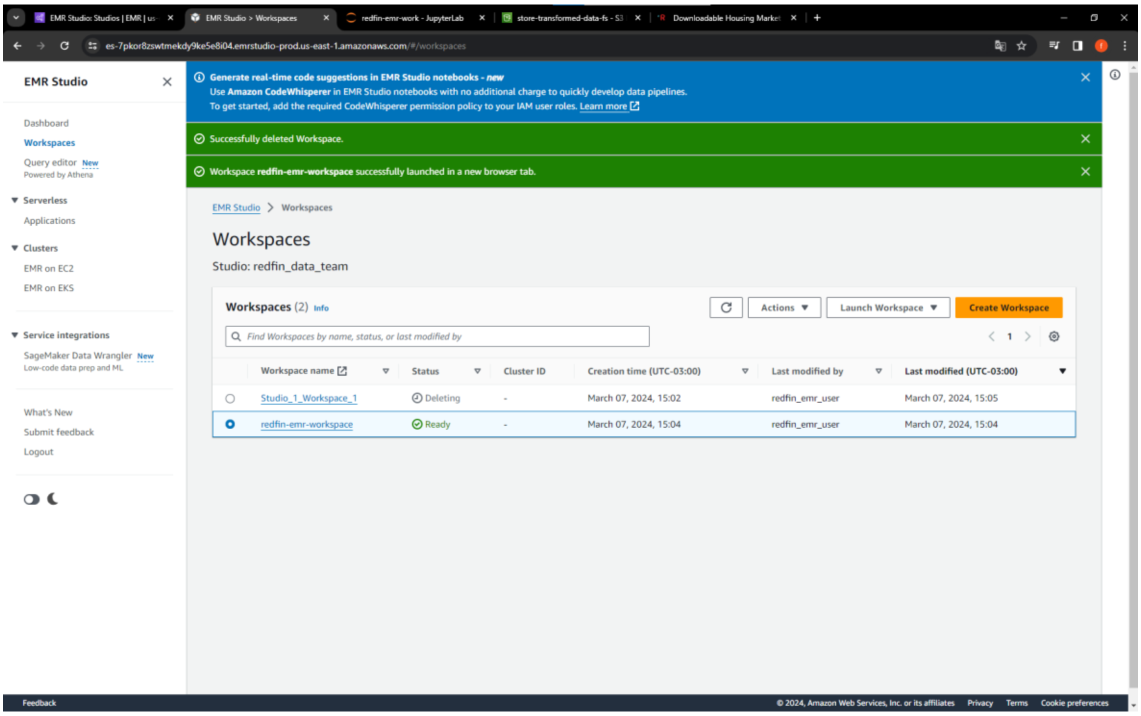
X.509 Standard

Use your own existing PKI infrastructure or use AWS Certificate Manager Private Certificate Authority to authenticate identities.

Temporary credentials

Use temporary credentials with ease and benefit from the enhanced security they provide.

© 2024, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie prefere



EMR Studio: Studios | EMR | us
EMR Studio > Workspaces
redfin-emr-work - JupyterLab
store-transformed-data-fs - S3
Downloadable Housing Market
es-7pkor8zswtmekdy9ke5e8i04.emrstudio-prod.us-east-1.amazonaws.com/#/workspaces
EMR Studio
Dashboard
Workspaces
Query editor New
Powered by Athena
Serverless
Applications
Clusters
EMR on EC2
EMR on EKS
Service integrations
SageMaker Data Wrangler New
Low-code data prep and ML
What's New
Submit feedback
Logout
Feedback

Generate real-time code suggestions in EMR Studio notebooks - *new*
Use **Amazon CodeWhisperer** in EMR Studio notebooks with no additional charge to quickly develop data pipelines.
To get started, add the required CodeWhisperer permission policy to your IAM user roles. Learn more

Successfully deleted Workspace.

Workspace **redfin-emr-workspace** successfully launched in a new browser tab.

EMR Studio > Workspaces

Workspaces

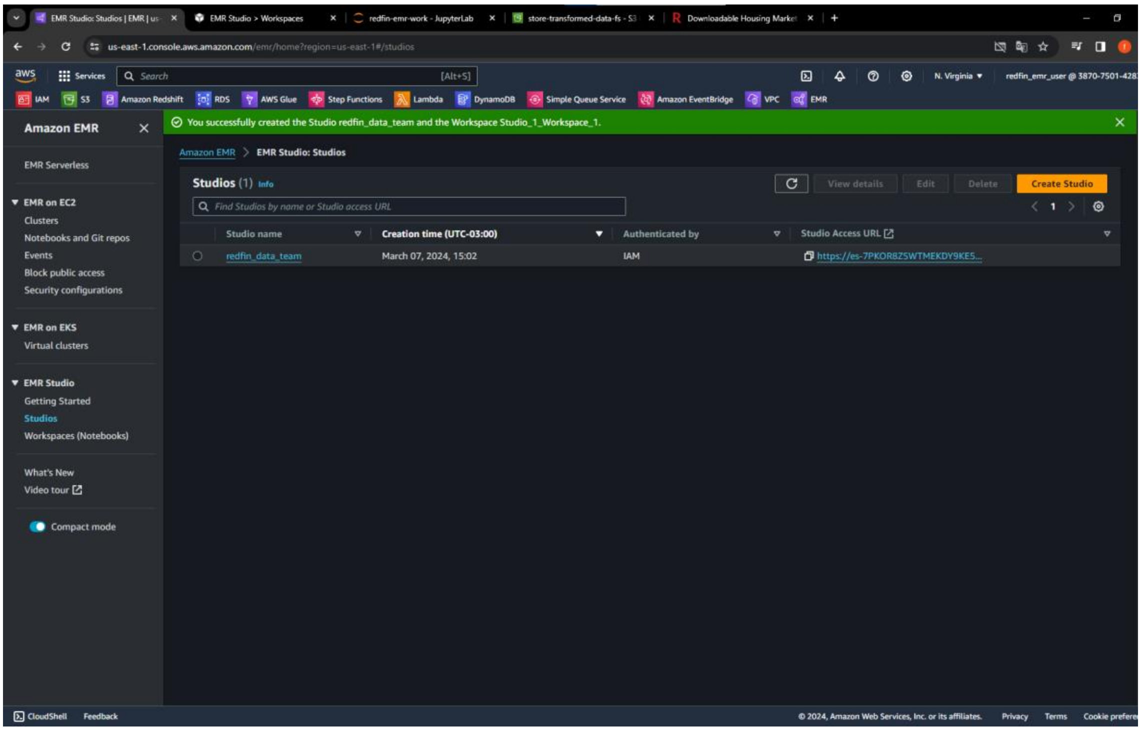
Studio: redfin_data_team

Workspaces (2) Info

Find Workspaces by name, status, or last modified by

	Workspace name	Status	Cluster ID	Creation time (UTC-03:00)	Last modified by	Last modified (UTC-03:00)
○	Studio_1_Workspace_1	Deleting	-	March 07, 2024, 15:02	redfin_emr_user	March 07, 2024, 15:05
●	redfin-emr-workspace	Ready	-	March 07, 2024, 15:04	redfin_emr_user	March 07, 2024, 15:04

© 2024, Amazon Web Services, Inc. or its affiliates Privacy Terms Cookie preferences



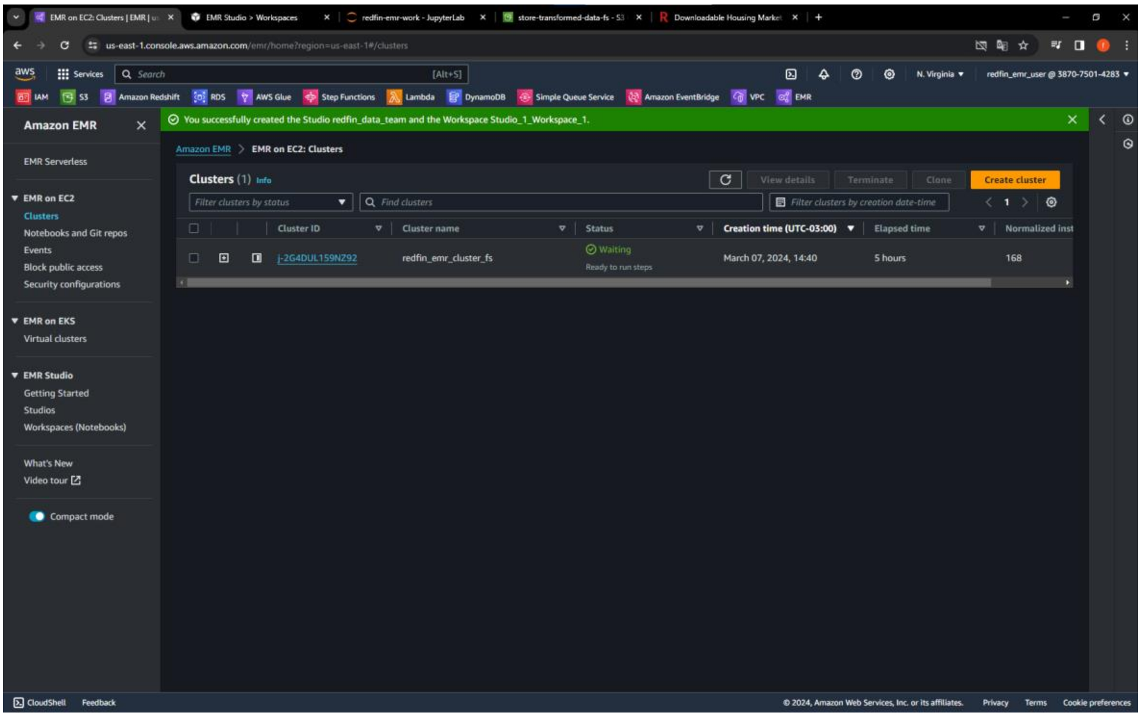
us-east-1.console.aws.amazon.com/emr/home?region=us-east-1#/studios

You successfully created the Studio redfin_data_team and the Workspace Studio_1_Workspace_1.

Amazon EMR > EMR Studio: Studios

Studios (1)

	Studio name	Creation time (UTC-03:00)	Authenticated by	Studio Access URL
○	redfin_data_team	March 07, 2024, 15:02	IAM	https://es-7PKOR8ZSWTMEKDY9KE5...



us-east-1.console.aws.amazon.com/emr/home?region=us-east-1#/clusters

You successfully created the Studio redfin_data_team and the Workspace Studio_1_Workspace_1.

Amazon EMR > EMR on EC2: Clusters

Clusters (1)

□			Cluster ID	Cluster name	Status	Creation time (UTC-03:00)	Elapsed time	Normalized inst
□	⊞	◧	j-2G4DUL159NZ92	redfin_emr_cluster_fs	✓ Waiting Ready to run steps	March 07, 2024, 14:40	5 hours	168