

Project: Movie Recommendation with MLlib - Collaborative Filtering

Student :

Presented by MANICKAM RAVISEKAR ,
Master of Science in Computer Science, 19599 , Fall Semester 2022

**Professor : Dr Henry Chung
TA : Liang**

SAN FRANCISCO BAY
UNIVERSITY
47671 WestingHouse Dr.,
Fremont, CA 94539

ACKNOWLEDGEMENT

One of our master's degree Project for Machine Learning using,

Colab notebooks which allows us to combine executable code and rich text in a single document and saving same cells from ipynb to py (python).

Is an Interesting, which made me to learn new things, it is useful in designing and applying on Google Cloud Platform.

Using **MLlib (machine learning library)** is built on top of Spark as part of the Spark package.

For deploying this project, I would like to thank Dr. Henry Chang an TA Liang for providing all the required input .

Also, for all I would like to always pray to Almighty for giving us wisdom and power to understand things.

Content

Index :

Recommender System / Utility Matrix

Explicit vs Implicit ratings

Data Sparsity and Cold Start

Dataset with Explicit Ratings (MovieLens)

Colab notebook steps to Find the best model and recommendations
from slides 15 - 35

PySpark on Google Cloud Platform (slide 36)

Conclusion

References

Abstract

MovieLens is a recommender system and virtual community website that recommends movies for its users to watch, based on their film preferences using collaborative filtering. MovieLens 100M dataset is taken from the MovieLens website, which customizes user recommendation based on the ratings given by the user. To understand the concept of recommendation system better, we will work with this dataset.

Collaborative filtering uses similarities between users and items simultaneously to provide recommendations.

I would like thank Dr Henry chung for providing the guidance and assigning this project .

Recommender System is an information filtering tool that seeks to predict which product a user will like, and based on that, recommends a few products to the users. For example, Amazon can recommend new shopping items to buy, Netflix can recommend new movies to watch, and Google can recommend news that a user might be interested in. The two widely used approaches for building a recommender system are content-based filtering (CBF) and collaborative filtering (CF).

To understand the concept of recommender systems, let us look at an example. The below table shows the user-item *utility matrix* Y where the value R_{ui} denotes how item i has been rated by user u on a scale of 1–5. The missing entries (shown ? in Table) are the items that have not been rated by the respective user.

	Item 1	Item 2	Item 3	Item 4
User 1	2	5	1	3
User 2	4	?	?	1
User 3	?	4	2	?
User 4	2	4	3	1
User 5	1	3	2	?

Explicit v.s. Implicit ratings

There are two ways to gather user preference data to recommend items, the first method is to ask for **explicit ratings** from a user, typically on a concrete rating scale (such as rating a movie from one to five stars) making it easier to make extrapolations from data to predict future ratings. However, the drawback with explicit data is that it puts the responsibility of data collection on the user, who may not want to take the time to enter ratings. On the other hand, **implicit data** is easy to collect in large quantities without any extra effort on the part of the user. Unfortunately, it is much more difficult to work with.

Data Sparsity and Cold Start

In real world problems, the utility matrix is expected to be very sparse, as each user only encounters a small fraction of items among the vast pool of options available. The cold-start problem can arise during the addition of a new user or a new item where both do not have a history in terms of ratings. Sparsity can be calculated using the below function.

```
def get_mat_sparsity(ratings):
```

```
def get_mat_sparsity(ratings):  
  
    # Count the total number of ratings in the dataset  
  
    count_nonzero = ratings.select("rating").count()  
  
    # Count the number of distinct userIds and distinct movieIds  
  
    total_elements = ratings.select("userId").distinct().count() * ratings.select("movieId").distinct().count()  
  
    # Divide the numerator by the denominator  
  
    sparsity = (1.0 - (count_nonzero * 1.0) / total_elements) * 100  
  
    print("The ratings dataframe is ", "% .2f" % sparsity + "% sparse.")  
  
get_mat_sparsity(ratings)
```

Colab Cells Steps to Find the best model and recommendations

The screenshot shows a Google Colab interface with several tabs at the top: 'Inbox (11) - manickar', 'My Drive - Google Dr...', 'Google Colab Tutorial', 'Install PySpark 3 on G...', 'als-recommender-pys...', 'Copy of Untitled5.ipynb', and 'Gafterdisconnection -'. The current tab is 'Copy of Untitled5.ipynb'.

The notebook content includes:

- A cell output showing Java version information:

```
[ ] ! java -version
openjdk version "11.0.17" 2022-10-18
OpenJDK Runtime Environment (build 11.0.17+8-post-Ubuntu-1ubuntu218.04)
OpenJDK 64-Bit Server VM (build 11.0.17+8-post-Ubuntu-1ubuntu218.04, mixed mode, sharing)
```
- A section titled 'Install PySpark' containing a code cell:

```
# Install pyspark
!pip install pyspark
```

This cell is currently executing, as indicated by the progress bar and the text '54s' next to it. The output shows the pip installation process for PySpark:

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Collecting pyspark
  Downloading pyspark-3.3.1.tar.gz (281.4 MB)
    |██████████| 281.4 MB 43 kB/s
Collecting py4j==0.10.9.5
  Downloading py4j-0.10.9.5-py2.py3-none-any.whl (199 kB)
    |██████████| 199 kB 50.1 MB/s
Building wheels for collected packages: pyspark
  Building wheel for pyspark (setup.py) ... done
    Created wheel for pyspark: filename=pyspark-3.3.1-py3.py3-none-any.whl size=281845514 sha256=c3fbf65813a8420dc32dd334242b541120ccde337aa9d9d0affd625eb520db9f
    Stored in directory: /root/.cache/pip/wheels/42/59/f5/79a5bf931714cd201b26025347785f087370a10a3329a899c
Successfully built pyspark
Installing collected packages: py4j, pyspark
Successfully installed py4j-0.10.9.5 pyspark-3.3.1
```

At the bottom, a status bar indicates '54s completed at 8:02 PM'.

Inbox (11) - manickan | My Drive - Google Dr | Google Colab Tutorial | Install PySpark 3 on G | als-recommender-pys | Copy of Untitled5.ipynb | Gafterdisconnection - +

colab.research.google.com/drive/1gXF4hPRj83HEX1cgtymJnankei4kJDj#scrollTo=Nxne7COWwif1

Copy of Untitled5.ipynb

File Edit View Insert Runtime Tools Help

+ Code + Text

RAM Disk Editing

Comment Share M

After installation, we can create a Spark session and check its information.

```
[3] # Import SparkSession
from pyspark.sql import SparkSession
```

```
[4] # Create a Spark Session
spark = SparkSession.builder.master("local[*]").getOrCreate()
```

```
# Check Spark Session Information
spark
```

SparkSession - in-memory
SparkContext
Spark UI
Version v3.3.1
Master local[*]
AppName pyspark-shell

1s completed at 8:04 PM

Copy of Untitled5.ipynb

File Edit View Insert Runtime Tools Help All changes saved

Comment Share M

RAM Disk Editing

+ Code + Text

We can also test the installation by importing a Spark library.

Double-click (or enter) to edit

```
[6] # Import a Spark function from library
from pyspark.sql.functions import col
```

Import libraries

```
[7] #https://grouplens.org/datasets/movielens/
```

```
[8] import pandas as pd
from pyspark.sql.functions import col, explode
from pyspark import SparkContext
```

Initiate spark session

```
[9] from pyspark.sql import SparkSession
sc = SparkContext
# sc.setCheckpointDir('checkpoint')
spark = SparkSession.builder.appName('Recommendations').getOrCreate()
```



Inbox (11) - manickan | My Drive - Google Dr | Google Colab Tutorial | Install PySpark 3 on G | als-recommender-pys | Copy of Untitled5.ipynb | Gafterdisconnection - +

colab.research.google.com/drive/1gXF4hPrj83HEX1cgtymJnankei4kJDj#scrollTo=8zzrlDd0yfO_

Copy of Untitled5.ipynb

File Edit View Insert Runtime Tools Help All changes saved

+ Code + Text

RAM Disk Editing

1. Load data

```
[10] from google.colab import drive
drive.mount('/content/drive')

Mounted at /content/drive
```

!ls -l /content/drive/MyDrive/

```
total 3035
drwx----- 2 root root 4096 Nov 16 15:31 'Colab Notebooks'
-rw----- 1 root root 143 Mar 23 2022 'Copy of CS540 - Project Group List.gsheets'
-rw----- 1 root root 143 Nov 17 04:08 Gafterdisconnection.gslides
drwx----- 2 root root 4096 Nov 17 00:16 GCollaborative-filtering
-rw----- 1 root root 494429 Nov 16 02:42 gmovies.csv
-rw----- 1 root root 2483721 Nov 16 02:45 gratings.csv
-rw----- 1 root root 118658 Nov 16 02:48 gtags.csv
-rw----- 1 root root 143 Nov 17 03:07 GWeek9-HW.gslides
```

1. Dataset with Explicit Ratings (MovieLens)

MovieLens is a recommender system and virtual community website that recommends movies for its users to watch, based on their film preferences using collaborative filtering. MovieLens 100M dataset is taken from the MovieLens website, which customizes user recommendation based on the ratings given by the user. To understand the concept of recommendation system better, we will work with this dataset. This dataset can be downloaded from [here](#).

There are 2 tuples, movies and ratings which contains variables such as MovieID::Genre::Title and UserID::MovieID::Rating::Timestamp respectively.

Let's load the data and explore the data. To load the data as a spark dataframe, import pyspark and instantiate a spark session.

As show in the next slide

```
from pyspark.sql import SparkSession
spark = SparkSession.builder.appName('Recommendations').getOrCreate()
movies = spark.read.csv("movies.csv",header=True)
ratings = spark.read.csv("ratings.csv",header=True)
ratings.show()
```

Inbox (11) - manickan | My Drive - Google Dr | Google Colab Tutorial | Install PySpark 3 on G | als-recommender-pys | Copy of Untitled5.ipynb | Gafterdisconnection | +

colab.research.google.com/drive/1gXF4hPrj83HEX1cgtiywmJnankei4kJDj#scrollTo=FxeVTTazdRW

File Edit View Insert Runtime Tools Help All changes saved

Comment Share Editing

RAM Disk

[11] [12] movies = spark.read.csv("/content/drive/MyDrive/gmovies.csv",header=True)

[13] ratings = spark.read.csv("/content/drive/MyDrive/gratings.csv",header=True)

ratings.show()

userId	movieId	rating	timestamp
1	1	4.0	964982703
1	31	4.0	964981247
1	6	4.0	964982224
1	47	5.0	964983815
1	501	5.0	964982931
1	701	3.0	964984001
1	101	5.0	964980868
1	110	4.0	964982176
1	151	5.0	964984041
1	157	5.0	964984100
1	163	5.0	964983650
1	216	5.0	964981208
1	223	3.0	964980985
1	231	5.0	964981179
1	235	4.0	964980908
1	260	5.0	964981680
1	296	3.0	964982967
1	316	3.0	964982310
1	333	5.0	964981179
1	349	4.0	964982563

only showing top 20 rows

✓ 0s completed at 8:17 PM

Inbox (11) - manickan | My Drive - Google Dr | Google Colab Tutorial | Install PySpark 3 on G | als-recommender-pys | Copy of Untitled5.ipynb | Gafterdisconnection - +

colab.research.google.com/drive/1gF4hPRj83HEX1cgtymJnankei4kJDj#scrollTo=AO4Xvb8Eze5

Copy of Untitled5.ipynb

File Edit View Insert Runtime Tools Help All changes saved

Comment Share M

+ Code + Text RAM Disk Editing

[15] ratings.printSchema()

```
root
 |-- userId: string (nullable = true)
 |-- movieId: string (nullable = true)
 |-- rating: string (nullable = true)
 |-- timestamp: string (nullable = true)
```

[16] ratings = ratings.\n withColumn('userId', col('userId').cast('integer')).\\
 withColumn('movieId', col('movieId').cast('integer')).\\
 withColumn('rating', col('rating').cast('float')).\\
 drop('timestamp')\nratings.show()

userId	movieId	rating
1	1	4.0
1	3	4.0
1	6	4.0
1	47	5.0
1	50	5.0
1	70	3.0
1	101	5.0
1	110	4.0
1	151	5.0
1	157	5.0
1	163	5.0
1	216	5.0
1	223	3.0
1	231	5.0
1	235	4.0
1	260	5.0
1	296	3.0
1	316	3.0

0s completed at 8:18 PM

 Copy of Untitled5.ipynb

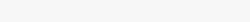
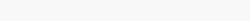
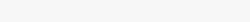
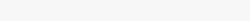
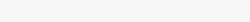
File Edit View Insert Runtime Tools Help All changes saved

Comment Share



+ Code + Text

RAM Disk



```
ratings = ratings.\n    withColumn('userId', col('userId').cast('integer')).\\n    withColumn('movieId', col('movieId').cast('integer')).\\n    withColumn('rating', col('rating').cast('float')).\\n    drop('timestamp')\nratings.show()
```

```
+-----+-----+\n|userId|movieId|rating|\n+-----+-----+\n|     1|       1|   4.0|\n|     1|       3|   4.0|\n|     1|       6|   4.0|\n|     1|      47|   5.0|\n|     1|      50|   5.0|\n|     1|      70|   3.0|\n|     1|     101|   5.0|\n|     1|     110|   4.0|\n|     1|     151|   5.0|\n|     1|     157|   5.0|\n|     1|     163|   5.0|\n|     1|     216|   5.0|\n|     1|     223|   3.0|\n|     1|     231|   5.0|\n|     1|     235|   4.0|\n|     1|     260|   5.0|\n|     1|     296|   3.0|\n|     1|     316|   3.0|\n|     1|     333|   5.0|\n|     1|     349|   4.0|\n+-----+-----+\nonly showing top 20 rows
```

Calculate sparsity

The screenshot shows a Google Colab notebook titled "Copy of Untitled5.ipynb". The code cell at the bottom has been executed, resulting in the output "The ratings dataframe is 98.30% empty." The status bar at the bottom indicates the execution took 0s and completed at 8:21 PM.

```
[17] # Count the total number of ratings in the dataset
numerator = ratings.select("rating").count()

[18] # Count the number of distinct userIds and distinct movieIds
num_users = ratings.select("userId").distinct().count()
num_movies = ratings.select("movieId").distinct().count()

[19] # Set the denominator equal to the number of users multiplied by the number of movies
denominator = num_users * num_movies

# Divide the numerator by the denominator
sparsity = (1.0 - (numerator * 1.0)/denominator)*100
print("The ratings dataframe is ", "%2f" % sparsity + "% empty.")

The ratings dataframe is 98.30% empty.
```

File Edit View Insert Runtime Tools Help

Comment Share M

RAM Disk Editing

Inbox (11) - manickan | My Drive - Google Dr | Google Colab Tutorial | Install PySpark 3 on | als-recommender-py | Copy of Untitled5.ipynb | Gafterdisconnection | + - ×

colab.research.google.com/drive/1gXF4hPRj83HEX1cgtywmJnankei4kJDj#scrollTo=VYfQEp8y0gWc

Code Text

Calculate sparsity

{x} Double-click (or enter) to edit

RAM Disk Editing

0s completed at 8:21 PM

Interpret ratings

Inbox (11) - mani | My Drive - Google | Google Colab Tutorials | Install PySpark 3 | als-recommender | Copy of Untitled5.ipynb | Copy of Copy of | Gafterdisconnect | +

colab.research.google.com/drive/1gXF4hPrj83HEX1cgtywmJnankei4kJDj#scrollTo=PW4g4MVQ03Gf

Copy of Untitled5.ipynb

File Edit View Insert Runtime Tools Help All changes saved

Comment Share M

+ Code + Text RAM Disk Editing

Interpret ratings

```
{x} # Group data by userId, count ratings
userId_ratings = ratings.groupBy("userId").count().orderBy('count', ascending=False)
userId_ratings.show()
```

userId	count
414	2698
599	2478
474	2108
448	1864
274	1346
610	1302
68	1260
380	1218
606	1115
288	1055
249	1046
387	1027
182	977
387	975
603	943
298	939
177	904
318	879
232	862
480	856

only showing top 20 rows

1s completed at 8:23 PM

Inbox (11) - manickan | My Drive - Google Dr | Google Colab Tutorial | Install PySpark 3 on G | als-recommender-pys | Copy of Untitled5.ipynb | Gafterdisconnection - +

colab.research.google.com/drive/1gXF4hPRj83HEx1cgtymJnankei4kJDj#scrollTo=r-CeD3Zz1PZz

Copy of Untitled5.ipynb

File Edit View Insert Runtime Tools Help All changes saved

+ Code + Text

[21] | 480| 836|
+-----+
only showing top 20 rows

```
# Group data by userId, count ratings
movieId_ratings = ratings.groupBy("movieId").count().orderBy('count', ascending=False)
movieId_ratings.show()
```

movieId	count
356	329
318	317
296	307
593	279
2574	278
260	251
480	238
110	237
589	224
527	220
2959	218
1	215
1196	211
2858	204
50	204
47	203
780	202
150	201
1198	200
4993	198

+-----+
only showing top 20 rows

RAM Disk ✓ Editing

Comment Share M

RAM Disk ✓ Editing

Comment Share M

Build Out An ALS Model

Inbox (11) - manickan | My Drive - Google Driv | Google Colab Tutorial | Install PySpark 3 on G | als-recommender-pys | Copy of Untitled5.ipynb | Gafterdisconnection - | +

colab.research.google.com/drive/1gXF4hPRj83HEX1cgtwmJnankei4kJDj#scrollTo=0l7ulqbW1yVm

Copy of Untitled5.ipynb

File Edit View Insert Runtime Tools Help

Comment Share M

+ Code + Text

RAM Disk Editing

[22]

1s	1120	200
	4993	198
	-----	-----
	only showing top 20 rows	

{x} Build Out An ALS Model

```
[23] # Import the required functions
from pyspark.ml.evaluation import RegressionEvaluator
from pyspark.ml.recommendation import ALS
from pyspark.ml.tuning import ParamGridBuilder, CrossValidator
```

```
[24] # Create test and train set
(train, test) = ratings.randomSplit([0.8, 0.2], seed = 1234)
```

```
[25] # Create ALS model
als = ALS(userCol="userId", itemCol="movieId", ratingCol="rating", nonnegative = True, implicitPrefs = False, coldStartStrategy="drop")
```

```
# Confirm that a model called "als" was created
type(als)
```

ans=pyspark.ml.recommendation.ALS

0s completed at 8:27 PM

Tell Spark how to tune your ALS model

The screenshot shows a Google Colab notebook titled "Copy of Untitled5.ipynb". The code cell at the bottom has a red box drawn around its output, which displays the message "Num models to be tested: 16".

```
# Import the requisite items
from pyspark.ml.evaluation import RegressionEvaluator
from pyspark.ml.tuning import ParamGridBuilder, CrossValidator

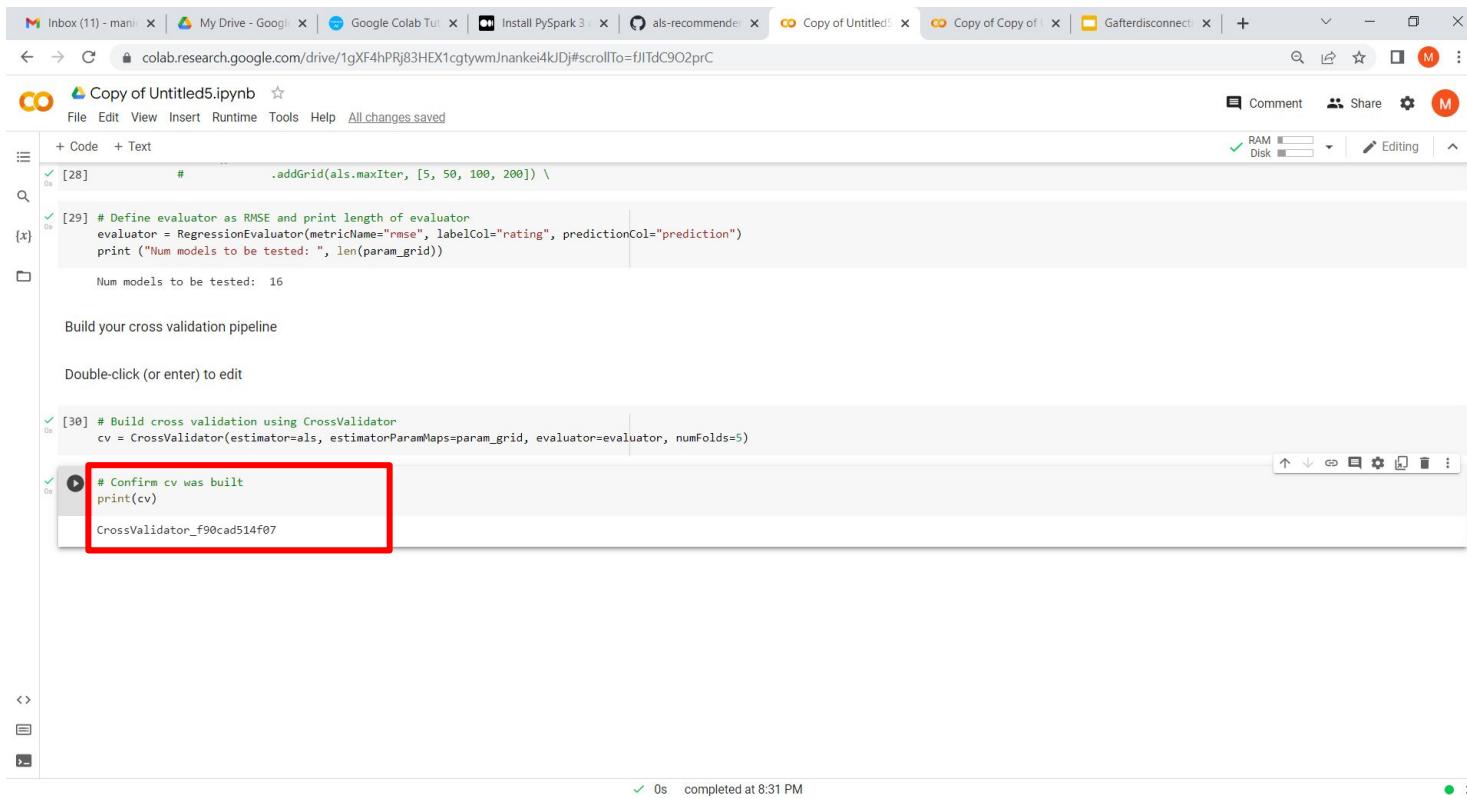
# Add hyperparameters and their respective values to param_grid
param_grid = ParamGridBuilder() \
    .addGrid(als.rank, [10, 50, 100, 150]) \
    .addGrid(als.regParam, [.01, .05, .1, .15]) \
    .build()

# Define evaluator as RMSE and print length of evaluator
evaluator = RegressionEvaluator(metricName="rmse", labelCol="rating", predictionCol="prediction")
print ("Num models to be tested: ", len(param_grid))

Num models to be tested: 16
```

0s completed at 8:29 PM

Build your cross validation pipeline



Inbox (11) - mail | My Drive - Google | Google Colab Tuto | Install PySpark 3 | als-recommender | Copy of Untitled | Copy of Copy of | Garterdisconnect | +

colab.research.google.com/drive/1gXF4hPRj83HEX1cgtymJnankei4kJDj#scrollTo=fjITdC9O2prC

File Edit View Insert Runtime Tools Help All changes saved

Comment Share M

RAM Disk Editing

```
[28] # .addGrid(als.maxIter, [5, 50, 100, 200]) \
```

```
[29] # Define evaluator as RMSE and print length of evaluator
evaluator = RegressionEvaluator(metricName="rmse", labelCol="rating", predictionCol="prediction")
print ("Num models to be tested: ", len(param_grid))
```

Num models to be tested: 16

Build your cross validation pipeline

Double-click (or enter) to edit

```
[30] # Build cross validation using CrossValidator
cv = CrossValidator(estimator=als, estimatorParamMaps=param_grid, evaluator=evaluator, numFolds=5)
```

```
# Confirm cv was built
print(cv)

CrossValidator_f90cad514f07
```

0s completed at 8:31 PM

Best Model and Best Model Parameters

The screenshot shows a Google Colab notebook titled "Copy of Untitled5.ipynb". The code cell [29] defines an evaluator and prints the number of models to be tested:

```
[29] # Define evaluator as RMSE and print length of evaluator
evaluator = RegressionEvaluator(metricName="rmse", labelCol="rating", predictionCol="prediction")
print ("Num models to be tested: ", len(param_grid))
```

The output of this cell is:

```
Num models to be tested: 16
```

The code cell [30] builds a cross-validation pipeline:

```
[30] # Build cross validation using CrossValidator
cv = CrossValidator(estimator=als, estimatorParamMaps=param_grid, evaluator=evaluator, numFolds=5)
```

The code cell [31] confirms the cv object was built:

```
[31] # Confirm cv was built
print(cv)
```

The output of this cell is:

```
CrossValidator_f90cad514f07
```

The code cell [32] fits the cross-validator to the training data:

```
[32] #Fit cross validator to the 'train' dataset
model = cv.fit(train)
```

The interface includes a toolbar at the top with various icons for file operations, a sidebar with navigation buttons, and a bottom toolbar with additional icons.

CO Copy of Untitled5.ipynb 

File Edit View Insert Runtime Tools Help

+ Code + Text  

[28] # .addGrid(als.maxIter, [5, 50, 100, 200]) \

[29] # Define evaluator as RMSE and print length of evaluator
evaluator = RegressionEvaluator(metricName="rmse", labelCol="rating", predictionCol="prediction")
print ("Num models to be tested: ", len(param_grid))

Num models to be tested: 16

Build your cross validation pipeline

Double-click (or enter) to edit

[30] # Build cross validation using CrossValidator
cv = CrossValidator(estimator=als, estimatorParamMaps=param_grid, evaluator=evaluator, numFolds=5)

[31] # Confirm cv was built
print(cv)

CrossValidator_f90cad514f07

Best Model and Best Model Parameters

1h  #Fit cross validator to the 'train' dataset
model = cv.fit(train)

Copy of Untitled5.ipynb ⌄

File Edit View Insert Runtime Tools Help All changes saved

Comment Share ⌄

M

+ Code + Text

RAM Disk ✓ Editing ⌄

```
✓ [28] # .addGrid(als.maxIter, [5, 50, 100, 200]) \n\n✓ [29] # Define evaluator as RMSE and print length of evaluator\n  evaluator = RegressionEvaluator(metricName="rmse", labelCol="rating", predictionCol="prediction")\n  print ("Num models to be tested: ", len(param_grid))\n\n  Num models to be tested: 16
```

Build your cross validation pipeline

Double-click (or enter) to edit

```
✓ [30] # Build cross validation using CrossValidator\n  cv = CrossValidator(estimator=als, estimatorParamMaps=param_grid, evaluator=evaluator, numFolds=5)\n\n✓ [31] # Confirm cv was built\n  print(cv)\n\n  CrossValidator_f90cad514f07
```

Best Model and Best Model Parameters

```
✓ [32] #Fit cross validator to the 'train' dataset\n  model = cv.fit(train)\n\n✓ [33] #Extract best model from the cv model above\n  best_model = model.bestModel
```



Copy of Untitled5.ipynb

File Edit View Insert Runtime Tools Help All changes saved

Comment Share M

+ Code + Text

RAM Disk Editing

```
✓ [30] # Build cross validation using CrossValidator
      cv = CrossValidator(estimator=als, estimatorParamMaps=param_grid, evaluator=evaluator, numFolds=5)

{x}
✓ [31] # Confirm cv was built
      print(cv)

CrossValidator_f90cad514f07
```

Best Model and Best Model Parameters

```
✓ [32] #Fit cross validator to the 'train' dataset
      model = cv.fit(train)

✓ [33] #Extract best model from the cv model above
      best_model = model.bestModel

# Print best_model
print(type(best_model))

<class 'pyspark.ml.recommendation.ALSModel'>
```



Print best model

The screenshot shows a Google Colab notebook titled "Copy of Untitled5.ipynb". The code cell [34] prints the type of the best model, which is an ALSModel. The subsequent code cells [35] through [38] extract and print various parameters of the ALS model, including Rank, MaxIter, and RegParam. The output for each parameter is shown in a red-bordered box.

```
#Fit cross validator to the "train" dataset
model = cv.fit(train)

#Extract best model from the cv model above
best_model = model.bestModel

# Print best_model
print(type(best_model))

<class 'pyspark.ml.recommendation.ALSModel'>

# Complete the code below to extract the ALS model parameters
print("**Best Model**")
**Best Model**

# # Print "Rank"
print(" Rank:", best_model._java_obj.parent().getRank())
Rank: 50

# Print "MaxIter"
print(" MaxIter:", best_model._java_obj.parent().getMaxIter())
MaxIter: 10

# Print "RegParam"
print(" RegParam:", best_model._java_obj.parent().getRegParam())
RegParam: 0.15
```

Print best model

****Best Model****
Rank: 50
MaxIter: 10
RegParam: 0.15

View the predictions

The screenshot shows a Google Colab notebook titled "Copy of Untitled5.ipynb". The code cell [39] contains the following Python code:

```
[39] # View the predictions
test_predictions = best_model.transform(test)
RMSE = evaluator.evaluate(test_predictions)
print(RMSE)
```

The output of this cell is highlighted with a red box and shows the value 0.8678589181737388.

At the bottom of the screen, there is a status bar with the text "completed at 10:25 PM".

Test predictions

Inbox (11) | My Drive - Go | Google Colab | Install PySpark | als-recomm... | Copy of Untitl... | Copy of Copy | Gafterdisconn... | New Tab | + | ×

colab.research.google.com/drive/1gXF4hPRj83HEX1cgtywmJnankei4kJDj#scrollTo=ve8J_jj_Q76i

Copy of Untitled5.ipynb

File Edit View Insert Runtime Tools Help

+ Code + Text

[39] # View the predictions
test_predictions = best_model.transform(test)
RMSE = evaluator.evaluate(test_predictions)
print(RMSE)

0.8678589181737388

test_predictions.show()

userId	movieId	rating	prediction
148	356	4.0	3.464672
148	4896	4.0	3.4670408
148	4993	3.0	3.52035
148	7153	3.0	3.4120665
148	8368	4.0	3.5930173
148	40629	5.0	3.2363656
148	50872	3.0	3.7235408
148	60069	4.5	3.6385822
148	69757	3.5	3.519815
148	72998	4.0	3.232824
148	81847	4.5	3.5208135
148	98491	5.0	3.7861571
148	115617	3.5	3.4901662
148	122886	3.5	3.4427116
463	296	4.0	4.1831555
463	527	4.0	3.736365
463	2019	4.0	3.9875493
471	527	4.5	3.8419898
471	6016	4.0	3.9658365
471	6333	2.5	3.255078

only showing top 20 rows

1s completed at 10:26 PM

Make Recommendations

The screenshot shows a Google Colab notebook interface. The top bar displays multiple tabs including 'Inbox (11) - m', 'My Drive - Go', 'Google Colab', 'Install PySpark', 'als-recomm...', 'Copy of Untitled...', 'Copy of Copy...', 'Gafterdisconn...', 'New Tab', and a '+' button. The main content area has a red box highlighting the title 'Make Recommendations'.

Code Block:

```
# generates n recommendations for all users
nrecommendations = best_model.recommendForAllUsers(10)
nrecommendations.limit(10).show()
```

Output:

```
+-----+|userId|  recommendations|+-----+| 1|[3379, 5.7355146...]| 3|[70946, 4.866154...]| 5|[5490, 4.5791554...]| 6|[3925, 4.8140335...]| 9|[3379, 4.935473]...|12|[42730, 5.655360...]|13|[33649, 4.978508...]|15|[3379, 4.527062]...|16|[3379, 4.614239]...|17|[3379, 5.1721935...|+-----+
```

At the bottom, a status bar indicates '11s completed at 10:28 PM'.

Inbox (11) - m | My Drive - Go | Google Colab | Install PySpark | als-recommen | Copy of Until | Copy of Copy | Gafterdisconn | New Tab | +

colab.research.google.com/drive/1gXF4hPrj83HEX1cgtymJnankei4kJDj#scrollTo=VWED1WCxRuc1

Copy of Untitled5.ipynb

File Edit View Insert Runtime Tools Help All changes saved

Comment Share M

+ Code + Text

RAM Disk ✓ Editing

[42] nrecommendations = best_model.recommendForAllUsers(10)
nrecommendations.limit(10).show()

userId	recommendations
1	[{3379, 5.7355146...}]
3	[{70946, 4.866154...}]
5	[{5490, 4.5791554...}]
6	[{3925, 4.8140335...}]
9	[{3379, 4.935473}...]
12	[{42730, 5.655360...}]
13	[{33649, 4.978508...}]
15	[{3379, 4.527062}...]
16	[{3379, 4.614239}...]
17	[{3379, 5.1721935...}]

nrecommendations = nrecommendations\
.withColumn("rec_exp", explode("recommendations"))\
.select('userId', col("rec_exp.movieId"), col("rec_exp.rating"))

nrecommendations.limit(10).show()

userId	movieId	rating
1	3379	5.7355146
1	33649	5.5564117
1	5490	5.47559
1	171495	5.4036317
1	5328	5.361952
1	5416	5.361952
1	3951	5.361952
1	78836	5.3408904
1	5915	5.309985
1	8477	5.3042073

6s completed at 10:29 PM

Do the recommendations make sense? Lets merge movie name and genres to the recommendation matrix for interpretability.

The screenshot shows a Jupyter Notebook interface in Google Colab. The notebook has two cells displayed:

Cell 1:

```
nrecommendations.join(movies, on='movieId').filter('userId = 100').show()
```

This cell displays a DataFrame with columns: movieId, userId, rating, title, and genres. The data includes various movies like Strictly Sexual, Saving Face, The Beach, Peacock Warrior, Glory Road, Anne of Green Gables, Watermark, Blue Planet II, Woman Under the I..., and Zeitgeist: Moving... with their respective ratings and genres.

Cell 2:

```
[45] ratings.join(movies, on='movieId').filter('userId = 100').sort('rating', ascending=False).limit(10).show()
```

This cell displays a DataFrame with columns: movieId, userId, rating, title, and genres. It shows the top 10 highest-rated movies for user ID 100, including Top Gun, Terms of Endearment, Christmas Vacation, Officer and a Gentleman, Sweet Home Alabama, Maverick, Father of the Bride, Sleepless in Seattle, Casino, and Tombstone.

 GCF-Wweek09Untitled5.ipynb ⭐

File Edit View Insert Runtime Tools Help All changes saved

Comment Share ⌂

M

+ Code + Text

RAM Disk ⌂

Editing ⌂

↑ ↓ ⌂ ⌂ ⌂ ⌂ ⌂ ⌂

nrecommendations.join(movies, on='movieId').filter('userId = 100').show()

movieId	userId	rating	title	genres
67618	100	5.15273	Strictly Sexual (... Comedy Drama Romance	
33649	100	5.095666	Saving Face (2004) Comedy Drama Romance	
3379	100	5.0508056	On the Beach (1959) Drama	
45503	100	4.9322023	Peaceful Warrior ... Drama	
42730	100	4.919008	Glory Road (2006) Drama	
74282	100	4.916703	Anne of Green Gab... Children Drama Ro...	
117531	100	4.9119444	Watermark (2014) Documentary	
179135	100	4.9119444	Blue Planet II (2... Documentary	
7071	100	4.9119444	Woman Under the I... Drama	
84273	100	4.9119444	Zeitgeist: Moving... Documentary	

[45] ratings.join(movies, on='movieId').filter('userId = 100').sort('rating', ascending=False).limit(10).show()

movieId	userId	rating	title	genres
1101	100	5.0	Top Gun (1986) Action Romance	
1958	100	5.0	Terms of Endearment (1981) Comedy Drama	
2423	100	5.0	Christmas Vacation (1989) Comedy	
4041	100	5.0	Officer and a Gentleman (1980) Drama Romance	
5620	100	5.0	Sweet Home Alabama (2002) Comedy Romance	
368	100	4.5	Maverick (1994) Adventure Comedy Romance	
934	100	4.5	Father of the Bride (1991) Comedy	
539	100	4.5	Sleepless in Seattle (1993) Comedy Drama Romance	
16	100	4.5	Casino (1995) Crime Drama	
553	100	4.5	Tombstone (1993) Action Drama Western	

✓ 1s completed at 10:32 PM

GCF_Wweek09U...ipynb ⌂

gcf_wweek09untitled5.ipynb ⌂

copy_of_untitled5.py ⌂

Copy_of_Untitled5.ipynb ⌂

Show all ⌂

Testing movielens.py on Google Cloud Platform

Save the colab cells notebooks program “movielens.ipynb” to python “movielens.py”
PySpark program and output screen to find the Find the best model and recommendations are enclosed in
Below slides.

```
# Import SparkSession
from pyspark.sql import SparkSession

# Create a Spark Session
spark = SparkSession.builder.master("local[*]").getOrCreate()

# Check Spark Session Information
spark

"""We can also test the installation by importing a Spark library."""

# Import a Spark function from library
from pyspark.sql.functions import col

"""Import libraries"""

#https://grouplens.org/datasets/movielens/

import pandas as pd
from pyspark.sql.functions import col, explode
from pyspark import SparkContext

"""Initiate spark session"""

from pyspark.sql import SparkSession
sc = SparkContext
# sc.setCheckpointDir('checkpoint')
spark = SparkSession.builder.appName('Recommendations').getOrCreate()
```

"""\\"1. Load data"""\\"

```
from google.colab import drive
drive.mount('/content/drive')
!ls -l /content/drive/MyDrive/
movies = spark.read.csv("/content/drive/MyDrive/gmovies.csv",header=True)
ratings = spark.read.csv("/content/drive/MyDrive/gratings.csv",header=True)
ratings.show()
ratings.printSchema()
ratings = ratings.\
    withColumn('userId', col('userId').cast('integer')).\\
    withColumn('movield', col('movield').cast('integer')).\\
    withColumn('rating', col('rating').cast('float')).\\
    drop('timestamp')
ratings.show()
```

"""\\"Calculate sparsity"""\\"

```
# Count the total number of ratings in the dataset
numerator = ratings.select("rating").count()
# Count the number of distinct userIds and distinct movields
num_users = ratings.select("userId").distinct().count()
num_movies = ratings.select("movield").distinct().count()
# Set the denominator equal to the number of users multiplied by the number of movies
denominator = num_users * num_movies
```

Divide the numerator by the denominator

```
sparsity = (1.0 - (numerator * 1.0) / denominator) * 100
```

```
print("The ratings dataframe is ", "%2f" % sparsity + "% empty.")
```

"""\\"Interpret ratings"""\\"

```
# Group data by userId, count ratings
```

```
userId_ratings = ratings.groupBy("userId").count().orderBy('count', ascending=False)
userId_ratings.show()
```

```
# Group data by movield, count ratings
```

```
movield_ratings = ratings.groupBy("movield").count().orderBy('count', ascending=False)
movield_ratings.show()
```

```
"""Build Out An ALS Model"""
# Import the required functions
from pyspark.ml.evaluation import RegressionEvaluator
from pyspark.ml.recommendation import ALS
from pyspark.ml.tuning import ParamGridBuilder, CrossValidator
# Create test and train set
(train, test) = ratings.randomSplit([0.8, 0.2], seed = 1234)
# Create ALS model
als = ALS(userCol="userId", itemCol="movielid", ratingCol="rating", nonnegative = True, implicitPrefs = False, coldStartStrategy="drop")
# Confirm that a model called "als" was created
type(als)
"""Tell Spark how to tune your ALS model"""
# Import the requisite items
from pyspark.ml.evaluation import RegressionEvaluator
from pyspark.ml.tuning import ParamGridBuilder, CrossValidator

# Add hyperparameters and their respective values to param_grid
param_grid = ParamGridBuilder() \
    .addGrid(als.rank, [10, 50, 100, 150]) \
    .addGrid(als.regParam, [.01, .05, .1, .15]) \
    .build()
#     .addGrid(als.maxIter, [5, 50, 100, 200]) \

# Define evaluator as RMSE and print length of evaluator
evaluator = RegressionEvaluator(metricName="rmse", labelCol="rating", predictionCol="prediction")
print ("Num models to be tested: ", len(param_grid))

"""Build your cross validation pipeline"""
# Build cross validation using CrossValidator
cv = CrossValidator(estimator=als, estimatorParamMaps=param_grid, evaluator=evaluator, numFolds=5)
# Confirm cv was built
print(cv)
```

```
"""Best Model and Best Model Parameters"""
#Fit cross validator to the 'train' dataset
model = cv.fit(train)
#Extract best model from the cv model above
best_model = model.bestModel
# Print best_model
print(type(best_model))
# Complete the code below to extract the ALS model parameters
print("**Best Model**")
## Print "Rank"
print(" Rank:", best_model._java_obj.parent().getRank())
# Print "MaxIter"
print(" MaxIter:", best_model._java_obj.parent().getMaxIter())
# Print "RegParam"
print(" RegParam:", best_model._java_obj.parent().getRegParam())
# View the predictions
test_predictions = best_model.transform(test)
RMSE = evaluator.evaluate(test_predictions)
print(RMSE)
test_predictions.show()
"""Make Recommendations"""
# Generate n Recommendations for all users
nrecommendations = best_model.recommendForAllUsers(10)
nrecommendations.limit(10).show()
nrecommendations = nrecommendations\
    .withColumn("rec_exp", explode("recommendations"))\
    .select('userId', col("rec_exp.movieId"), col("rec_exp.rating"))
nrecommendations.limit(10).show()
"""Do the recommendations make sense? Lets merge movie name and genres to th recommendation matrix for interpretability. """
nrecommendations.join(movies, on='movieId').filter('userId = 100').show()
ratings.join(movies, on='movieId').filter('userId = 100').sort('rating', ascending=False).limit(10).show()
```

Google Cloud Platform VM INSTANCE

Inbox (11) - manickam@student... | GGCP - Google Slides | VM instances – Compute Engine

console.cloud.google.com/compute/instances?onCreate=true&project=scala-3

Start your Free Trial with \$300 in credit. Don't worry—you won't be charged if you run out of credits. [Learn more](#)

DISMISS ACTIVATE

Google Cloud Compute Engine VM instances CREATE INSTANCE IMPORT VM REFRESH OPERATIONS HELP ASSISTANT LEARN SHOW INFO PANEL

Virtual machines

- VM instances
- Instance templates
- Sole-tenant nodes
- Machine images
- TPUs
- Committed use discounts
- Migrate to Virtual Machin...

VM instances are highly configurable virtual machines for running workloads on Google infrastructure. [Learn more](#)

Filter Enter property name or value

Status	Name	Zone	Recommendations	In use by	Internal IP	External IP	Connect
<input type="checkbox"/>	<input checked="" type="checkbox"/> instance-for-mlib	us-central1-a			10.128.0.3 (nic0)	34.134.118.89 (nic0)	SSH

Related actions

- Explore Backup and DR NEW Back up your VMs and set up disaster recovery
- View billing report View and manage your Compute Engine billing
- Monitor VMs View outlier VMs across metrics like CPU and network
- Explore VM logs View, search, analyze, and download VM instance logs
- Set up firewall rules Control traffic to and from a VM instance
- Patch management Schedule patch updates and view patch compliance on VM instances

Show all

GGCP (1).pdf GGCP.pdf GGCP.pptx

Inbox (11) - manickam@student: ~ | GGCP - Google Slides | instance-for-mlib – Compute Eng | +

console.cloud.google.com/compute/instancesDetail/zones/us-central1-a/instances/instance-for-mlib?project=scala-3

Start your Free Trial with \$300 in credit. Don't worry—you won't be charged if you run out of credits. [Learn more](#)

DISMISS ACTIVATE

Google Cloud scala-3 Search for resources, docs, products, and more (/) Search

Compute Engine instance-for-... EDIT RESET CREATE MACHINE IMAGE CREATE SIMILAR OPERATIONS HELP ASSISTANT LEARN

Virtual machines VM instances Instance templates Sole-tenant nodes Machine images TPUs Committed use discounts Migrate to Virtual Machin...

DETAILS OBSERVABILITY OS INFO SCREENSHOT

SSH CONNECT TO SERIAL CONSOLE

Connecting to serial ports is disabled

Logs

Cloud Logging Serial port 1 (console)

SHOW MORE

Basic information

Name	instance-for-mlib
Instance Id	6631235624795226025
Description	None
Type	Instance
Status	Running
Creation time	Nov 18, 2022, 2:03:51 PM UTC-08:00
Zone	us-central1-a
Instance template	None
Instance type	None

GGCP (1).pdf GGCP.pdf GGCP.pptx Show all

Inbox (11) - manickam@student: ~ | GGCP - Google Slides | instance-for-mlib – Compute Eng | +

console.cloud.google.com/compute/instancesDetail/zones/us-central1-a/instances/instance-for-mlib?project=scala-3

Start your Free Trial with \$300 in credit. Don't worry—you won't be charged if you run out of credits. [Learn more](#)

DISMISS ACTIVATE

Google Cloud scala-3 Search for resources, docs, products, and more (/) Search

Compute Engine instance-for... EDIT RESET CREATE MACHINE IMAGE CREATE SIMILAR OPERATIONS HELP ASSISTANT LEARN

Virtual machines

- VM instances
- Instance templates
- Sole-tenant nodes
- Machine images
- TPUs
- Committed use discounts
- Migrate to Virtual Machin...

Storage

- Disks
- Marketplace
- Release Notes

DETAILS OBSERVABILITY OS INFO Screenshot

Status	Running
Creation time	Nov 18, 2022, 2:03:51 PM UTC-08:00
Zone	us-central1-a
Instance template	None
In use by	None
Reservations	Automatically choose
Labels	None
Tags	-
Deletion protection	Disabled
Confidential VM service	Disabled
Preserved state size	0 GB

Machine configuration

Machine type	e2-medium
CPU platform	Intel Broadwell
Architecture	x86/64
vCPUs to core ratio	-
Custom visible cores	-

GGCP (1).pdf GGCP.pdf GGCP.pptx Show all

Inbox (11) - manickam@student: ~ | GGCP - Google Slides | instance-for-mlib – Compute Eng | +

console.cloud.google.com/compute/instancesDetail/zones/us-central1-a/instances/instance-for-mlib?project=scala-3&pageState={"duration":("groupValue":"PT1H","customValue":null)}

Start your Free Trial with \$300 in credit. Don't worry—you won't be charged if you run out of credits. [Learn more](#)

DISMISS ACTIVATE

Google Cloud scala-3 Search for resources, docs, products, and more (/) Search

Compute Engine instance-for-... EDIT RESET CREATE MACHINE IMAGE CREATE SIMILAR OPERATIONS HELP ASSISTANT LEARN

Virtual machines VM instances Instance templates Sole-tenant nodes Machine images TPUs Committed use discounts Migrate to Virtual Machin...

DETAILS OBSERVABILITY OS INFO Screenshot

Display device Disabled
Enable to use screen capturing and recording tools

GPUs None

Networking

Public DNS PTR Record None

Total egress bandwidth tier —

NIC type —

→ VIEW IN NETWORK TOPOLOGY

Firewalls

HTTP traffic On

HTTPS traffic On

Network tags

http-server https-server

GGCP (1).pdf GGCP.pdf GGCP.pptx Show all

Inbox (11) - manickam@student: ~ | GGCP - Google Slides | instance-for-mlib – Compute Eng | +

console.cloud.google.com/compute/instancesDetail/zones/us-central1-a/instances/instance-for-mlib?project=scala-3&pageState={"duration":("groupValue":"PT1H","customValue":null)}

Start your Free Trial with \$300 in credit. Don't worry—you won't be charged if you run out of credits. [Learn more](#)

DISMISS ACTIVATE

Google Cloud scala-3 Search for resources, docs, products, and more (/) Search

Compute Engine instance-for-... EDIT RESET CREATE MACHINE IMAGE CREATE SIMILAR OPERATIONS HELP ASSISTANT LEARN

Virtual machines VM instances Instance templates Sole-tenant nodes Machine images TPUs Committed use discounts Migrate to Virtual Machin...

DETAILS OBSERVABILITY OS INFO SCREENSHOT

Network tags

http-server https-server

Network interfaces

Name	Network	Subnetwork	Primary internal IP address	Alias IP ranges	Stack Type	External IP address	Network
nic0	default	default	10.128.0.3		IPv4	34.134.118.89 (Ephemeral)	Premium

Storage

Boot disk

Name	Image	Interface type	Size (GB)	Device name	Type	Architecture	Encryption	Mode	W
instance-for-mlib	debian-11-bullseye-v20221102	SCSI	10	instance-for-mlib	Balanced persistent disk	x86/64	Google-managed	Boot, read/write	Del

GGCP (1).pdf GGCP.pdf GGCP.pptx Show all

Inbox (11) - manickam@student: ~ | GGCP - Google Slides | instance-for-mlib – Compute Eng | +

console.cloud.google.com/compute/instancesDetail/zones/us-central1-a/instances/instance-for-mlib?project=scala-3&pageState={"duration":("groupValue":"PT1H","customValue":null)}

Start your Free Trial with \$300 in credit. Don't worry—you won't be charged if you run out of credits. [Learn more](#)

DISMISS ACTIVATE

Google Cloud scala-3 Search for resources, docs, products, and more (/) Search

Compute Engine instance-for-... EDIT RESET CREATE MACHINE IMAGE CREATE SIMILAR OPERATIONS HELP ASSISTANT LEARN

Virtual machines VM instances Instance templates Sole-tenant nodes Machine images TPUs Committed use discounts Migrate to Virtual Machin...

Local disks None

Additional disks None

Security and access

Shielded VM

Secure Boot	Off
vTPM	On
Integrity Monitoring	On

SSH Keys

GGCP (1).pdf GGCP.pdf GGCP.pptx Show all

Inbox (11) - manickam@student: ~ | GGCP - Google Slides | instance-for-mlib – Compute Eng | +

console.cloud.google.com/compute/instancesDetail/zones/us-central1-a/instances/instance-for-mlib?project=scala-3&pageState={"duration":{"groupValue":"PT1H","customValue":null})

Start your Free Trial with \$300 in credit. Don't worry—you won't be charged if you run out of credits. [Learn more](#)

DISMISS ACTIVATE

Google Cloud scala-3 Search for resources, docs, products, and more (/) Search

Compute Engine instance-for... EDIT RESET CREATE MACHINE IMAGE CREATE SIMILAR OPERATIONS HELP ASSISTANT LEARN

Virtual machines VM instances Instance templates Sole-tenant nodes Machine images TPUs Committed use discounts Migrate to Virtual Machin...

SSH Keys

Username	Key
manickam	ecdsa-sha2-nistp256 AAAAE2VjZHNhLXNoYTItbmlzdHAyNTYAAAlbmlzdHAyNTYAAABBM7D7aDQzSWwgsp0XQAls nzWggq...
manickam	ssh-rsa AAAAB3NzaC1yc2EAAAQABAAQDuEn5Ww/mygCzHcoGZEMcdep2opgP9IThcXFChOKlrZpUuzGT4caC1BDCvi4...

Block project-wide SSH keys Off

API and identity management

Service account	179637453175-compute@developer.gserviceaccount.com
Cloud API access scopes	Allow full access to all Cloud APIs

Management

Availability policies

VM provisioning model Standard

GGCP (1).pdf GGCP.pdf GGCP.pptx Show all

Inbox (11) - manickam@student: ~ | GGCP - Google Slides | instance-for-mlib – Compute Eng | +

console.cloud.google.com/compute/instancesDetail/zones/us-central1-a/instances/instance-for-mlib?project=scala-3&pageState={"duration":("groupValue":"PT1H","customValue":null)}

Start your Free Trial with \$300 in credit. Don't worry—you won't be charged if you run out of credits. [Learn more](#)

DISMISS ACTIVATE

Google Cloud scala-3 Search for resources, docs, products, and more (/) Search

Compute Engine instance-for-... EDIT RESET CREATE MACHINE IMAGE CREATE SIMILAR OPERATIONS HELP ASSISTANT LEARN

Virtual machines VM instances Instance templates Sole-tenant nodes Machine images TPUs Committed use discounts Migrate to Virtual Machin...

DETAILS OBSERVABILITY OS INFO SCREENSHOT

Management

Availability policies

VM provisioning model	Standard
Max duration	None
Preemptibility	Off (Recommended)
On VM termination	—
On host maintenance	Migrate VM instance (Recommended)
Automatic restart	On (Recommended)
Customer Managed Encryption Key (CMEK) revocation policy	Do nothing

Sole-tenancy

Affinity labels	None
CPU Overcommit	Disabled

Show all

Inbox (11) - manickam@student: ~ | GGCP - Google Slides | instance-for-mlib – Compute Eng | +

console.cloud.google.com/compute/instancesDetail/zones/us-central1-a/instances/instance-for-mlib?project=scala-3&pageState={"duration":("groupValue":"PT1H","customValue":null)}

Start your Free Trial with \$300 in credit. Don't worry—you won't be charged if you run out of credits. [Learn more](#)

DISMISS ACTIVATE

Google Cloud scala-3 Search for resources, docs, products, and more (/) Search

Compute Engine instance-for-... EDIT RESET CREATE MACHINE IMAGE CREATE SIMILAR OPERATIONS HELP ASSISTANT LEARN

Virtual machines VM instances Instance templates Sole-tenant nodes Machine images TPUs Committed use discounts Migrate to Virtual Machin...

DETAILS OBSERVABILITY OS INFO SCREENSHOT

Availability policies

VM provisioning model	Standard
Max duration	None
Preemptibility	Off (Recommended)
On VM termination	—
On host maintenance	Migrate VM instance (Recommended)
Automatic restart	On (Recommended)
Customer Managed Encryption Key (CMEK) revocation policy	Do nothing

Sole-tenancy

Affinity labels	None
CPU Overcommit	Disabled

Custom metadata

None

EQUIVALENT REST

GGCP (1).pdf GGCP.pdf GGCP.pptx Show all

Inbox (11) - manickam@student: ~ | GGCP - Google Slides | instance-for-mlib – Compute Eng | +

console.cloud.google.com/compute/instancesDetail/zones/us-central1-a/instances/instance-for-mlib?project=scala-3&pageState={"duration":("groupValue":"PT1H","customValue":null)}

Start your Free Trial with \$300 in credit. Don't worry—you won't be charged if you run out of credits. [Learn more](#)

DISMISS ACTIVATE

Google Cloud scala-3 Search for resources, docs, products, and more (/) Search

Compute Engine instance-for-... EDIT RESET CREATE MACHINE IMAGE CREATE SIMILAR OPERATIONS HELP ASSISTANT LEARN

Virtual machines VM instances Instance templates Sole-tenant nodes Machine images TPUs Committed use discounts Migrate to Virtual Machin...

DETAILS OBSERVABILITY OS INFO SCREENSHOT

http-server https-server

Network interfaces

Name	Network	Subnetwork	Primary internal IP address	Alias IP ranges	Stack Type	External IP address	Network
nic0	default	default	10.128.0.3		IPv4	Ephemeral	Premium

Storage

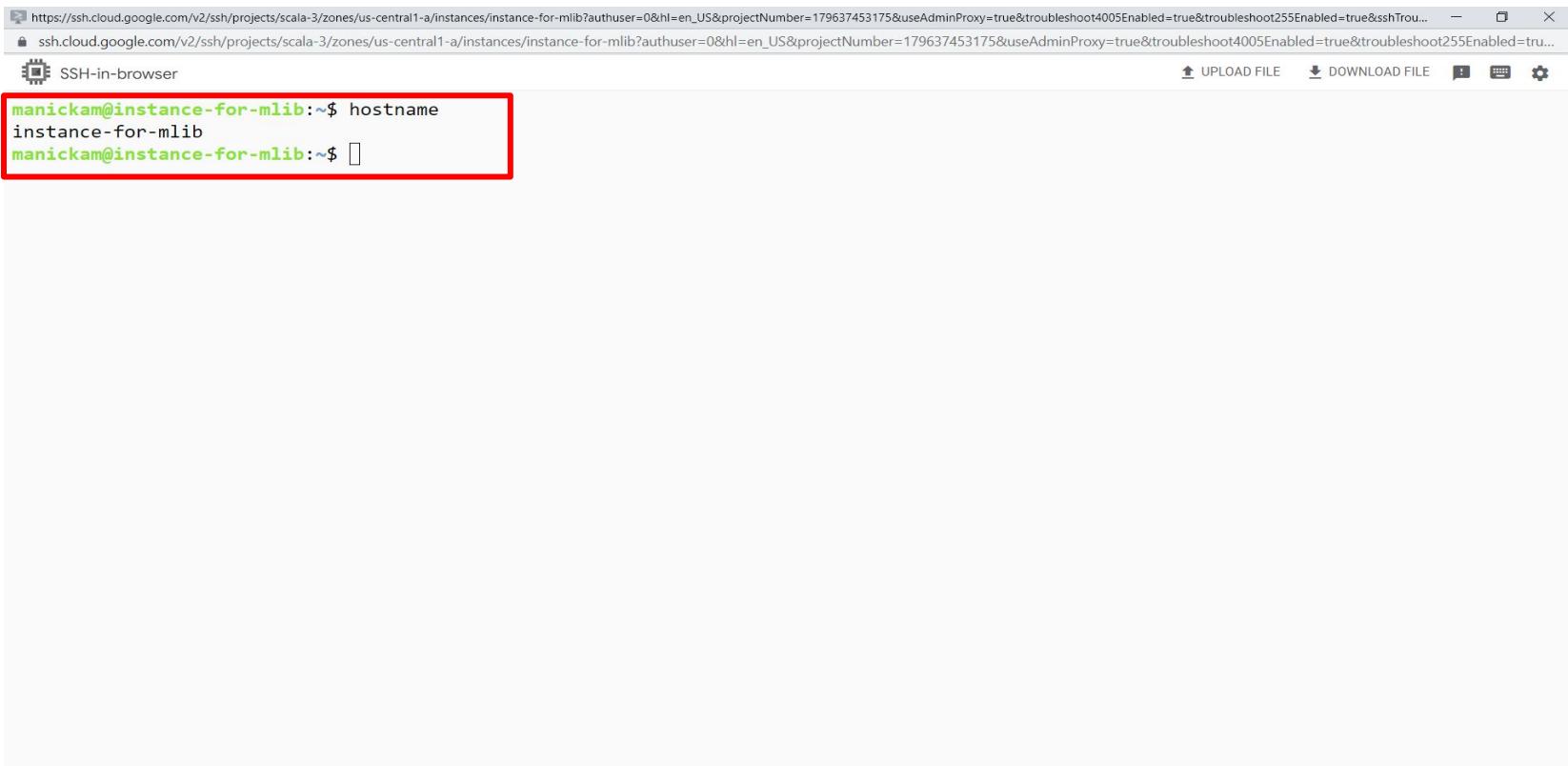
Boot disk

Name	Image	Interface type	Size (GB)	Device name	Type	Architecture	Encryption	Mode	Wb
instance-for-mlib	debian-11-bullseye-v20221102	SCSI	10	instance-for-mlib	Balanced persistent disk	x86/64	Google-managed	Boot, read/write	Del

Local disks

GGCP (1).pdf GGCP.pdf GGCP.pptx Show all

Vm instance console



The screenshot shows a browser-based SSH interface for a Google Cloud VM instance. The title bar reads "Vm instance console". The address bar shows the URL for the instance's SSH connection. The interface includes standard browser controls (back, forward, search) and a toolbar with icons for "SSH-in-browser", "UPLOAD FILE", "DOWNLOAD FILE", and settings.

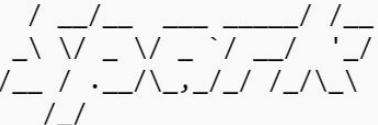
In the terminal window, the user has run the command "hostname", which outputs "instance-for-mlib". The command prompt "manickam@instance-for-mlib:~\$" is visible at the end of the output, followed by a cursor icon.

```
manickam@instance-for-mlib:~$ hostname
instance-for-mlib
manickam@instance-for-mlib:~$ 
```

Pyspark Installations

```
https://ssh.cloud.google.com/v2/ssh/projects/scala-3/zones/us-central1-a/instances/instance-for-mlib?authuser=0&hl=en_US&projectNu... ━ □ ×  
ssh.cloud.google.com/v2/ssh/projects/scala-3/zones/us-central1-a/instances/instance-for-mlib?authuser=0&hl=en_US&projectNumber...  
 SSH-in-browser  UPLOAD FILE  DOWNLOAD FILE     
manickam@instance-for-mlib:~$ sudo pip install pyspark  
Collecting pyspark  
  Downloading pyspark-3.3.1.tar.gz (281.4 MB)  
   |████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████| 281.4 MB 24 kB/s  
Collecting py4j==0.10.9.5  
  Downloading py4j-0.10.9.5-py2.py3-none-any.whl (199 kB)  
   |████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████| 199 kB 37.4 MB/s  
Building wheels for collected packages: pyspark  
  Building wheel for pyspark (setup.py) ... done  
  Created wheel for pyspark: filename=pyspark-3.3.1-py2.py3-none-any.whl size=2  
81845510 sha256=18aef36984d3d930955fb53fabe5c5a9abed7a0b39dabd670efb89435189500  
9  
  Stored in directory: /root/.cache/pip/wheels/51/c8/18/298a4ced8ebb3ab8a7d26a7  
198c0cc7035abb906bde94a4c4b  
Successfully built pyspark  
Installing collected packages: py4j, pyspark  
Successfully installed py4j-0.10.9.5 pyspark-3.3.1  
manickam@instance-for-mlib:~$
```

Spark installation checks

```
https://ssh.cloud.google.com/v2/ssh/projects/scala-3/zones/us-central1-a/instances/instance-for-mlib?authuser=0&hl=en_US&projectNu... ━ ━ X
ssh.cloud.google.com/v2/ssh/projects/scala-3/zones/us-central1-a/instances/instance-for-mlib?authuser=0&hl=en_US&projectNumber...
SSH-in-browser  UPLOAD FILE DOWNLOAD FILE   
  
version 3.3.1  
  
Using Python version 3.9.2 (default, Feb 28 2021 17:03:44)  
Spark context Web UI available at http://instance-for-mlib.us-central1-a.c.scala-3.internal:4040  
Spark context available as 'sc' (master = local[*], app id = local-166881044979).  
SparkSession available as 'spark'.  
>>> # Import SparkSession  
>>> from pyspark.sql import SparkSession  
>>> # Create a Spark Session  
>>> spark = SparkSession.builder.master("local[*]").getOrCreate()  
>>> # Check Spark Session Information  
>>> spark  
<pyspark.sql.session.SparkSession object at 0x7f14f18952b0>  
>>> # Import a Spark function from library  
>>> from pyspark.sql.functions import col  
>>> █
```

Routings.show()

Gresults-1-reference.prn - Notepad

File Edit Format View Help

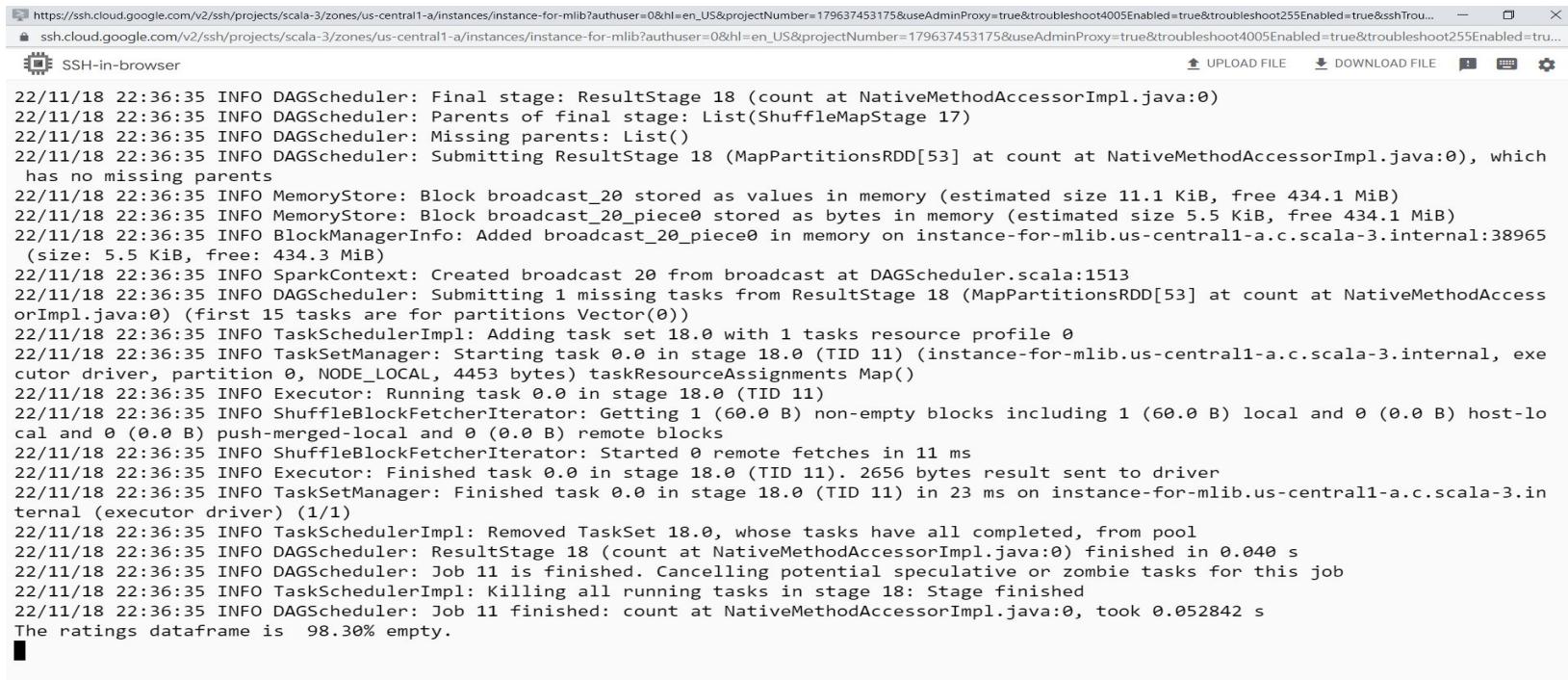
```
22/11/18 22:57:49 INFO TaskSchedulerImpl: Killing all running tasks in stage 2: Stage finished
22/11/18 22:57:49 INFO DAGScheduler: Job 2 finished: showString at NativeMethodAccessorImpl.java:0, took
0.281905 s
22/11/18 22:57:49 INFO CodeGenerator: Code generated in 41.783209 ms
+-----+-----+-----+
|userId|moviId|rating|timestamp|
+-----+-----+-----+
| 1| 1| 4.0|964982703|
| 1| 3| 4.0|964981247|
| 1| 6| 4.0|964982224|
| 1| 47| 5.0|964983815|
| 1| 50| 5.0|964982931|
| 1| 70| 3.0|964982400|
| 1| 101| 5.0|964980868|
| 1| 110| 4.0|964982176|
| 1| 151| 5.0|964984041|
| 1| 157| 5.0|964984100|
| 1| 163| 5.0|964983650|
| 1| 216| 5.0|964981208|
| 1| 223| 3.0|964980985|
| 1| 231| 5.0|964981179|
| 1| 235| 4.0|964980908|
| 1| 260| 5.0|964981680|
| 1| 296| 3.0|964982967|
| 1| 316| 3.0|964982310|
| 1| 333| 5.0|964981179|
| 1| 349| 4.0|964982563|
+-----+-----+-----+
only showing top 20 rows
```

Ln 148, Col 1 100% Unix (LF) UTF-8

Ratings.show()

```
https://ssh.cloud.google.com/v2/ssh/projects/scala-3/zones/us-central1-a/instances/instance-for-mllib?authuser=0&hl=en_US&projectNumber=179637453175&useAdminProxy=true&troubleshoot4005Enabled=true&troubleshoot255Enabled=true&sshT... ━ ━ X
ssh.cloud.google.com/v2/ssh/projects/scala-3/zones/us-central1-a/instances/instance-for-mllib?authuser=0&hl=en_US&projectNumber=179637453175&useAdminProxy=true&troubleshoot4005Enabled=true&troubleshoot255Enabled=true...
SSH-in-browser UPLOAD FILE DOWNLOAD FILE ⚡ ⚡ ⚡ ⚡ ⚡
22/11/18 22:36:01 INFO DAGScheduler: Job 3 finished: showString at NativeMethodAccessorImpl.java:0, took 0.207684 s
22/11/18 22:36:01 INFO CodeGenerator: Code generated in 26.098897 ms
+-----+
|userId|movieId|rating|
+-----+
| 1 | 1 | 4.0 |
| 1 | 3 | 4.0 |
| 1 | 6 | 4.0 |
| 1 | 47 | 5.0 |
| 1 | 50 | 5.0 |
| 1 | 70 | 3.0 |
| 1 | 101 | 5.0 |
| 1 | 110 | 4.0 |
| 1 | 151 | 5.0 |
| 1 | 157 | 5.0 |
| 1 | 163 | 5.0 |
| 1 | 216 | 5.0 |
| 1 | 223 | 3.0 |
| 1 | 231 | 5.0 |
| 1 | 235 | 4.0 |
| 1 | 260 | 5.0 |
| 1 | 296 | 3.0 |
| 1 | 316 | 3.0 |
| 1 | 333 | 5.0 |
| 1 | 349 | 4.0 |
+-----+
only showing top 20 rows
```

Calculate sparsity



The screenshot shows an SSH session in a browser window titled "SSH-in-browser". The URL is https://ssh.cloud.google.com/v2/ssh/projects/scala-3/zones/us-central1-a/instances/instance-for-mllib?authuser=0&hl=en_US&projectNumber=179637453175&useAdminProxy=true&troubleshoot4005Enabled=true&troubleshoot255Enabled=true&sshT...". The session is connected to an instance with IP 10.128.0.3. The terminal window displays a log of Java application output. The log details the execution of a DAGScheduler, including the creation of broadcast variables, the submission of tasks, and the completion of the job. Key messages include:

```
22/11/18 22:36:35 INFO DAGScheduler: Final stage: ResultStage 18 (count at NativeMethodAccessorImpl.java:0)
22/11/18 22:36:35 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 17)
22/11/18 22:36:35 INFO DAGScheduler: Missing parents: List()
22/11/18 22:36:35 INFO DAGScheduler: Submitting ResultStage 18 (MapPartitionsRDD[53] at count at NativeMethodAccessorImpl.java:0), which
has no missing parents
22/11/18 22:36:35 INFO MemoryStore: Block broadcast_20 stored as values in memory (estimated size 11.1 KiB, free 434.1 MiB)
22/11/18 22:36:35 INFO MemoryStore: Block broadcast_20_piece0 stored as bytes in memory (estimated size 5.5 KiB, free 434.1 MiB)
22/11/18 22:36:35 INFO BlockManagerInfo: Added broadcast_20_piece0 in memory on instance-for-mllib.us-central1-a.c.scala-3.internal:38965
(size: 5.5 KiB, free: 434.3 MiB)
22/11/18 22:36:35 INFO SparkContext: Created broadcast 20 from broadcast at DAGScheduler.scala:1513
22/11/18 22:36:35 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 18 (MapPartitionsRDD[53] at count at NativeMethodAccess
orImpl.java:0) (first 15 tasks are for partitions Vector(0))
22/11/18 22:36:35 INFO TaskSchedulerImpl: Adding task set 18.0 with 1 tasks resource profile 0
22/11/18 22:36:35 INFO TaskSetManager: Starting task 0.0 in stage 18.0 (TID 11) (instance-for-mllib.us-central1-a.c.scala-3.internal, exe
cutor driver, partition 0, NODE_LOCAL, 4453 bytes) taskResourceAssignments Map()
22/11/18 22:36:35 INFO Executor: Running task 0.0 in stage 18.0 (TID 11)
22/11/18 22:36:35 INFO ShuffleBlockFetcherIterator: Getting 1 (60.0 B) non-empty blocks including 1 (60.0 B) local and 0 (0.0 B) host-lo
cal and 0 (0.0 B) push-merged-local and 0 (0.0 B) remote blocks
22/11/18 22:36:35 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 11 ms
22/11/18 22:36:35 INFO Executor: Finished task 0.0 in stage 18.0 (TID 11). 2656 bytes result sent to driver
22/11/18 22:36:35 INFO TaskSetManager: Finished task 0.0 in stage 18.0 (TID 11) in 23 ms on instance-for-mllib.us-central1-a.c.scala-3.in
ternal (executor driver) (1/1)
22/11/18 22:36:35 INFO TaskSchedulerImpl: Removed TaskSet 18.0, whose tasks have all completed, from pool
22/11/18 22:36:35 INFO DAGScheduler: ResultStage 18 (count at NativeMethodAccessorImpl.java:0) finished in 0.040 s
22/11/18 22:36:35 INFO DAGScheduler: Job 11 is finished. Cancelling potential speculative or zombie tasks for this job
22/11/18 22:36:35 INFO TaskSchedulerImpl: Killing all running tasks in stage 18: Stage finished
22/11/18 22:36:35 INFO DAGScheduler: Job 11 finished: count at NativeMethodAccessorImpl.java:0, took 0.052842 s
The ratings dataframe is 98.30% empty.
```

Interpret ratings

https://ssh.cloud.google.com/v2/ssh/projects/scala-3/zones/us-central1-a/instances/instance-for-mllib?authuser=0&hl=en_US&projectNumber=179637453175&useAdminProxy=true&troubleshoot4005Enabled=true&troubleshoot255Enabled=true&sshTrou... X

ssh.cloud.google.com/v2/ssh/projects/scala-3/zones/us-central1-a/instances/instance-for-mllib?authuser=0&hl=en_US&projectNumber=179637453175&useAdminProxy=true&troubleshoot4005Enabled=true&troubleshoot255Enabled=true

 SSH-in-browser UPLOAD FILE DOWNLOAD FILE ! 📊⚙️

userId	count
414	2698
599	2478
474	2108
448	1864
274	1346
610	1302
68	1260
380	1218
606	1115
288	1055
249	1046
387	1027
182	977
307	975
603	943
298	939
177	904
318	879
232	862
480	836

only showing top 20 rows

```
22/11/18 22:37:07 INFO BlockManagerInfo: Removed broadcast_17_piece0 on instance-for-mllib.us-central1-a.c.scala-3.internal:38965 in memory (size: 34.0 KiB, free: 434.3 MiB)
```

moviel_ ratings.show()

```
https://ssh.cloud.google.com/v2/ssh/projects/scala-3/zones/us-central1-a/instances/instance-for-mllib?authuser=0&hl=en_US&projectNumber=179637453175&useAdminProxy=true&troubleshoot4005Enabled=true&troubleshoot255Enabled=true&sshT... ━ ━ ━
ssh.cloud.google.com/v2/ssh/projects/scala-3/zones/us-central1-a/instances/instance-for-mllib?authuser=0&hl=en_US&projectNumber=179637453175&useAdminProxy=true&troubleshoot4005Enabled=true&troubleshoot255Enabled=true... ━ ━ ━
SSH-in-browser ━ ━ ━
↑ UPLOAD FILE ↓ DOWNLOAD FILE ━ ━ ━
+-----+-----+
| 356 | 329 |
| 318 | 317 |
| 296 | 307 |
| 593 | 279 |
| 2571| 278 |
| 260 | 251 |
| 480 | 238 |
| 110 | 237 |
| 589 | 224 |
| 527 | 220 |
| 2959| 218 |
| 1   | 215 |
| 1196| 211 |
| 2858| 204 |
| 50  | 204 |
| 47  | 203 |
| 780 | 202 |
| 150 | 201 |
| 1198| 200 |
| 4993| 198 |
+-----+
only showing top 20 rows

22/11/18 22:37:38 INFO BlockManagerInfo: Removed broadcast_21_piece0 on instance-for-mllib.us-central1-a.c.scala-3.internal:38965 in memory (size: 34.0 KiB, free: 434.3 MiB)
22/11/18 22:37:38 INFO BlockManagerInfo: Removed broadcast_25_piece0 on instance-for-mllib.us-central1-a.c.scala-3.internal:38965 in memory (size: 17.6 KiB, free: 434.3 MiB)
```

Build Out An ALS Model

Gresults-1-reference.prn - Notepad

File Edit Format View Help

```
a.c.scala-3.internal:46767 (size: 203.5 KiB, free: 423.2 MiB)
22/11/19 02:00:56 INFO Executor: Finished task 9.0 in stage 8889.0 (TID 21820). 2236 bytes result sent to driver
22/11/19 02:00:56 INFO TaskSetManager: Finished task 9.0 in stage 8889.0 (TID 21820) in 494 ms on instance-for-mllib.us-central1-a.c.scala-3.internal (executor driver) (10/10)
22/11/19 02:00:56 INFO TaskSchedulerImpl: Removed TaskSet 8889.0, whose tasks have all completed, from pool
22/11/19 02:00:56 INFO DAGScheduler: ResultStage 8889 (count at ALS.scala:1080) finished in 2.971 s
22/11/19 02:00:56 INFO DAGScheduler: Job 945 is finished. Cancelling potential speculative or zombie tasks for this job
22/11/19 02:00:56 INFO TaskSchedulerImpl: Killing all running tasks in stage 8889: Stage finished
22/11/19 02:00:56 INFO DAGScheduler: Job 945 finished: count at ALS.scala:1080, took 2.977674 s
22/11/19 02:00:56 INFO MapPartitionsRDD: Removing RDD 19202 from persistence list
22/11/19 02:00:56 INFO BlockManager: Removing RDD 19202
22/11/19 02:00:56 INFO MapPartitionsRDD: Removing RDD 19028 from persistence list
22/11/19 02:00:56 INFO BlockManager: Removing RDD 19028
22/11/19 02:00:56 INFO Instrumentation: [13ed0081] training finished
<class 'pyspark.ml.recommendation.ALSModel'>
**Best Model**
Rank: 50
MaxIter: 10
RegParam: 0.15
22/11/19 02:03:56 INFO FileSourceStrategy: Pushed Filters:
22/11/19 02:03:56 INFO FileSourceStrategy: Post-Scan Filters:
22/11/19 02:03:56 INFO FileSourceStrategy: Output Data Schema: struct<userId: string, movieId: string, rating: string ... 1 more fields>
22/11/19 02:03:56 INFO CodeGenerator: Code generated in 21.203661 ms
22/11/19 02:03:56 INFO MemoryStore: Block broadcast_2992 stored as values in memory (estimated size 199.1 KiB, free 420.9 MiB)
22/11/19 02:03:56 INFO MemoryStore: Block broadcast_2992_piece0 stored as bytes in memory (estimated size 34 0 KiB, free 420.9 MiB)
```

Tell Spark how to tune your ALS model

Build your cross validation pipeline

```
https://ssh.cloud.google.com/v2/ssh/projects/scala-3/zones/us-central1-a/instances/instance-for-mlib?authuser=0&hl=en_US&projectNumber=179637453175&useAdminProxy=true&troubleshoot4005Enabled=true&troubleshoot255Enabled=true&sshTrous... — □ ×
ssh.cloud.google.com/v2/ssh/projects/scala-3/zones/us-central1-a/instances/instance-for-mlib?authuser=0&hl=en_US&projectNumber=179637453175&useAdminProxy=true&troubleshoot4005Enabled=true&troubleshoot255Enabled=true...
SSH-in-browser UPLOAD FILE DOWNLOAD FILE ! ⌨ ⚙

```
296 | 307 |
593 | 279 |
2571| 278 |
260 | 251 |
480 | 238 |
110 | 237 |
589 | 224 |
527 | 220 |
2959| 218 |
1 | 215 |
1196| 211 |
2858| 204 |
50 | 204 |
47 | 203 |
780 | 202 |
150 | 201 |
1198| 200 |
4993| 198 |
+----+----+
only showing top 20 rows

22/11/18 22:37:38 INFO BlockManagerInfo: Removed broadcast_21_piece0 on instance-for-mlib.us-central1-a.c.scala-3.internal:38965 in memory (size: 34.0 KiB, free: 434.3 MiB)
22/11/18 22:37:38 INFO BlockManagerInfo: Removed broadcast_25_piece0 on instance-for-mlib.us-central1-a.c.scala-3.internal:38965 in memory (size: 17.6 KiB, free: 434.3 MiB)
Num models to be tested: 16
CrossValidator_d9b0b0e10a06
Until Cross Validation before fit model
```


```

Output Of Python Program on GCP

```
https://ssh.cloud.google.com/v2/ssh/projects/scala-3/zones/us-central1-a/instances/instance-for-mllib?authuser=0&hl=en_US&projectNumber=179637453175&useAdminProxy=true&troubleshoot4005Enabled=true&troubleshoot255Enabled=true&sshT... — X
ssh.cloud.google.com/v2/ssh/projects/scala-3/zones/us-central1-a/instances/instance-for-mllib?authuser=0&hl=en_US&projectNumber=179637453175&useAdminProxy=true&troubleshoot4005Enabled=true&troubleshoot255Enabled=true...
SSH-in-browser
 UPLOAD FILE DOWNLOAD FILE ⚡ 📋 🛠️ 🛡️
1|      3| 4.0|964981247|
1|      6| 4.0|964982224|
1|     47| 5.0|964983815|
1|     50| 5.0|964982931|
1|     70| 3.0|964982400|
1|    101| 5.0|964980868|
1|    110| 4.0|964982176|
1|    151| 5.0|964984041|
1|    157| 5.0|964984100|
1|    163| 5.0|964983650|
1|    216| 5.0|964981208|
1|    223| 3.0|964980985|
1|    231| 5.0|964981179|
1|    235| 4.0|964980908|
1|    260| 5.0|964981680|
1|    296| 3.0|964982967|
1|    316| 3.0|964982310|
1|    333| 5.0|964981179|
1|    349| 4.0|964982563|
+---+---+---+---+
only showing top 20 rows

root
|-- userId: string (nullable = true)
|-- movieId: string (nullable = true)
|-- rating: string (nullable = true)
|-- timestamp: string (nullable = true)
```

```
22/11/18 22:58:20 INFO DAGScheduler: Job 3 finished: showString at NativeMethodAccessorImpl.java:0, took 0.197561 s
22/11/18 22:58:20 INFO CodeGenerator: Code generated in 19.586172 ms
```

userId	movieId	rating
1	1	4.0
1	3	4.0
1	6	4.0
1	47	5.0
1	50	5.0
1	70	3.0
1	101	5.0
1	110	4.0
1	151	5.0
1	157	5.0
1	163	5.0
1	216	5.0
1	223	3.0
1	231	5.0
1	235	4.0
1	260	5.0
1	296	3.0
1	316	3.0
1	333	5.0
1	349	4.0

only showing top 20 rows

22/11/18 22:58:20 INFO DAGScheduler: Job 3 finished: showString at NativeMethodAccessorImpl.java:0, took 0.197561 s

22/11/18 22:58:20 INFO CodeGenerator: Code generated in 19.586172 ms

+-----+-----+

|userId|movieId|rating|

+-----+-----+

1	1	4.0
1	3	4.0
1	6	4.0
1	47	5.0
1	50	5.0
1	70	3.0
1	101	5.0
1	110	4.0
1	151	5.0
1	157	5.0
1	163	5.0
1	216	5.0
1	223	3.0
1	231	5.0
1	235	4.0
1	260	5.0
1	296	3.0
1	316	3.0
1	333	5.0
1	349	4.0

+-----+-----+

only showing top 20 rows

https://ssh.cloud.google.com/v2/ssh/projects/scala-3/zones/us-central1-a/instances/instance-for-mlib?authuser=0&hl=en_US&projectNumber=179637453175&useAdminProxy=true&troubleshoot4005Enabled=true&troubleshoot255Enabled=true&sshT... — X

SSH-in-browser

UPLOAD FILE DOWNLOAD FILE

```
22/11/18 22:58:55 INFO DAGScheduler: Final stage: ResultStage 18 (count at NativeMethodAccessorImpl.java:0)
22/11/18 22:58:55 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 17)
22/11/18 22:58:55 INFO DAGScheduler: Missing parents: List()
22/11/18 22:58:55 INFO DAGScheduler: Submitting ResultStage 18 (MapPartitionsRDD[53] at count at NativeMethodAccessorImpl.java:0), which has no missing parents
22/11/18 22:58:55 INFO MemoryStore: Block broadcast_20 stored as values in memory (estimated size 11.1 KiB, free 434.1 MiB)
22/11/18 22:58:55 INFO MemoryStore: Block broadcast_20_piece0 stored as bytes in memory (estimated size 5.5 KiB, free 434.1 MiB)
22/11/18 22:58:55 INFO BlockManagerInfo: Added broadcast_20_piece0 in memory on instance-for-mlib.us-central1-a.c.scala-3.internal:46767 (size: 5.5 KiB, free: 434.3 MiB)
22/11/18 22:58:55 INFO SparkContext: Created broadcast 20 from broadcast at DAGScheduler.scala:1513
22/11/18 22:58:55 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 18 (MapPartitionsRDD[53] at count at NativeMethodAccess orImpl.java:0) (first 15 tasks are for partitions Vector(0))
22/11/18 22:58:55 INFO TaskSchedulerImpl: Adding task set 18.0 with 1 tasks resource profile 0
22/11/18 22:58:55 INFO TaskSetManager: Starting task 0.0 in stage 18.0 (TID 11) (instance-for-mlib.us-central1-a.c.scala-3.internal, executor driver, partition 0, NODE_LOCAL, 4453 bytes) taskResourceAssignments Map()
22/11/18 22:58:55 INFO Executor: Running task 0.0 in stage 18.0 (TID 11)
22/11/18 22:58:55 INFO ShuffleBlockFetcherIterator: Getting 1 (60.0 B) non-empty blocks including 1 (60.0 B) local and 0 (0.0 B) host-local and 0 (0.0 B) push-merged-local and 0 (0.0 B) remote blocks
22/11/18 22:58:55 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 1 ms
22/11/18 22:58:55 INFO Executor: Finished task 0.0 in stage 18.0 (TID 11). 2656 bytes result sent to driver
22/11/18 22:58:55 INFO TaskSetManager: Finished task 0.0 in stage 18.0 (TID 11) in 13 ms on instance-for-mlib.us-central1-a.c.scala-3.in ternal (executor driver) (1/1)
22/11/18 22:58:55 INFO TaskSchedulerImpl: Removed TaskSet 18.0, whose tasks have all completed, from pool
22/11/18 22:58:55 INFO DAGScheduler: ResultStage 18 (count at NativeMethodAccessorImpl.java:0) finished in 0.024 s
22/11/18 22:58:55 INFO DAGScheduler: Job 11 is finished. Cancelling potential speculative or zombie tasks for this job
22/11/18 22:58:55 INFO TaskSchedulerImpl: Killing all running tasks in stage 18: Stage finished
22/11/18 22:58:55 INFO DAGScheduler: Job 11 finished: count at NativeMethodAccessorImpl.java:0, took 0.035534 s
The ratings dataframe is 98.30% empty.
```

```
22/11/18 22:59:27 INFO CodeGenerator: Code generated in 28.59071 ms
22/11/18 22:59:27 INFO CodeGenerator: Code generated in 16.680332 ms
```

userId	count
414	2698
599	2478
474	2108
448	1864
274	1346
610	1302
68	1260
380	1218
606	1115
288	1055
249	1046
387	1027
182	977
307	975
603	943
298	939
177	904
318	879
232	862
480	836

only showing top 20 rows

SSH-in-browser

UPLOAD FILE DOWNLOAD FILE ! 📈 🗃

movieId	count
356	329
318	317
296	307
593	279
2571	278
260	251
480	238
110	237
589	224
527	220
2959	218
1	215
1196	211
2858	204
50	204
47	203
780	202
150	201
1198	200
4993	198

only showing top 20 rows

22/11/18 22:59:58 INFO BlockManagerInfo: Removed broadcast_25_piece0 on instance-for-mlib.us-central1-a.c.scala-3.internal:46767 in memory (size: 17.6 KiB, free: 434.3 MiB)

356	329
318	317
296	307
593	279
2571	278
260	251
480	238
110	237
589	224
527	220
2959	218
1	215
1196	211
2858	204
50	204
47	203
780	202
150	201
1198	200
4993	198

+-----+

only showing top 20 rows

22/11/18 22:59:58 INFO BlockManagerInfo: Removed broadcast_25_piece0 on instance-for-mlib.us-central1-a.c.scala-3.internal:46767 in memory (size: 17.6 KiB, free: 434.3 MiB)

22/11/18 23:00:28 INFO BlockManagerInfo: Removed broadcast_26_piece0 on instance-for-mlib.us-central1-a.c.scala-3.internal:46767 in memory (size: 18.6 KiB, free: 434.4 MiB)

Confirm that a model called als was created

```
593| 279|
2571| 278|
260| 251|
480| 238|
110| 237|
589| 224|
527| 220|
2959| 218|
1| 215|
1196| 211|
2858| 204|
50| 204|
47| 203|
780| 202|
150| 201|
1198| 200|
4993| 198|
+-----+-----+
only showing top 20 rows
```

```
22/11/18 22:59:58 INFO BlockManagerInfo: Removed broadcast_25_piece0 on instance-for-mlib.us-central1-a.c.scala-3.internal:46767 in memory (size: 17.6 KiB, free: 434.3 MiB)
22/11/18 23:00:28 INFO BlockManagerInfo: Removed broadcast_26_piece0 on instance-for-mlib.us-central1-a.c.scala-3.internal:46767 in memory (size: 18.6 KiB, free: 434.4 MiB)
```

Confirm that a model called als was created

Num models to be tested: 16

CrossVidator

CrossValidator_259aeee816f1

Cross Validator

```
https://ssh.cloud.google.com/v2/ssh/projects/scala-3/zones/us-central1-a/instances/instance-for-mllib?authuser=0&hl=en_US&projectNumber=179637453175&useAdminProxy=true&troubleshoot4005Enabled=true&troubleshoot255Enabled=true&sshTou... ━ ━ ×
ssh.cloud.google.com/v2/ssh/projects/scala-3/zones/us-central1-a/instances/instance-for-mllib?authuser=0&hl=en_US&projectNumber=179637453175&useAdminProxy=true&troubleshoot4005Enabled=true&troubleshoot255Enabled=true...
SSH-in-browser ━ UPLOAD FILE ━ DOWNLOAD FILE ━ ━ ━
2571| 278|
 260| 251|
 480| 238|
 110| 237|
 589| 224|
 527| 220|
2959| 218|
  1| 215|
1196| 211|
2858| 204|
   50| 204|
    47| 203|
 780| 202|
 150| 201|
1198| 200|
 4993| 198|
+-----+
only showing top 20 rows

22/11/18 22:59:58 INFO BlockManagerInfo: Removed broadcast_25_piece0 on instance-for-mllib.us-central1-a.c.scala-3.internal:46767 in memory (size: 17.6 KiB, free: 434.3 MiB)
22/11/18 23:00:28 INFO BlockManagerInfo: Removed broadcast_26_piece0 on instance-for-mllib.us-central1-a.c.scala-3.internal:46767 in memory (size: 18.6 KiB, free: 434.4 MiB)
Confirm that a model called als was created
Num models to be tested: 16
CrossValidator
CrossValidator_259aeee816f1
Until Cross Validation before fit model
```

```
22/11/19 01:09:43 INFO Executor: Running task 3.0 in stage 6631.0 (TID 16221)
22/11/19 01:09:43 INFO BlockManager: Found block rdd_14292_3 locally
22/11/19 01:09:43 INFO BlockManager: Found block rdd_14291_3 locally
22/11/19 01:09:43 INFO ShuffleBlockFetcherIterator: Getting 10 (359.5 KiB) non-empty blocks including 10 (359.5 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-merged-local and 0 (0.0 B) remote blocks
22/11/19 01:09:43 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
22/11/19 01:09:46 INFO Executor: Finished task 2.0 in stage 6631.0 (TID 16220). 2694 bytes result sent to driver
22/11/19 01:09:46 INFO TaskSetManager: Starting task 4.0 in stage 6631.0 (TID 16222) (instance-for-mlib.us-central1-a.c.scala-3.internal, executor driver, partition 4, PROCESS_LOCAL, 4550 bytes) taskResourceAssignments Map()
22/11/19 01:09:46 INFO TaskSetManager: Finished task 2.0 in stage 6631.0 (TID 16220) in 4271 ms on instance-for-mlib.us-central1-a.c.scala-3.internal (executor driver) (3/10)
22/11/19 01:09:46 INFO Executor: Running task 4.0 in stage 6631.0 (TID 16222)
22/11/19 01:09:46 INFO BlockManager: Found block rdd_14292_4 locally
22/11/19 01:09:46 INFO BlockManager: Found block rdd_14291_4 locally
22/11/19 01:09:46 INFO ShuffleBlockFetcherIterator: Getting 10 (349.7 KiB) non-empty blocks including 10 (349.7 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-merged-local and 0 (0.0 B) remote blocks
22/11/19 01:09:46 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
22/11/19 01:09:47 INFO Executor: Finished task 3.0 in stage 6631.0 (TID 16221). 2694 bytes result sent to driver
22/11/19 01:09:47 INFO TaskSetManager: Starting task 5.0 in stage 6631.0 (TID 16223) (instance-for-mlib.us-central1-a.c.scala-3.internal, executor driver, partition 5, PROCESS_LOCAL, 4550 bytes) taskResourceAssignments Map()
22/11/19 01:09:47 INFO Executor: Running task 5.0 in stage 6631.0 (TID 16223)
22/11/19 01:09:47 INFO BlockManager: Found block rdd_14292_5 locally
22/11/19 01:09:47 INFO BlockManager: Found block rdd_14291_5 locally
22/11/19 01:09:47 INFO ShuffleBlockFetcherIterator: Getting 10 (352.5 KiB) non-empty blocks including 10 (352.5 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-merged-local and 0 (0.0 B) remote blocks
22/11/19 01:09:47 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
22/11/19 01:09:47 INFO TaskSetManager: Finished task 3.0 in stage 6631.0 (TID 16221) in 3839 ms on instance-for-mlib.us-central1-a.c.scala-3.internal (executor driver) (4/10)
```

https://ssh.cloud.google.com/v2/ssh/projects/scala-3/zones/us-central1-a/instances/instance-for-mlib?authuser=0&hl=en_US&projectNumber=179637453175&useAdminProxy=true&troubleshoot4005Enabled=true&troubleshoot255Enabled=true&sshT... — □ ×

ssh.cloud.google.com/v2/ssh/projects/scala-3/zones/us-central1-a/instances/instance-for-mlib?authuser=0&hl=en_US&projectNumber=179637453175&useAdminProxy=true&troubleshoot4005Enabled=true&troubleshoot255Enabled=true

 SSH-in-browser  UPLOAD FILE  DOWNLOAD FILE   

```
22/11/19 01:10:41 INFO Executor: Running task 3.0 in stage 6634.0 (TID 16251)
22/11/19 01:10:41 INFO BlockManager: Found block rdd_14287_3 locally
22/11/19 01:10:41 INFO BlockManager: Found block rdd_14286_3 locally
22/11/19 01:10:41 INFO ShuffleBlockFetcherIterator: Getting 10 (1289.4 KiB) non-empty blocks including 10 (1289.4 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-merged-local and 0 (0.0 B) remote blocks
22/11/19 01:10:41 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
22/11/19 01:10:45 INFO Executor: Finished task 2.0 in stage 6634.0 (TID 16250). 2694 bytes result sent to driver
22/11/19 01:10:45 INFO TaskSetManager: Starting task 4.0 in stage 6634.0 (TID 16252) (instance-for-mlib.us-central1-a.c.scala-3.internal, executor driver, partition 4, PROCESS_LOCAL, 4550 bytes) taskResourceAssignments Map()
22/11/19 01:10:45 INFO TaskSetManager: Finished task 2.0 in stage 6634.0 (TID 16250) in 4039 ms on instance-for-mlib.us-central1-a.c.scala-3.internal (executor driver) (3/10)
22/11/19 01:10:45 INFO Executor: Running task 4.0 in stage 6634.0 (TID 16252)
22/11/19 01:10:45 INFO BlockManager: Found block rdd_14287_4 locally
22/11/19 01:10:45 INFO BlockManager: Found block rdd_14286_4 locally
22/11/19 01:10:45 INFO ShuffleBlockFetcherIterator: Getting 10 (2.1 MiB) non-empty blocks including 10 (2.1 MiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-merged-local and 0 (0.0 B) remote blocks
22/11/19 01:10:45 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
22/11/19 01:10:46 INFO Executor: Finished task 3.0 in stage 6634.0 (TID 16251). 2694 bytes result sent to driver
22/11/19 01:10:46 INFO TaskSetManager: Starting task 5.0 in stage 6634.0 (TID 16253) (instance-for-mlib.us-central1-a.c.scala-3.internal, executor driver, partition 5, PROCESS_LOCAL, 4550 bytes) taskResourceAssignments Map()
22/11/19 01:10:46 INFO TaskSetManager: Finished task 3.0 in stage 6634.0 (TID 16251) in 4695 ms on instance-for-mlib.us-central1-a.c.scala-3.internal (executor driver) (4/10)
22/11/19 01:10:46 INFO Executor: Running task 5.0 in stage 6634.0 (TID 16253)
22/11/19 01:10:46 INFO BlockManager: Found block rdd_14287_5 locally
22/11/19 01:10:46 INFO BlockManager: Found block rdd_14286_5 locally
22/11/19 01:10:46 INFO ShuffleBlockFetcherIterator: Getting 10 (1379.8 KiB) non-empty blocks including 10 (1379.8 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-merged-local and 0 (0.0 B) remote blocks
22/11/19 01:10:46 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
```

Best Model , Rank, MaxIter, RegParam

```
https://ssh.cloud.google.com/v2/ssh/projects/scala-3/zones/us-central1-a/instances/instance-for-mllib?authuser=0&hl=en_US&projectNumber=179637453175&useAdminProxy=true&troubleshoot4005Enabled=true&troubleshoot255Enabled=true&sshT... — X
ssh.cloud.google.com/v2/ssh/projects/scala-3/zones/us-central1-a/instances/instance-for-mllib?authuser=0&hl=en_US&projectNumber=179637453175&useAdminProxy=true&troubleshoot4005Enabled=true&troubleshoot255Enabled=true...
SSH-in-browser UPLOAD FILE DOWNLOAD FILE ⚙️
host-local and 0 (0.0 B) push-merged-local and 0 (0.0 B) remote blocks
22/11/19 02:00:55 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
22/11/19 02:00:55 INFO MemoryStore: Block rdd_19221_8 stored as values in memory (estimated size 203.5 KiB, free 420.7 MiB)
22/11/19 02:00:55 INFO BlockManagerInfo: Added rdd_19221_8 in memory on instance-for-mllib.us-central1-a.c.scala-3.internal:46767 (size: 203.5 KiB, free: 423.4 MiB)
22/11/19 02:00:55 INFO Executor: Finished task 8.0 in stage 8889.0 (TID 21819). 2236 bytes result sent to driver
22/11/19 02:00:55 INFO TaskSetManager: Finished task 8.0 in stage 8889.0 (TID 21819) in 303 ms on instance-for-mllib.us-central1-a.c.scala-3.internal (executor driver) (9/10)
22/11/19 02:00:56 INFO MemoryStore: Block rdd_19221_9 stored as values in memory (estimated size 203.5 KiB, free 420.5 MiB)
22/11/19 02:00:56 INFO BlockManagerInfo: Added rdd_19221_9 in memory on instance-for-mllib.us-central1-a.c.scala-3.internal:46767 (size: 203.5 KiB, free: 423.2 MiB)
22/11/19 02:00:56 INFO Executor: Finished task 9.0 in stage 8889.0 (TID 21820). 2236 bytes result sent to driver
22/11/19 02:00:56 INFO TaskSetManager: Finished task 9.0 in stage 8889.0 (TID 21820) in 494 ms on instance-for-mllib.us-central1-a.c.scala-3.internal (executor driver) (10/10)
22/11/19 02:00:56 INFO TaskSchedulerImpl: Removed TaskSet 8889.0, whose tasks have all completed, from pool
22/11/19 02:00:56 INFO DAGScheduler: ResultStage 8889 (count at ALS.scala:1080) finished in 2.971 s
22/11/19 02:00:56 INFO DAGScheduler: Job 945 is finished. Cancelling potential speculative or zombie tasks for this job
22/11/19 02:00:56 INFO TaskSchedulerImpl: Killing all running tasks in stage 8889: Stage finished
22/11/19 02:00:56 INFO DAGScheduler: Job 945 finished: count at ALS.scala:1080, took 2.977674 s
22/11/19 02:00:56 INFO MapPartitionsRDD: Removing RDD 19202 from persistence list
22/11/19 02:00:56 INFO BlockManager: Removing RDD 19202
22/11/19 02:00:56 INFO MapPartitionsRDD: Removing RDD 19028 from persistence list
22/11/19 02:00:56 INFO BlockManager: Removing RDD 19028
22/11/19 02:00:56 INFO Instrumentation: [13ed0081] training finished
<class 'pyspark.ml.recommendation.ALSModel'>
**Best Model**
Rank: 50
MaxIter: 10
```

View the Predictions

Print(RMSE)

```
https://ssh.cloud.google.com/v2/ssh/projects/scala-3/zones/us-central1-a/instances/instance-for-mlib?authuser=0&hl=en_US&projectNumber=179637453175&useAdminProxy=true&troubleshoot4005Enabled=true&troubleshoot255Enabled=true&sshT... — X
ssh.cloud.google.com/v2/ssh/projects/scala-3/zones/us-central1-a/instances/instance-for-mlib?authuser=0&hl=en_US&projectNumber=179637453175&useAdminProxy=true&troubleshoot4005Enabled=true&troubleshoot255Enabled=true... — X
SSH-in-browser UPLOAD FILE DOWNLOAD FILE
22/11/19 02:03:57 INFO Executor: Running task 0.0 in stage 8947.0 (TID 21845)
22/11/19 02:03:57 INFO Executor: Running task 1.0 in stage 8947.0 (TID 21846)
22/11/19 02:03:57 INFO ShuffleBlockFetcherIterator: Getting 1 (1735.1 KiB) non-empty blocks including 1 (1735.1 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-merged-local and 0 (0.0 B) remote blocks
22/11/19 02:03:57 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
22/11/19 02:03:57 INFO ShuffleBlockFetcherIterator: Getting 1 (1696.0 KiB) non-empty blocks including 1 (1696.0 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-merged-local and 0 (0.0 B) remote blocks
22/11/19 02:03:57 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 11 ms
22/11/19 02:03:57 INFO Executor: Finished task 0.0 in stage 8947.0 (TID 21845). 7908 bytes result sent to driver
22/11/19 02:03:58 INFO TaskSetManager: Finished task 0.0 in stage 8947.0 (TID 21845) in 260 ms on instance-for-mlib.us-central1-a.c.scala-3.internal (executor driver) (1/2)
22/11/19 02:03:58 INFO BlockManagerInfo: Removed broadcast_2998_piece0 on instance-for-mlib.us-central1-a.c.scala-3.internal:46767 in memory (size: 15.6 KiB, free: 421.8 MiB)
22/11/19 02:03:58 INFO Executor: Finished task 1.0 in stage 8947.0 (TID 21846). 7908 bytes result sent to driver
22/11/19 02:03:58 INFO TaskSetManager: Finished task 1.0 in stage 8947.0 (TID 21846) in 294 ms on instance-for-mlib.us-central1-a.c.scala-3.internal (executor driver) (2/2)
22/11/19 02:03:58 INFO TaskSchedulerImpl: Removed TaskSet 8947.0, whose tasks have all completed, from pool
22/11/19 02:03:58 INFO DAGScheduler: ResultStage 8947 (treeAggregate at Statistics.scala:58) finished in 0.328 s
22/11/19 02:03:58 INFO DAGScheduler: Job 952 is finished. Cancelling potential speculative or zombie tasks for this job
22/11/19 02:03:58 INFO TaskSchedulerImpl: Killing all running tasks in stage 8947: Stage finished
22/11/19 02:03:58 INFO DAGScheduler: Job 952 finished: treeAggregate at Statistics.scala:58, took 0.328835 s
0.8685666257690707
22/11/19 02:03:58 INFO BlockManagerInfo: Removed broadcast_2995_piece0 on instance-for-mlib.us-central1-a.c.scala-3.internal:46767 in memory (size: 32.1 KiB, free: 421.9 MiB)
22/11/19 02:03:58 INFO BlockManagerInfo: Removed broadcast_2996_piece0 on instance-for-mlib.us-central1-a.c.scala-3.internal:46767 in memory (size: 3.8 KiB, free: 421.9 MiB)
22/11/19 02:03:58 INFO BlockManagerInfo: Removed broadcast_2999_piece0 on instance-for-mlib.us-central1-a.c.scala-3.internal:46767 in memory (size: 3.8 KiB, free: 421.9 MiB)
```

Best Model , Rank, MaxIter, RegParam

```
https://ssh.cloud.google.com/v2/ssh/projects/scala-3/zones/us-central1-a/instances/instance-for-mlib?authuser=0&hl=en_US&projectNumber=179637453175&useAdminProxy=true&troubleshoot4005Enabled=true&troubleshoot255Enabled=true&sshT... — X
ssh.cloud.google.com/v2/ssh/projects/scala-3/zones/us-central1-a/instances/instance-for-mlib?authuser=0&hl=en_US&projectNumber=179637453175&useAdminProxy=true&troubleshoot4005Enabled=true&troubleshoot255Enabled=true...
SSH-in-browser UPLOAD FILE DOWNLOAD FILE ... ☰ 🔍 ⚙️

22/11/19 02:00:55 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
22/11/19 02:00:55 INFO MemoryStore: Block rdd_19221_8 stored as values in memory (estimated size 203.5 KiB, free 420.7 MiB)
22/11/19 02:00:55 INFO BlockManagerInfo: Added rdd_19221_8 in memory on instance-for-mlib.us-central1-a.c.scala-3.internal:46767 (size: 203.5 KiB, free: 423.4 MiB)
22/11/19 02:00:55 INFO Executor: Finished task 8.0 in stage 8889.0 (TID 21819). 2236 bytes result sent to driver
22/11/19 02:00:55 INFO TaskSetManager: Finished task 8.0 in stage 8889.0 (TID 21819) in 303 ms on instance-for-mlib.us-central1-a.c.scala-3.internal (executor driver) (9/10)
22/11/19 02:00:56 INFO MemoryStore: Block rdd_19221_9 stored as values in memory (estimated size 203.5 KiB, free 420.5 MiB)
22/11/19 02:00:56 INFO BlockManagerInfo: Added rdd_19221_9 in memory on instance-for-mlib.us-central1-a.c.scala-3.internal:46767 (size: 203.5 KiB, free: 423.2 MiB)
22/11/19 02:00:56 INFO Executor: Finished task 9.0 in stage 8889.0 (TID 21820). 2236 bytes result sent to driver
22/11/19 02:00:56 INFO TaskSetManager: Finished task 9.0 in stage 8889.0 (TID 21820) in 494 ms on instance-for-mlib.us-central1-a.c.scala-3.internal (executor driver) (10/10)
22/11/19 02:00:56 INFO TaskSchedulerImpl: Removed TaskSet 8889.0, whose tasks have all completed, from pool
22/11/19 02:00:56 INFO DAGScheduler: ResultStage 8889 (count at ALS.scala:1080) finished in 2.971 s
22/11/19 02:00:56 INFO DAGScheduler: Job 945 is finished. Cancelling potential speculative or zombie tasks for this job
22/11/19 02:00:56 INFO TaskSchedulerImpl: Killing all running tasks in stage 8889: Stage finished
22/11/19 02:00:56 INFO DAGScheduler: Job 945 finished: count at ALS.scala:1080, took 2.977674 s
22/11/19 02:00:56 INFO MapPartitionsRDD: Removing RDD 19202 from persistence list
22/11/19 02:00:56 INFO BlockManager: Removing RDD 19202
22/11/19 02:00:56 INFO MapPartitionsRDD: Removing RDD 19028 from persistence list
22/11/19 02:00:56 INFO BlockManager: Removing RDD 19028
22/11/19 02:00:56 INFO Instrumentation: [13ed0081] training finished
<class 'pyspark.ml.recommendation.ALSModel'>
**Best Model**
Rank: 50
MaxIter: 10
RegParam: 0.15
```

Test_predictions.show()

```
https://ssh.cloud.google.com/v2/ssh/projects/scala-3/zones/us-central1-a/instances/instance-for-mlib?authuser=0&hl=en_US&projectNumber=179637453175&useAdminProxy=true&troubleshoot4005Enabled=true&troubleshoot255Enabled=true&sshT... ━ X
ssh.cloud.google.com/v2/ssh/projects/scala-3/zones/us-central1-a/instances/instance-for-mlib?authuser=0&hl=en_US&projectNumber=179637453175&useAdminProxy=true&troubleshoot4005Enabled=true&troubleshoot255Enabled=true...
SSH-in-browser  UPLOAD FILE DOWNLOAD FILE ⌂ ⌂ ⌂ ⌂
22/11/19 02:04:29 INFO DAGScheduler: Job 959 finished: showString at NativeMethodAccessorImpl.java:0, took 0.024115 s
22/11/19 02:04:29 INFO CodeGenerator: Code generated in 9.295771 ms
+-----+-----+-----+
|userId|movieId|rating|prediction|
+-----+-----+-----+
| 580| 1580| 4.0| 3.4476714|
| 580| 44022| 3.5| 3.2499707|
| 597| 471| 2.0| 4.2078404|
| 108| 1959| 5.0| 3.929421|
| 368| 2122| 2.0| 1.8601143|
| 436| 471| 3.0| 3.6853335|
| 587| 1580| 4.0| 3.798573|
| 27| 1580| 3.0| 3.4053383|
| 606| 1580| 2.5| 3.169431|
| 606| 44022| 4.0| 2.8594952|
| 91| 2122| 4.0| 2.4488945|
| 157| 3175| 2.0| 3.3310113|
| 232| 1580| 3.5| 3.3821797|
| 232| 44022| 3.0| 3.142664|
| 246| 1645| 4.0| 3.80756|
| 599| 2366| 3.0| 2.8970969|
| 111| 1088| 3.0| 3.2433865|
| 111| 3175| 3.5| 3.0296626|
| 47| 1580| 1.5| 2.6835701|
| 140| 1580| 3.0| 3.4079206|
+-----+-----+-----+
only showing top 20 rows
```

'nrecommendations.limit(10).show()

```
https://ssh.cloud.google.com/v2/ssh/projects/scala-3/zones/us-central1-a/instances/instance-for-mlib?authuser=0&hl=en_US&projectNumber=179637453175&useAdminProxy=true&troubleshoot4005Enabled=true&troubleshoot255Enabled=true&sshT... ━ ━ ×  
ssh.cloud.google.com/v2/ssh/projects/scala-3/zones/us-central1-a/instances/instance-for-mlib?authuser=0&hl=en_US&projectNumber=179637453175&useAdminProxy=true&troubleshoot4005Enabled=true&troubleshoot255Enabled=true...  
 SSH-in-browser  UPLOAD FILE  DOWNLOAD FILE    22/11/19 02:05:12 INFO Executor: Running task 0.0 in stage 9155.0 (TID 22077)  
22/11/19 02:05:12 INFO ShuffleBlockFetcherIterator: Getting 2 (342.0 B) non-empty blocks including 2 (342.0 B) local and 0 (0.0 B) host-local and 0 (0.0 B) push-merged-local and 0 (0.0 B) remote blocks  
22/11/19 02:05:12 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms  
22/11/19 02:05:12 INFO Executor: Finished task 0.0 in stage 9155.0 (TID 22077). 2421 bytes result sent to driver  
22/11/19 02:05:12 INFO TaskSetManager: Finished task 0.0 in stage 9155.0 (TID 22077) in 5 ms on instance-for-mlib.us-central1-a.c.scala-3.internal (executor driver) (1/1)  
22/11/19 02:05:12 INFO TaskSchedulerImpl: Removed TaskSet 9155.0, whose tasks have all completed, from pool  
22/11/19 02:05:12 INFO DAGScheduler: ResultStage 9155 (showString at NativeMethodAccessorImpl.java:0) finished in 0.011 s  
22/11/19 02:05:12 INFO DAGScheduler: Job 965 is finished. Cancelling potential speculative or zombie tasks for this job  
22/11/19 02:05:12 INFO TaskSchedulerImpl: Killing all running tasks in stage 9155: Stage finished  
22/11/19 02:05:12 INFO DAGScheduler: Job 965 finished: showString at NativeMethodAccessorImpl.java:0, took 0.013687 s  
22/11/19 02:05:12 INFO CodeGenerator: Code generated in 10.814836 ms  
+-----+  
|userId|movieId| rating|  
+-----+-----+-----+  
| 1| 3379| 5.763239|  
| 1| 33649| 5.598928|  
| 1| 5490| 5.5296617|  
| 1| 171495| 5.416649|  
| 1| 5416| 5.4002886|  
| 1| 5328| 5.4002886|  
| 1| 3951| 5.4002886|  
| 1| 131724| 5.363606|  
| 1| 5915| 5.3629932|  
| 1| 177593| 5.356516|  
+-----+
```

```
nrecommendations.join(movies, on='movieId').filter('userId = 100').show()
```

https://ssh.cloud.google.com/v2/ssh/projects/scala-3/zones/us-central1-a/instances/instance-for-mlib?authuser=0&hl=en_US&projectNumber=179637453175&useAdminProxy=true&troubleshoot4005Enabled=true&troubleshoot255Enabled=true&sshT... — ○ ×
ssh.cloud.google.com/v2/ssh/projects/scala-3/zones/us-central1-a/instances/instance-for-mlib?authuser=0&hl=en_US&projectNumber=179637453175&useAdminProxy=true&troubleshoot4005Enabled=true&troubleshoot255Enabled=true...
 SSH-in-browser  UPLOAD FILE  DOWNLOAD FILE   

```
22/11/19 02:05:45 INFO Executor: Running task 0.0 in stage 9205.0 (TID 22179)
22/11/19 02:05:45 INFO ShuffleBlockFetcherIterator: Getting 10 (3.8 KiB) non-empty blocks including 10 (3.8 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-merged-local and 0 (0.0 B) remote blocks
22/11/19 02:05:45 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
22/11/19 02:05:45 INFO Executor: Finished task 0.0 in stage 9205.0 (TID 22179). 5133 bytes result sent to driver
22/11/19 02:05:45 INFO TaskSetManager: Finished task 0.0 in stage 9205.0 (TID 22179) in 31 ms on instance-for-mlib.us-central1-a.c.scala-3.internal (executor driver) (1/1)
22/11/19 02:05:45 INFO TaskSchedulerImpl: Removed TaskSet 9205.0, whose tasks have all completed, from pool
22/11/19 02:05:45 INFO DAGScheduler: ResultStage 9205 (showString at NativeMethodAccessorImpl.java:0) finished in 0.060 s
22/11/19 02:05:45 INFO DAGScheduler: Job 968 is finished. Cancelling potential speculative or zombie tasks for this job
22/11/19 02:05:45 INFO TaskSchedulerImpl: Killing all running tasks in stage 9205: Stage finished
22/11/19 02:05:45 INFO DAGScheduler: Job 968 finished: showString at NativeMethodAccessorImpl.java:0, took 0.064482 s
22/11/19 02:05:45 INFO CodeGenerator: Code generated in 9.458525 ms
```

movieId	userId	rating	title	genres
67618	100	5.120143	Strictly Sexual (...)	Comedy Drama Romance
3379	100	5.064743	On the Beach (1959)	Drama
42730	100	5.042285	Glory Road (2006)	Drama
33649	100	5.0216565	Saving Face (2004)	Comedy Drama Romance
184245	100	4.9267745	De platte jungle ...	Documentary
117531	100	4.9267745	Watermark (2014)	Documentary
84273	100	4.9267745	Zeitgeist: Moving...	Documentary
179135	100	4.9267745	Blue Planet II (2...	Documentary
7071	100	4.9267745	Woman Under the I...	Drama
26073	100	4.9267745	Human Condition I...	Drama War

```
ratings.join(movis, on='movieId').filter('userId=100').sort('rating',  
ascending=False).limit(10).show()
```

https://ssh.cloud.google.com/v2/ssh/projects/scala-3/zones/us-central1-a/instances/instance-for-mllib?authuser=0&hl=en_US&projectNumber=179637453175&useAdminProxy=true&troubleshoot4005Enabled=true&troubleshoot255Enabled=true&sshT... — X

ssh.cloud.google.com/v2/ssh/projects/scala-3/zones/us-central1-a/instances/instance-for-mllib?authuser=0&hl=en_US&projectNumber=179637453175&useAdminProxy=true&troubleshoot4005Enabled=true&troubleshoot255Enabled=true...

SSH-in-browser UPLOAD FILE DOWNLOAD FILE ≡ ⚙️

```
, executor driver, partition 0, PROCESS_LOCAL, 4911 bytes) taskResourceAssignments Map()
22/11/19 02:06:15 INFO Executor: Running task 0.0 in stage 9207.0 (TID 22181)
22/11/19 02:06:15 INFO FileScanRDD: Reading File path: file:///home/manickam/gmlib/gratings.csv, range: 0-2483721, partition values: [empty row]
22/11/19 02:06:16 INFO Executor: Finished task 0.0 in stage 9207.0 (TID 22181). 4080 bytes result sent to driver
22/11/19 02:06:16 INFO TaskSetManager: Finished task 0.0 in stage 9207.0 (TID 22181) in 481 ms on instance-for-mllib.us-central1-a.c.scala-3.internal (executor driver) (1/1)
22/11/19 02:06:16 INFO TaskSchedulerImpl: Removed TaskSet 9207.0, whose tasks have all completed, from pool
22/11/19 02:06:16 INFO DAGScheduler: ResultStage 9207 (showString at NativeMethodAccessorImpl.java:0) finished in 0.487 s
22/11/19 02:06:16 INFO DAGScheduler: Job 970 is finished. Cancelling potential speculative or zombie tasks for this job
22/11/19 02:06:16 INFO TaskSchedulerImpl: Killing all running tasks in stage 9207: Stage finished
22/11/19 02:06:16 INFO DAGScheduler: Job 970 finished: showString at NativeMethodAccessorImpl.java:0, took 0.489254 s
22/11/19 02:06:16 INFO CodeGenerator: Code generated in 16.496616 ms
+-----+-----+-----+-----+
| movieId | userId | rating | title | genres |
+-----+-----+-----+-----+
| 1101 | 100 | 5.0 | Top Gun (1986) | Action|Romance|
| 1958 | 100 | 5.0 | Terms of Endearment | Comedy|Drama|
| 2423 | 100 | 5.0 | Christmas Vacation | Comedy |
| 4041 | 100 | 5.0 | Officer and a Gentleman | Drama|Romance|
| 5620 | 100 | 5.0 | Sweet Home Alabama | Comedy|Romance|
| 368 | 100 | 4.5 | Maverick (1994) | Adventure|Comedy|...
| 934 | 100 | 4.5 | Father of the Bride | Comedy | | |
| 539 | 100 | 4.5 | Sleepless in Seattle | Comedy|Drama|Romance|
| 16 | 100 | 4.5 | Casino (1995) | Crime|Drama |
| 553 | 100 | 4.5 | Tombstone (1993) | Action|Drama|Western |
+-----+-----+-----+-----+
```

Inbox (11) - manickam@student: ~ | GGCP - Google Slides | VM instances – Compute Engine | +

console.cloud.google.com/compute/instances?onCreate=true&project=scala-3

Start your Free Trial with \$300 in credit. Don't worry—you won't be charged if you run out of credits. [Learn more](#)

DISMISS ACTIVATE

Google Cloud scala-3 Search for resources, docs, products, and more (/) Search

Compute Engine VM instances CREATE INSTANCE IMPORT VM REFRESH OPERATIONS HELP ASSISTANT LEARN SHOW INFO PANEL

Virtual machines

- VM instances
- Instance templates
- Sole-tenant nodes
- Machine images
- TPUs
- Committed use discounts
- Migrate to Virtual Machine

INSTANCE SCHEDULES

Get better visibility into your instance schedules.

VM instances are highly configurable infrastructure. [Learn more](#)

Filter Enter property name or value

Status Name IP Address

instance-for-mllib

Stop instance-for-mllib?

Stop shuts down the instance. If the shutdown doesn't complete within 90 seconds, the instance is forced to halt. This can lead to file-system corruption. Do you want to stop instance "instance-for-mllib"?

CANCEL STOP

External IP Connect
4.134.118.89 (nic0) SSH ...

Related actions

Explore Backup and DR NEW

View billing report

Monitor VMs

Explore VM logs

Set up firewall rules

Patch management

Release Notes

Show all

GGCP (1).pdf GGCP.pdf GGCP.pptx

Inbox (11) - manickam@student: ~ | GGCP - Google Slides | VM instances – Compute Engine +

console.cloud.google.com/compute/instances?onCreate=true&project=scala-3

Start your Free Trial with \$300 in credit. Don't worry—you won't be charged if you run out of credits. [Learn more](#)

DISMISS ACTIVATE

Google Cloud scala-3 Search for resources, docs, products, and more (/) Search

Compute Engine VM instances CREATE INSTANCE IMPORT VM REFRESH OPERATIONS HELP ASSISTANT LEARN SHOW INFO PANEL

Virtual machines

- VM instances
- Instance templates
- Sole-tenant nodes
- Machine images
- TPUs
- Committed use discounts
- Migrate to Virtual Machine...

INSTANCE SCHEDULES

Get better visibility into your VMs by installing Ops Agent - aggregate logs and metrics in one place. [Learn more](#)

DISMISS

INSTANCES

VM instances are highly configurable infrastructure. [Learn more](#)

Filter Enter property name

Status Name

Status Name

CANCEL DELETE

External IP Connect

SSH

Related actions

Explore Backup and DR NEW

View billing report

Monitor VMs

Explore VM logs

Set up firewall rules

Patch management

HIDE

GGCP (1).pdf GGCP.pdf GGCP.pptx Show all

Conclusion

Many recommendation systems suggest items to users by utilizing the techniques of collaborative filtering (CF) based on historical records of items that the users have viewed, purchased, or rated. Two major problems that most CF approaches must contend with are scalability and sparseness of the user profiles. To tackle these issues, in this program, we followed a CF algorithm Alternating-least-squares,

Similar logic is followed by amazon, Netflix, google.

References: SFBU exercises materials