Kadirbek Sharau

Detailed Report on Collaborative Filtering with PySpark on Google Cloud Dataproc

Objective

The goal of this assignment is to perform collaborative filtering using the Alternating Least Squares (ALS) algorithm with PySpark on Google Cloud Dataproc. The tasks involve preparing data, uploading it to Google Cloud Storage, creating and uploading a PySpark script, and submitting a PySpark job to a Dataproc cluster.

Step-by-Step Instructions

Step 1: Prepare and Transform Data

Description: Transform the `u.data` file to the required format (UserID, MovieID, rating) using a shell script and upload it to your Cloud Storage bucket.

Code:

1. **Create the `u.data` File:**
   - **Create a file named `u.data` and populate it with your data.**

Transform Data Using Shell Script:

```
# Create transform_data.sh
echo '#!/bin/bash
cat u.data | while read userid movieid rating timestamp
do
    echo "${userid},${movieid},${rating}"
done > u_data_transformed.csv' > transform_data.sh

# Make the script executable
chmod +x transform_data.sh

# Run the script
./transform_data.sh
```

2.

Explanation: The shell script reads the `u.data` file, trims extra spaces, extracts the first three fields (UserID, MovieID, rating), and replaces spaces with commas. The transformed data is saved in `u_data_transformed.csv`.

Step 2: Upload Data to Cloud Storage Bucket

Description: Upload the transformed data file `u_data_transformed.csv` to your Cloud Storage bucket.

Code:

```
# Upload the transformed data to Cloud Storage
gsutil cp u_data_transformed.csv gs://big_data_ml_recommendation_sys/
```

Explanation: The `gsutil cp` command copies the `u_data_transformed.csv` file from your local machine to your specified Cloud Storage bucket.

Step 3: Create and Upload the PySpark Script

Description: Create a PySpark script to perform collaborative filtering using MLlib and upload it to your Cloud Storage bucket.

Code:

1. **Create the PySpark Script:**
   - **Create a file named `recommendation_example.py` with the following content:**

```
from pyspark import SparkContext
from pyspark.mllib.recommendation import ALS,
MatrixFactorizationModel, Rating

if __name__ == "__main__":
    sc = SparkContext(appName="PythonCollaborativeFilteringExample")
    data =
sc.textFile("gs://big_data_ml_recommendation_sys/u_data_transformed.cs
v")
    ratings = data.map(lambda l: l.split(','))\
                  .map(lambda l: Rating(int(l[0]), int(l[1]),
float(l[2])))
```

```
    rank = 10
    numIterations = 10
    model = ALS.train(ratings, rank, numIterations)

    testdata = ratings.map(lambda p: (p[0], p[1]))
    predictions = model.predictAll(testdata).map(lambda r: ((r[0],
r[1]), r[2]))
    ratesAndPreds = ratings.map(lambda r: ((r[0], r[1]),
r[2])).join(predictions)
    MSE = ratesAndPreds.map(lambda r: (r[1][0] - r[1][1])**2).mean()
    print("Mean Squared Error = " + str(MSE))

    model.save(sc,
"gs://big_data_ml_recommendation_sys/myCollaborativeFilter")
    sameModel = MatrixFactorizationModel.load(sc,
"gs://big_data_ml_recommendation_sys/myCollaborativeFilter")
```

2.

Upload the PySpark Script:

```
gsutil cp recommendation_example.py
gs://big_data_ml_recommendation_sys/
```

3.

Explanation: The PySpark script loads the transformed data from Cloud Storage, trains a collaborative filtering model using ALS, evaluates the model by calculating the mean squared error, and saves the model back to Cloud Storage. The script is then uploaded to the Cloud Storage bucket.

Step 4: Submit the PySpark Job to Dataproc

Description: Submit the PySpark job to your Dataproc cluster to execute the collaborative filtering task.

Code:

```
gcloud dataproc jobs submit pyspark \
    gs://big_data_ml_recommendation_sys/recommendation_example.py \
    --cluster=spark-cluster \
    --region=us-west1
```

Explanation: The `gcloud dataproc jobs submit pyspark` command submits the PySpark script stored in Cloud Storage to the Dataproc cluster named `spark-cluster` located in the `us-west1` region for execution.

Troubleshooting

If you encounter an error indicating that no Dataproc cluster exists, follow these steps:

Create a Dataproc Cluster:

```
gcloud dataproc clusters create spark-cluster \
    --region us-west1 \
    --zone us-west1-a \
    --single-node
```

1.

Submit the PySpark Job:

```
gcloud dataproc jobs submit pyspark \
    gs://big_data_ml_recommendation_sys/recommendation_example.py \
    --cluster=spark-cluster \
    --region=us-west1
```

2.

Explanation: First, create a Dataproc cluster named `spark-cluster` in the `us-west1` region. Then, submit the PySpark job to the newly created cluster. Ensure to replace `spark-cluster` with the actual name of your Dataproc cluster if you choose a different name. The cluster creation step might take a few minutes. Once it's running, you can then submit your job.

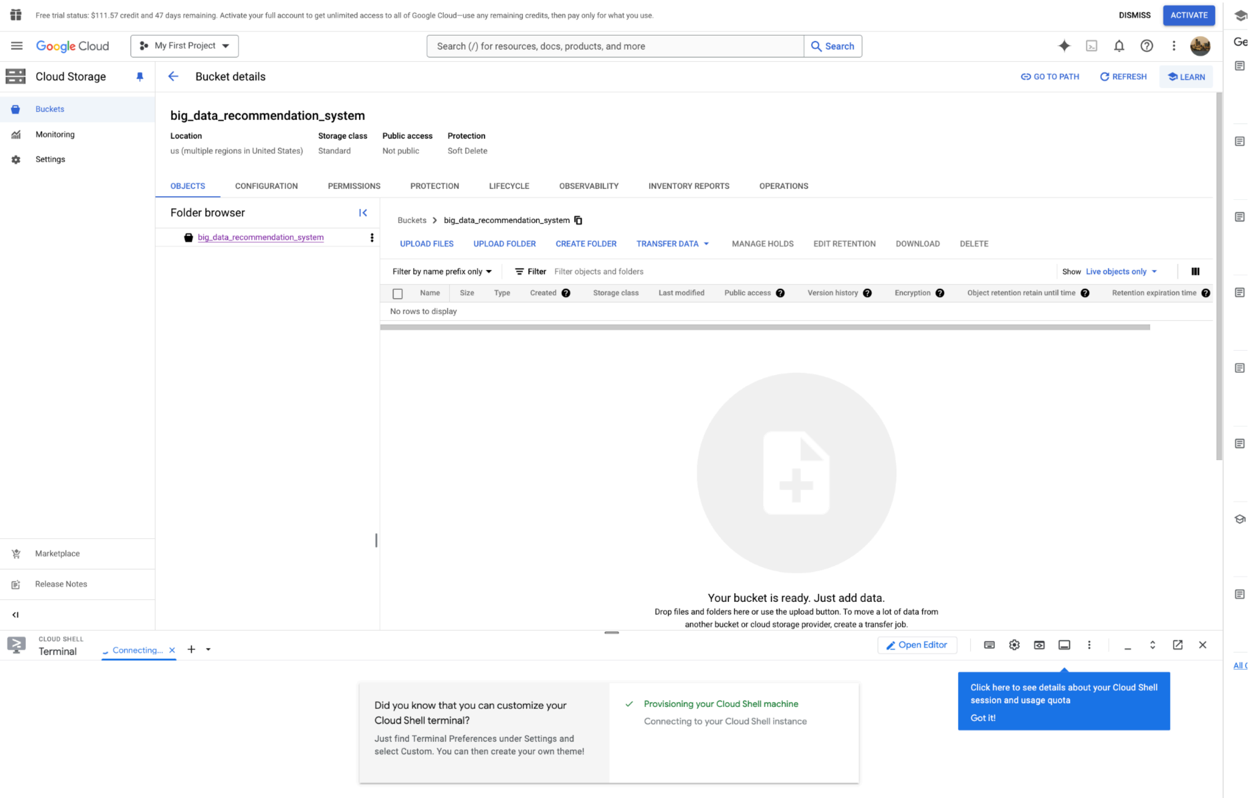
Result

After submitting the job, the output will indicate the Mean Squared Error (MSE) of the model:

```
24/08/01 20:45:41 INFO org.apache.hadoop.mapred.FileInputFormat: Total
input files to process : 1
Mean Squared Error = 0.48149423210378404
24/08/01 20:46:21 INFO
com.google.cloud.hadoop.repackaged.gcs.com.google.cloud.hadoop.gcsio.G
oogleCloudStorageFileSystem: Successfully repaired
```

```
'gs://big_data_ml_recommendation_sys/myCollaborativeFilter/metadata/'
directory.
24/08/01 20:46:35 INFO
com.google.cloud.hadoop.repackaged.gcs.com.google.cloud.hadoop.gcsio.G
oogleCloudStorageFileSystem: Successfully repaired
'gs://big_data_ml_recommendation_sys/myCollaborativeFilter/data/user/'
directory.
24/08/01 20:46:35 INFO
com.google.cloud.hadoop.repackaged.gcs.com.google.cloud.hadoop.gcsio.G
oogleCloudStorageFileSystem: Successfully repaired
'gs://big_data_ml_recommendation_sys/myCollaborativeFilter/data/produc
t/' directory.
24/08/01 20:46:36 INFO org.apache.hadoop.mapred.FileInputFormat: Total
input files to process : 1
24/08/01 20:46:36 WARN
org.apache.spark.mllib.recommendation.MatrixFactorizationModel: User
factor does not have a partitioner. Prediction on individual records
could be slow.
24/08/01 20:46:36 WARN
org.apache.spark.mllib.recommendation.MatrixFactorizationModel:
Product factor is not cached. Prediction could be slow.
24/08/01 20:46:42 WARN
org.apache.spark.mllib.recommendation.MatrixFactorizationModel: User
factor is not cached. Prediction on individual records could be slow.
24/08/01 20:46:42 WARN
org.apache.spark.mllib.recommendation.MatrixFactorizationModel:
Product factor is not cached. Prediction could be slow.
24/08/01 20:46:42 INFO org.apache.spark.SparkContext: Stopped
Spark3b5e47ed
```

By following these steps, you will be able to successfully complete your assignment using your Dataproc cluster and Cloud Storage bucket on GCP.



Free trial status: $111.57 credit and 47 days remaining. Activate your full account to get unlimited access to all of Google Cloud—use any remaining credits, then pay only for what you use.

big_data_recommendation_system

Location	Storage class	Public access	Protection
us (multiple regions in United States)	Standard	Not public	Soft Delete

□	Name	Size	Type	Created	Storage class	Last modified	Public access	Version history	Encryption	Object retention retain until time	Retention expiration time

No rows to display

Your bucket is ready. Just add data.

Drop files and folders here or use the upload button. To move a lot of data from another bucket or cloud storage provider, create a transfer job.

Did you know that you can customize your Cloud Shell terminal?

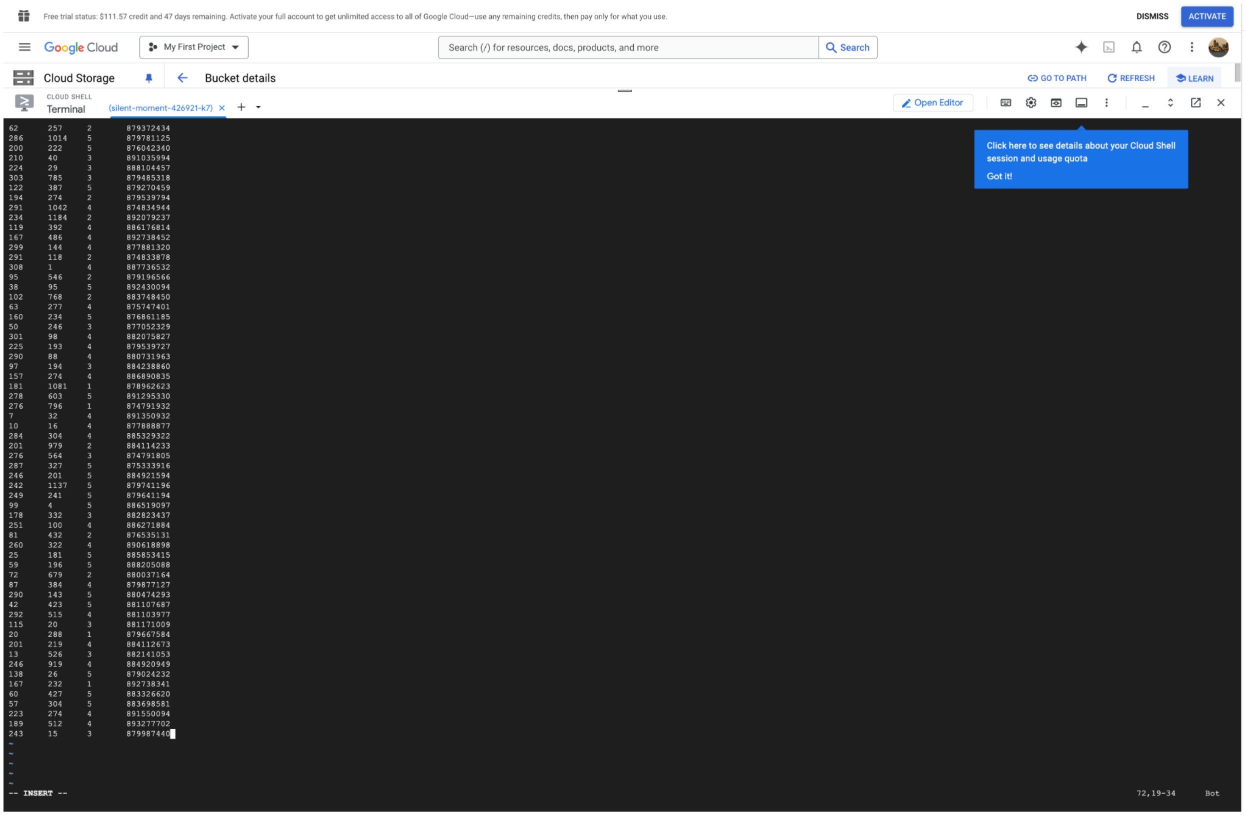
Just find Terminal Preferences under Settings and select Custom. You can then create your own theme!

✓ Provisioning your Cloud Shell machine

Connecting to your Cloud Shell instance

Click here to see details about your Cloud Shell session and usage quota

Got it!

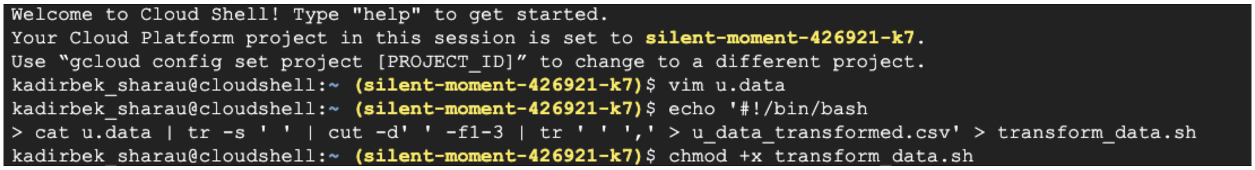


Free trial status: $111.57 credit and 47 days remaining. Activate your full account to get unlimited access to all of Google Cloud—use any remaining credits, then pay only for what you use.

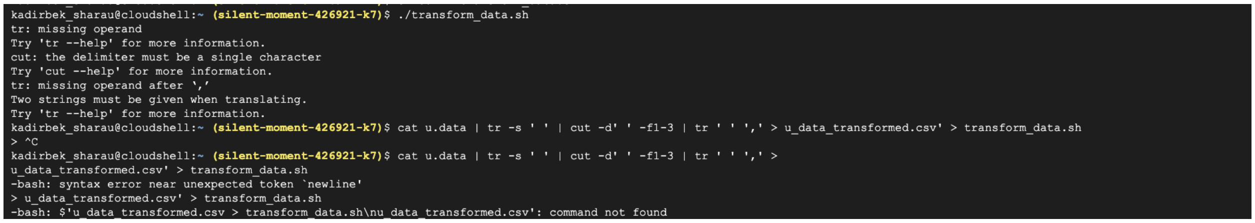
Click here to see details about your Cloud Shell session and usage quota

Got it!

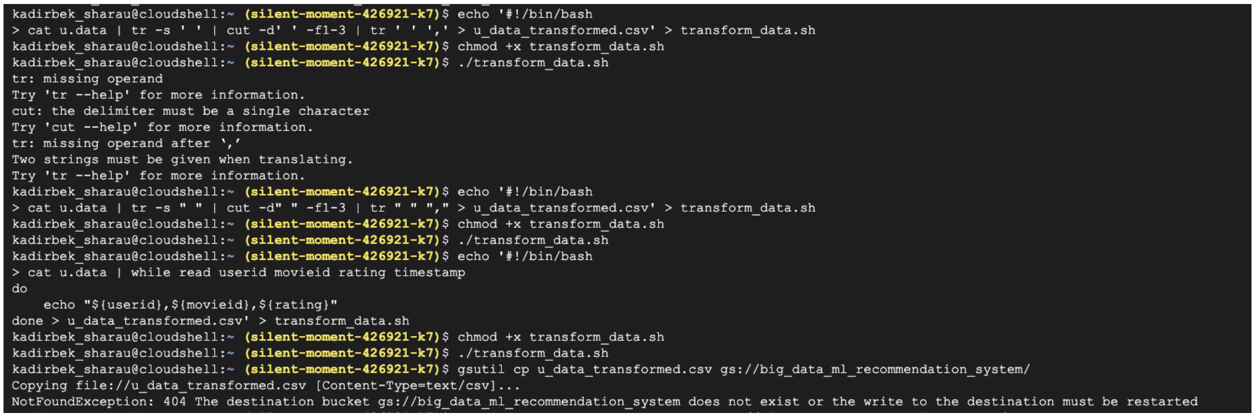
```
62      257     2       879372434
286     1014    5       879781125
200     222     5       876042340
210     40      3       891035994
224     29      3       888104457
303     785     3       879485318
122     387     5       879270459
194     274     2       879539794
291     1042    4       874834944
234     1184    2       892079237
119     392     4       886176814
167     486     4       892738452
299     144     4       877881320
291     118     2       874833878
308     1       4       887736532
95      546     2       879196566
38      95      5       892430094
102     768     2       883748450
63      277     4       875747401
160     234     5       876861185
50      246     3       877052329
301     98      4       882075827
225     193     4       879539727
290     88      4       880731963
97      194     3       884238860
157     274     4       886890835
181     1081    1       878962623
278     603     5       891295330
276     796     1       874791932
7       32      4       891350932
10      16      4       877888877
284     304     4       885329322
201     979     2       884114233
276     564     3       874791805
287     327     5       875333916
246     201     5       884921594
242     1137    5       879741196
249     241     5       879641194
99      4       5       886519097
178     332     3       882823437
251     100     4       886271884
81      432     2       876535131
260     322     4       890618898
25      181     5       885853415
59      196     5       888205088
72      679     2       880037164
87      384     4       879877127
290     143     5       880474293
42      423     5       881107687
292     515     4       881103977
115     20      3       881171009
20      288     1       879667584
201     219     4       884112673
13      526     3       882141053
246     919     4       884920949
138     26      5       879024232
167     232     1       892738341
60      427     5       883326620
57      304     5       883698581
223     274     4       891550094
189     512     4       893277702
243     15      3       879987440
~
~
~
~
~
-- INSERT --                                                72,19-34      Bot
```



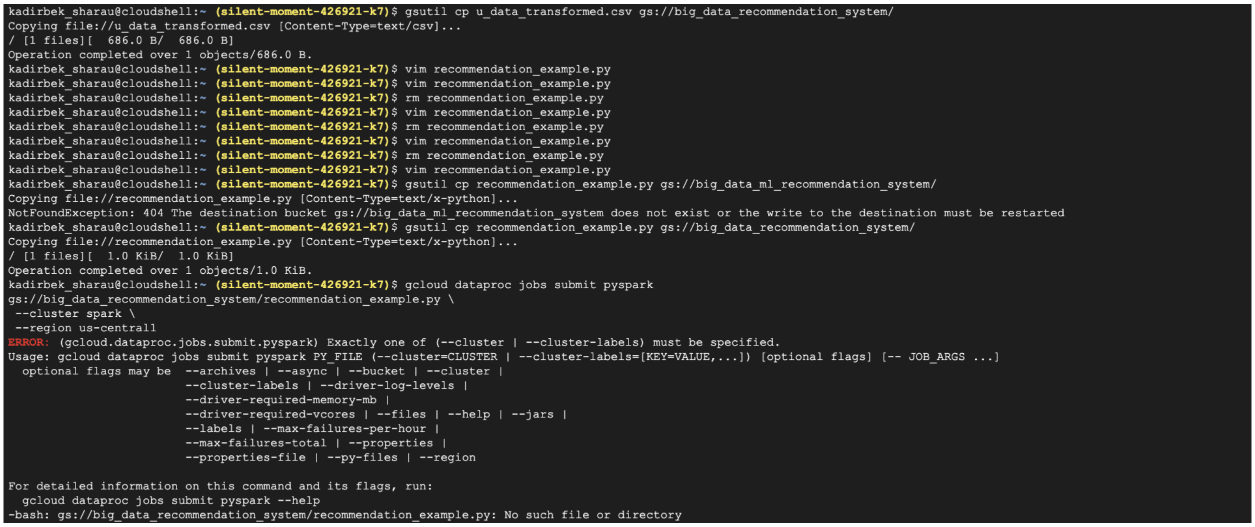
```
Welcome to Cloud Shell! Type "help" to get started.
Your Cloud Platform project in this session is set to silent-moment-426921-k7.
Use “gcloud config set project [PROJECT_ID]” to change to a different project.
kadirbek_sharau@cloudshell:~ (silent-moment-426921-k7)$ vim u.data
kadirbek_sharau@cloudshell:~ (silent-moment-426921-k7)$ echo '#!/bin/bash
> cat u.data | tr -s ' ' | cut -d' ' -f1-3 | tr ' ' ',' > u_data_transformed.csv' > transform_data.sh
kadirbek_sharau@cloudshell:~ (silent-moment-426921-k7)$ chmod +x transform_data.sh
```



```
kadirbek_sharau@cloudshell:~ (silent-moment-426921-k7)$ ./transform_data.sh
tr: missing operand
Try 'tr --help' for more information.
cut: the delimiter must be a single character
Try 'cut --help' for more information.
tr: missing operand after ‘,’
Two strings must be given when translating.
Try 'tr --help' for more information.
kadirbek_sharau@cloudshell:~ (silent-moment-426921-k7)$ cat u.data | tr -s ' ' | cut -d' ' -f1-3 | tr ' ' ',' > u_data_transformed.csv' > transform_data.sh
> ^C
kadirbek_sharau@cloudshell:~ (silent-moment-426921-k7)$ cat u.data | tr -s ' ' | cut -d' ' -f1-3 | tr ' ' ',' >
u_data_transformed.csv' > transform_data.sh
-bash: syntax error near unexpected token `newline'
> u_data_transformed.csv' > transform_data.sh
-bash: $'u_data_transformed.csv > transform_data.sh\nu_data_transformed.csv': command not found
```

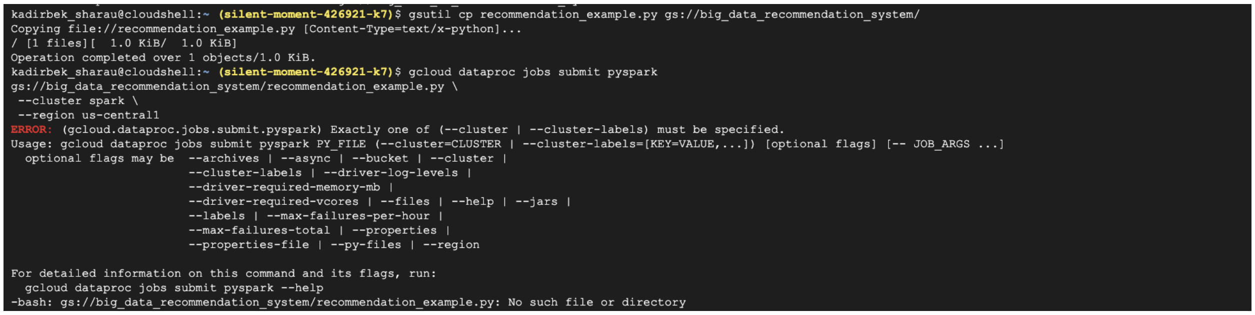


```
kadirbek_sharau@cloudshell:~ (silent-moment-426921-k7)$ echo '#!/bin/bash
> cat u.data | tr -s ' ' | cut -d' ' -f1-3 | tr ' ' ',' > u_data_transformed.csv' > transform_data.sh
kadirbek_sharau@cloudshell:~ (silent-moment-426921-k7)$ chmod +x transform_data.sh
kadirbek_sharau@cloudshell:~ (silent-moment-426921-k7)$ ./transform_data.sh
tr: missing operand
Try 'tr --help' for more information.
cut: the delimiter must be a single character
Try 'cut --help' for more information.
tr: missing operand after ‘,’
Two strings must be given when translating.
Try 'tr --help' for more information.
kadirbek_sharau@cloudshell:~ (silent-moment-426921-k7)$ echo '#!/bin/bash
> cat u.data | tr -s " " | cut -d" " -f1-3 | tr " " "," > u_data_transformed.csv' > transform_data.sh
kadirbek_sharau@cloudshell:~ (silent-moment-426921-k7)$ chmod +x transform_data.sh
kadirbek_sharau@cloudshell:~ (silent-moment-426921-k7)$ ./transform_data.sh
kadirbek_sharau@cloudshell:~ (silent-moment-426921-k7)$ echo '#!/bin/bash
> cat u.data | while read userid movieid rating timestamp
do
    echo "${userid},${movieid},${rating}"
done > u_data_transformed.csv' > transform_data.sh
kadirbek_sharau@cloudshell:~ (silent-moment-426921-k7)$ chmod +x transform_data.sh
kadirbek_sharau@cloudshell:~ (silent-moment-426921-k7)$ ./transform_data.sh
kadirbek_sharau@cloudshell:~ (silent-moment-426921-k7)$ gsutil cp u_data_transformed.csv gs://big_data_ml_recommendation_system/
Copying file://u_data_transformed.csv [Content-Type=text/csv]...
NotFoundException: 404 The destination bucket gs://big_data_ml_recommendation_system does not exist or the write to the destination must be restarted
```



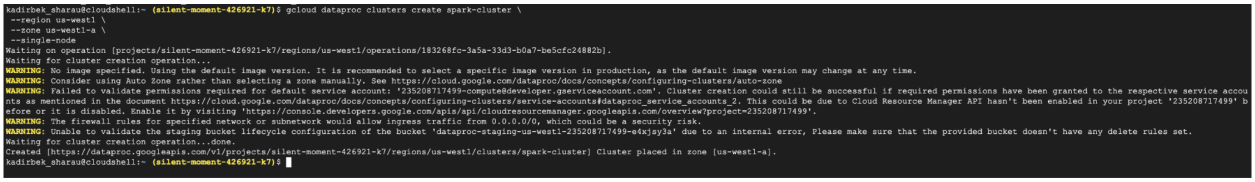
```
kadirbek_sharau@cloudshell:~ (silent-moment-426921-k7)$ gsutil cp u_data_transformed.csv gs://big_data_recommendation_system/
Copying file://u_data_transformed.csv [Content-Type=text/csv]...
/ [1 files][  686.0 B/  686.0 B]
Operation completed over 1 objects/686.0 B.
kadirbek_sharau@cloudshell:~ (silent-moment-426921-k7)$ vim recommendation_example.py
kadirbek_sharau@cloudshell:~ (silent-moment-426921-k7)$ vim recommendation_example.py
kadirbek_sharau@cloudshell:~ (silent-moment-426921-k7)$ rm recommendation_example.py
kadirbek_sharau@cloudshell:~ (silent-moment-426921-k7)$ vim recommendation_example.py
kadirbek_sharau@cloudshell:~ (silent-moment-426921-k7)$ rm recommendation_example.py
kadirbek_sharau@cloudshell:~ (silent-moment-426921-k7)$ vim recommendation_example.py
kadirbek_sharau@cloudshell:~ (silent-moment-426921-k7)$ rm recommendation_example.py
kadirbek_sharau@cloudshell:~ (silent-moment-426921-k7)$ vim recommendation_example.py
kadirbek_sharau@cloudshell:~ (silent-moment-426921-k7)$ gsutil cp recommendation_example.py gs://big_data_ml_recommendation_system/
Copying file://recommendation_example.py [Content-Type=text/x-python]...
NotFoundException: 404 The destination bucket gs://big_data_ml_recommendation_system does not exist or the write to the destination must be restarted
kadirbek_sharau@cloudshell:~ (silent-moment-426921-k7)$ gsutil cp recommendation_example.py gs://big_data_recommendation_system/
Copying file://recommendation_example.py [Content-Type=text/x-python]...
/ [1 files][  1.0 KiB/  1.0 KiB]
Operation completed over 1 objects/1.0 KiB.
kadirbek_sharau@cloudshell:~ (silent-moment-426921-k7)$ gcloud dataproc jobs submit pyspark
gs://big_data_recommendation_system/recommendation_example.py \
 --cluster spark \
 --region us-central1
ERROR: (gcloud.dataproc.jobs.submit.pyspark) Exactly one of (--cluster | --cluster-labels) must be specified.
Usage: gcloud dataproc jobs submit pyspark PY_FILE (--cluster=CLUSTER | --cluster-labels=[KEY=VALUE,...]) [optional flags] [-- JOB_ARGS ...]
  optional flags may be  --archives | --async | --bucket | --cluster |
                         --cluster-labels | --driver-log-levels |
                         --driver-required-memory-mb |
                         --driver-required-vcores | --files | --help | --jars |
                         --labels | --max-failures-per-hour |
                         --max-failures-total | --properties |
                         --properties-file | --py-files | --region

For detailed information on this command and its flags, run:
  gcloud dataproc jobs submit pyspark --help
-bash: gs://big_data_recommendation_system/recommendation_example.py: No such file or directory
```

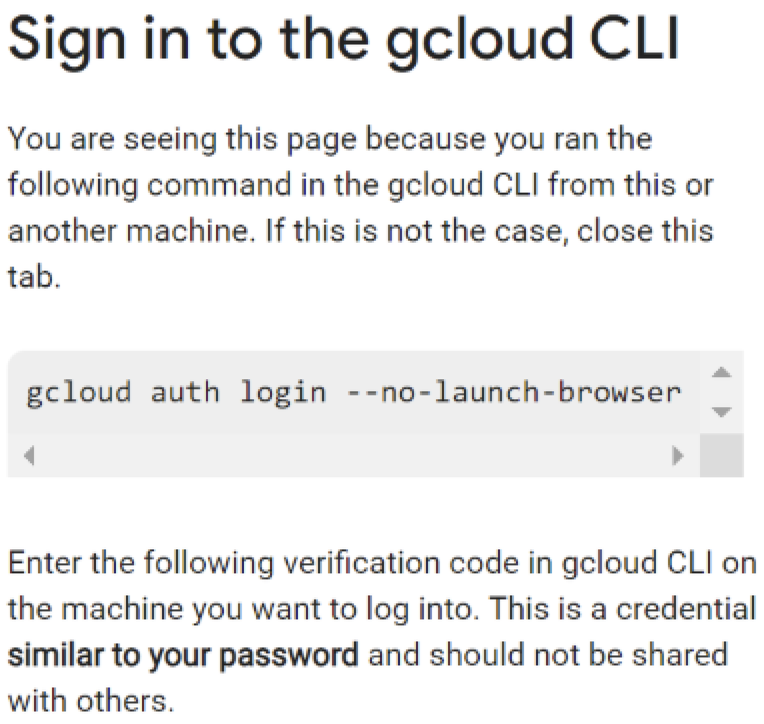


```
kadirbek_sharau@cloudshell:~ (silent-moment-426921-k7)$ gsutil cp recommendation_example.py gs://big_data_recommendation_system/
Copying file://recommendation_example.py [Content-Type=text/x-python]...
/ [1 files][  1.0 KiB/  1.0 KiB]
Operation completed over 1 objects/1.0 KiB.
kadirbek_sharau@cloudshell:~ (silent-moment-426921-k7)$ gcloud dataproc jobs submit pyspark
gs://big_data_recommendation_system/recommendation_example.py \
 --cluster spark \
 --region us-central1
ERROR: (gcloud.dataproc.jobs.submit.pyspark) Exactly one of (--cluster | --cluster-labels) must be specified.
Usage: gcloud dataproc jobs submit pyspark PY_FILE (--cluster=CLUSTER | --cluster-labels=[KEY=VALUE,...]) [optional flags] [-- JOB_ARGS ...]
  optional flags may be  --archives | --async | --bucket | --cluster |
                         --cluster-labels | --driver-log-levels |
                         --driver-required-memory-mb |
                         --driver-required-vcores | --files | --help | --jars |
                         --labels | --max-failures-per-hour |
                         --max-failures-total | --properties |
                         --properties-file | --py-files | --region

For detailed information on this command and its flags, run:
  gcloud dataproc jobs submit pyspark --help
-bash: gs://big_data_recommendation_system/recommendation_example.py: No such file or directory
```



```
kadirbek_sharau@cloudshell:~ (silent-moment-426921-k7)$ gcloud dataproc clusters create spark-cluster \
 --region us-west1 \
 --zone us-west1-a \
 --single-node
Waiting on operation [projects/silent-moment-426921-k7/regions/us-west1/operations/183268fc-3a5a-33d3-b0a7-be5cfc24882b].
Waiting for cluster creation operation...
WARNING: No image specified. Using the default image version. It is recommended to select a specific image version in production, as the default image version may change at any time.
WARNING: Consider using Auto Zone rather than selecting a zone manually. See https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/auto-zone
WARNING: Failed to validate permissions required for default service account: '235208717499-compute@developer.gserviceaccount.com'. Cluster creation could still be successful if required permissions have been granted to the respective service accou
nts as mentioned in the document https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/service-accounts#dataproc_service_accounts_2. This could be due to Cloud Resource Manager API hasn't been enabled in your project '235208717499' b
efore or it is disabled. Enable it by visiting 'https://console.developers.google.com/apis/api/cloudresourcemanager.googleapis.com/overview?project=235208717499'.
WARNING: The firewall rules for specified network or subnetwork would allow ingress traffic from 0.0.0.0/0, which could be a security risk.
WARNING: Unable to validate the staging bucket lifecycle configuration of the bucket 'dataproc-staging-us-west1-235208717499-e4xjsy3a' due to an internal error, Please make sure that the provided bucket doesn't have any delete rules set.
Waiting for cluster creation operation...done.
Created [https://dataproc.googleapis.com/v1/projects/silent-moment-426921-k7/regions/us-west1/clusters/spark-cluster] Cluster placed in zone [us-west1-a].
kadirbek_sharau@cloudshell:~ (silent-moment-426921-k7)$ 
```

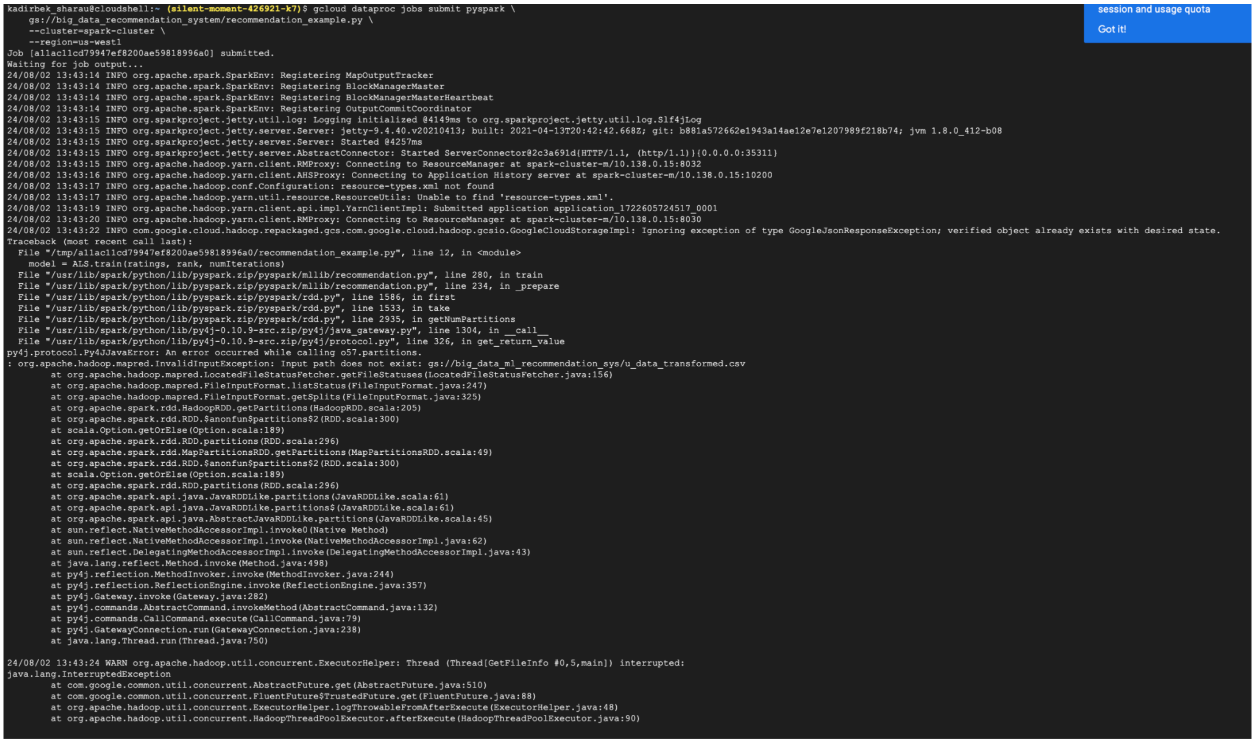


Sign in to the gcloud CLI

You are seeing this page because you ran the following command in the gcloud CLI from this or another machine. If this is not the case, close this tab.

```
gcloud auth login --no-launch-browser
```

Enter the following verification code in gcloud CLI on the machine you want to log into. This is a credential **similar to your password** and should not be shared with others.



```
kadirbek_sharau@cloudshell:~ (silent-moment-426921-k7)$ gcloud dataproc jobs submit pyspark \
    gs://big_data_recommendation_system/recommendation_example.py \
    --cluster=spark-cluster \
    --region=us-west1
Job [a11ac11cd79947ef8200ae59818996a0] submitted.
Waiting for job output...
24/08/02 13:43:14 INFO org.apache.spark.SparkEnv: Registering MapOutputTracker
24/08/02 13:43:14 INFO org.apache.spark.SparkEnv: Registering BlockManagerMaster
24/08/02 13:43:14 INFO org.apache.spark.SparkEnv: Registering BlockManagerMasterHeartbeat
24/08/02 13:43:14 INFO org.apache.spark.SparkEnv: Registering OutputCommitCoordinator
24/08/02 13:43:15 INFO org.sparkproject.jetty.util.log: Logging initialized @4149ms to org.sparkproject.jetty.util.log.Slf4jLog
24/08/02 13:43:15 INFO org.sparkproject.jetty.server.Server: jetty-9.4.40.v20210413; built: 2021-04-13T20:42:42.668Z; git: b881a572662e1943a14ae12e7e1207989f218b74; jvm 1.8.0_412-b08
24/08/02 13:43:15 INFO org.sparkproject.jetty.server.Server: Started @4257ms
24/08/02 13:43:15 INFO org.sparkproject.jetty.server.AbstractConnector: Started ServerConnector@2c3a691d{HTTP/1.1, (http/1.1)}{0.0.0.0:35311}
24/08/02 13:43:15 INFO org.apache.hadoop.yarn.client.RMProxy: Connecting to ResourceManager at spark-cluster-m/10.138.0.15:8032
24/08/02 13:43:16 INFO org.apache.hadoop.yarn.client.AHSProxy: Connecting to Application History server at spark-cluster-m/10.138.0.15:10200
24/08/02 13:43:17 INFO org.apache.hadoop.conf.Configuration: resource-types.xml not found
24/08/02 13:43:17 INFO org.apache.hadoop.yarn.util.resource.ResourceUtils: Unable to find 'resource-types.xml'.
24/08/02 13:43:19 INFO org.apache.hadoop.yarn.client.api.impl.YarnClientImpl: Submitted application application_1722605724517_0001
24/08/02 13:43:20 INFO org.apache.hadoop.yarn.client.RMProxy: Connecting to ResourceManager at spark-cluster-m/10.138.0.15:8030
24/08/02 13:43:22 INFO com.google.cloud.hadoop.repackaged.gcs.com.google.cloud.hadoop.gcsio.GoogleCloudStorageImpl: Ignoring exception of type GoogleJsonResponseException; verified object already exists with desired state.
Traceback (most recent call last):
  File "/tmp/a11ac11cd79947ef8200ae59818996a0/recommendation_example.py", line 12, in <module>
    model = ALS.train(ratings, rank, numIterations)
  File "/usr/lib/spark/python/lib/pyspark.zip/pyspark/mllib/recommendation.py", line 280, in train
  File "/usr/lib/spark/python/lib/pyspark.zip/pyspark/mllib/recommendation.py", line 234, in _prepare
  File "/usr/lib/spark/python/lib/pyspark.zip/pyspark/rdd.py", line 1586, in first
  File "/usr/lib/spark/python/lib/pyspark.zip/pyspark/rdd.py", line 1533, in take
  File "/usr/lib/spark/python/lib/pyspark.zip/pyspark/rdd.py", line 2935, in getNumPartitions
  File "/usr/lib/spark/python/lib/py4j-0.10.9-src.zip/py4j/java_gateway.py", line 1304, in __call__
  File "/usr/lib/spark/python/lib/py4j-0.10.9-src.zip/py4j/protocol.py", line 326, in get_return_value
py4j.protocol.Py4JJavaError: An error occurred while calling o57.partitions.
: org.apache.hadoop.mapred.InvalidInputException: Input path does not exist: gs://big_data_ml_recommendation_sys/u_data_transformed.csv
        at org.apache.hadoop.mapred.LocatedFileStatusFetcher.getFileStatuses(LocatedFileStatusFetcher.java:156)
        at org.apache.hadoop.mapred.FileInputFormat.listStatus(FileInputFormat.java:247)
        at org.apache.hadoop.mapred.FileInputFormat.getSplits(FileInputFormat.java:325)
        at org.apache.spark.rdd.HadoopRDD.getPartitions(HadoopRDD.scala:205)
        at org.apache.spark.rdd.RDD.$anonfun$partitions$2(RDD.scala:300)
        at scala.Option.getOrElse(Option.scala:189)
        at org.apache.spark.rdd.RDD.partitions(RDD.scala:296)
        at org.apache.spark.rdd.MapPartitionsRDD.getPartitions(MapPartitionsRDD.scala:49)
        at org.apache.spark.rdd.RDD.$anonfun$partitions$2(RDD.scala:300)
        at scala.Option.getOrElse(Option.scala:189)
        at org.apache.spark.rdd.RDD.partitions(RDD.scala:296)
        at org.apache.spark.api.java.JavaRDDLike.partitions(JavaRDDLike.scala:61)
        at org.apache.spark.api.java.JavaRDDLike.partitions$(JavaRDDLike.scala:61)
        at org.apache.spark.api.java.AbstractJavaRDDLike.partitions(JavaRDDLike.scala:45)
        at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
        at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:498)
        at py4j.reflection.MethodInvoker.invoke(MethodInvoker.java:244)
        at py4j.reflection.ReflectionEngine.invoke(ReflectionEngine.java:357)
        at py4j.Gateway.invoke(Gateway.java:282)
        at py4j.commands.AbstractCommand.invokeMethod(AbstractCommand.java:132)
        at py4j.commands.CallCommand.execute(CallCommand.java:79)
        at py4j.GatewayConnection.run(GatewayConnection.java:238)
        at java.lang.Thread.run(Thread.java:750)

24/08/02 13:43:24 WARN org.apache.hadoop.util.concurrent.ExecutorHelper: Thread (Thread[GetFileInfo #0,5,main]) interrupted:
java.lang.InterruptedException
        at com.google.common.util.concurrent.AbstractFuture.get(AbstractFuture.java:510)
        at com.google.common.util.concurrent.FluentFuture$TrustedFuture.get(FluentFuture.java:88)
        at org.apache.hadoop.util.concurrent.ExecutorHelper.logThrowableFromAfterExecute(ExecutorHelper.java:48)
        at org.apache.hadoop.util.concurrent.HadoopThreadPoolExecutor.afterExecute(HadoopThreadPoolExecutor.java:90)
```