Movie Recommendation with MLlib - Collaborative Filtering

CS570 Big Data Processing Project
By Feven Araya
Instructor: Dr. Chang, Henry

Table of contents

1. Introduction
2. Design
3. Implementation
4. Testing
5. Enhancement
6. Conclusion
7. References

01 Introduction

Project Context

- Introduction to the use of machine learning in the entertainment industry, specifically in recommending movies to users based on their preferences.
- Brief overview of the significance of personalized recommendations in enhancing user experience and engagement on movie platforms.

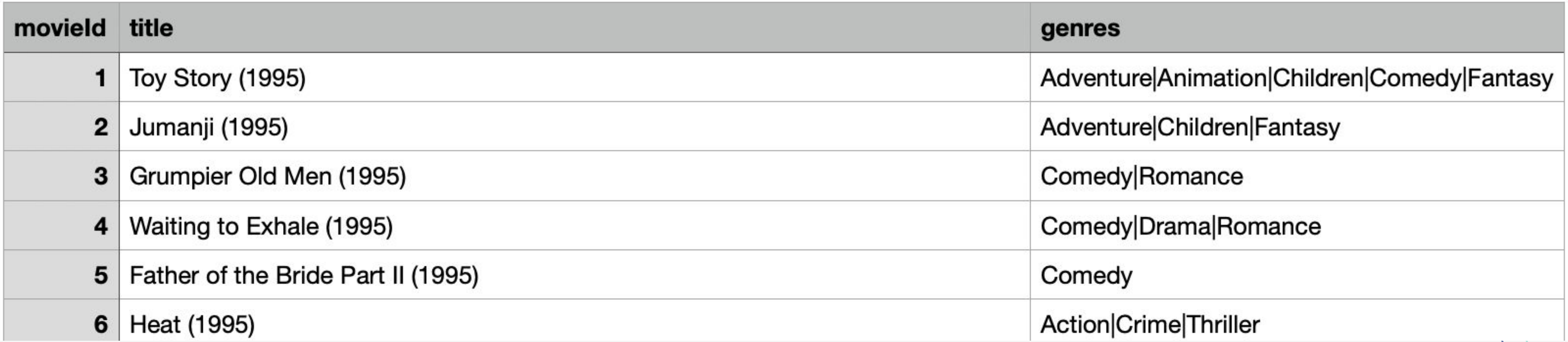
Objective of the Project

- To develop a machine learning model that accurately predicts user preferences and recommends movies using Collaborative Filtering techniques.
- Aim to leverage Apache Spark’s MLlib for efficient processing of large-scale movie rating data from the MovieLens dataset.

Title: Datasets

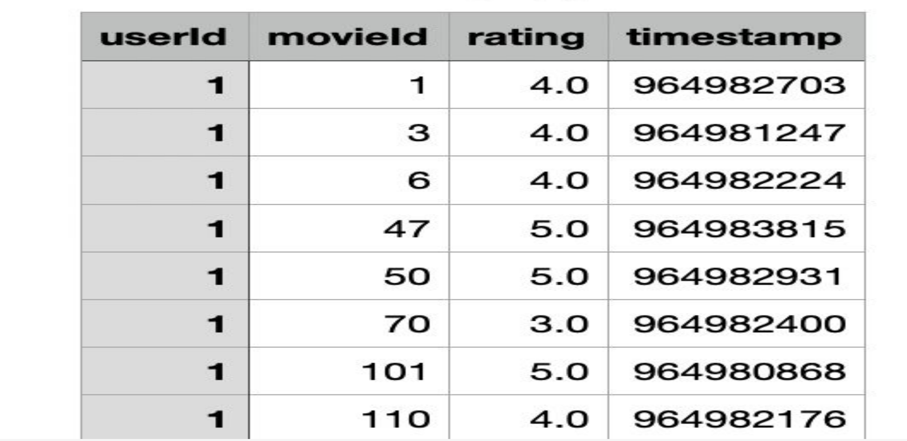
Content:

- **Movies Dataset (movies.csv)**:
  - Contains metadata about movies.
  - Columns: movieId, title, genres.



movieId	title	genres
1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy
2	Jumanji (1995)	Adventure|Children|Fantasy
3	Grumpier Old Men (1995)	Comedy|Romance
4	Waiting to Exhale (1995)	Comedy|Drama|Romance
5	Father of the Bride Part II (1995)	Comedy
6	Heat (1995)	Action|Crime|Thriller

- **Ratings Dataset (ratings.csv)**:
  - Contains user ratings for movies.
  - Columns: userId, movieId, rating, timestamp.



userId	movieId	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

02 Design

- ○ Google Colab
- ○ GCP

Identifying and Understanding the Problems

- Dataset sparsity
- Cold start problem
- Objective: Provide accurate movie recommendations despite these challenges

Investigating Solutions

- Traditional collaborative filtering techniques
- Content-based filtering
- Hybrid models
- Our approach: Collaborative Filtering with ALS

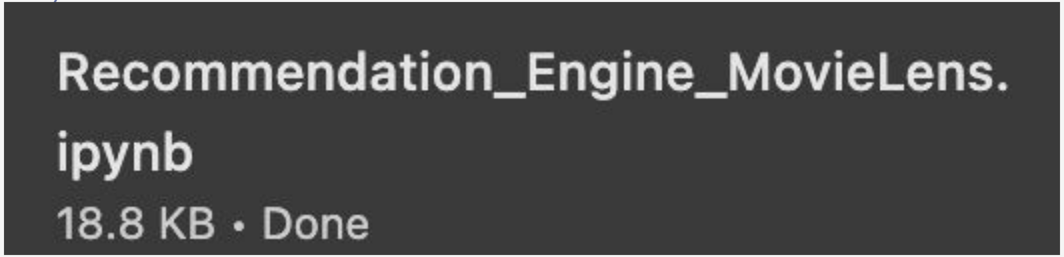
Theoretical Comparison and Selection

- Comparison of different recommendation algorithms
- Benefits of ALS:
  - Handles large datasets efficiently
  - Suitable for implicit feedback
- Selection rationale: ALS for its scalability and effectiveness

03

Implementation

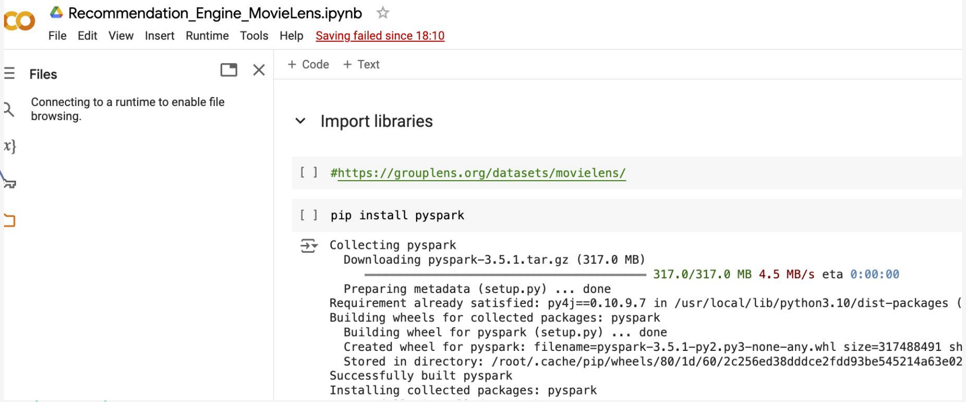
- **Step 3.1: Download the Pyspark code (ipynb)**



Recommendation_Engine_MovieLens.ipynb

18.8 KB • Done

- **Step 3.2: Upload the ipynb file to your Colab**



Recommendation_Engine_MovieLens.ipynb
File Edit View Insert Runtime Tools Help Saving failed since 18:10
Files
+ Code + Text

Connecting to a runtime to enable file browsing.

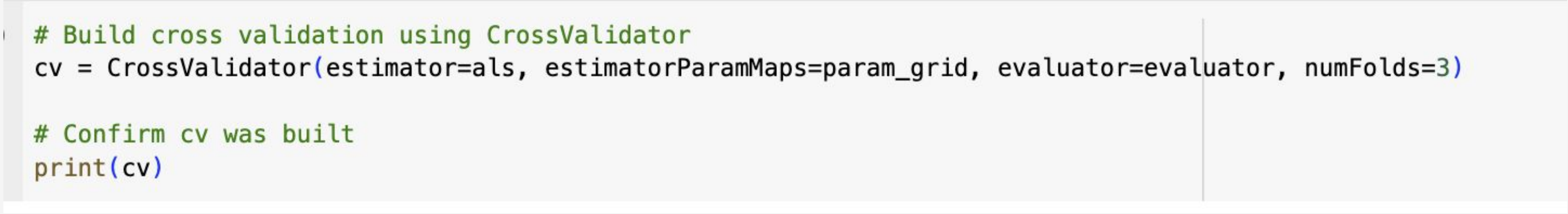
Import libraries

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

```
[ ] pip install pyspark
```

```
Collecting pyspark
  Downloading pyspark-3.5.1.tar.gz (317.0 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 317.0/317.0 MB 4.5 MB/s eta 0:00:00
  Preparing metadata (setup.py) ... done
Requirement already satisfied: py4j==0.10.9.7 in /usr/local/lib/python3.10/dist-packages (
Building wheels for collected packages: pyspark
  Building wheel for pyspark (setup.py) ... done
  Created wheel for pyspark: filename=pyspark-3.5.1-py2.py3-none-any.whl size=317488491 sh
  Stored in directory: /root/.cache/pip/wheels/80/1d/60/2c256ed38dddce2fdd93be545214a63e02
Successfully built pyspark
Installing collected packages: pyspark
```

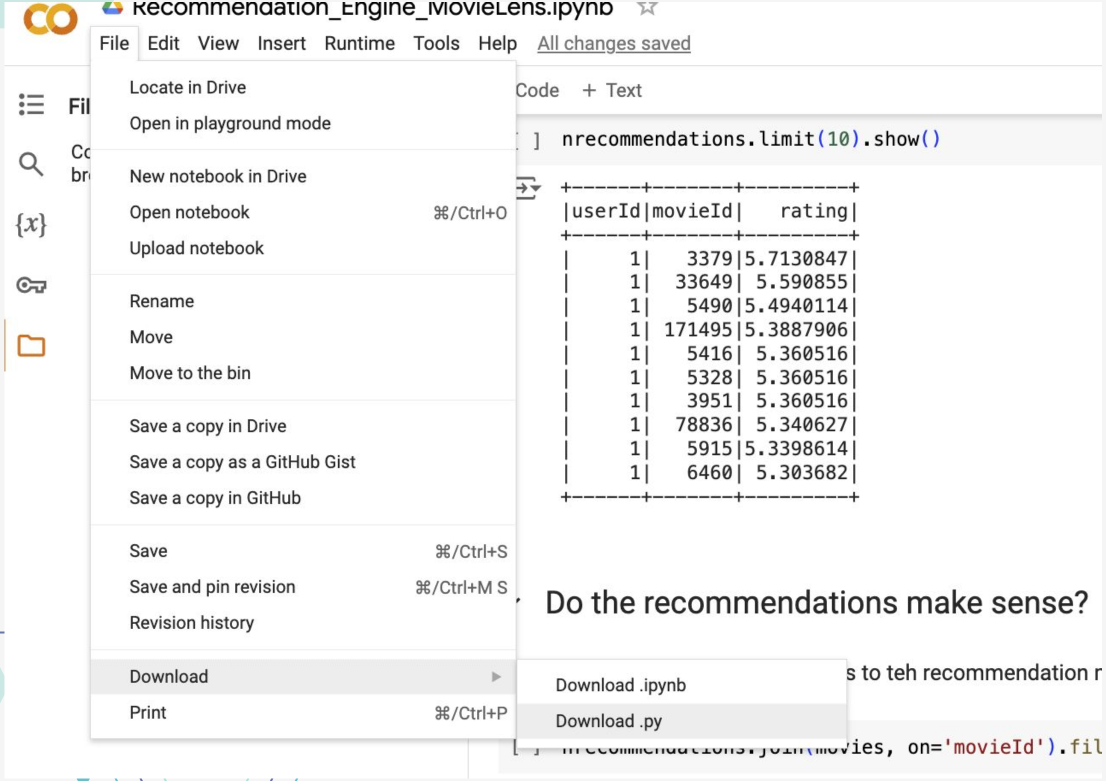
- Step 3.3: Experiment Pyspark code (ipynb) by modifying the ipynb file



```
# Build cross validation using CrossValidator
cv = CrossValidator(estimator=als, estimatorParamMaps=param_grid, evaluator=evaluator, numFolds=3)

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

Change number of folds to 5

Save the modified ipynb file as py and HTML format



```
nrecommendations.limit(10).show()
```

```
+------+-------+---------+
|userId|movieId|   rating|
+------+-------+---------+
|     1|   3379|5.7130847|
|     1|  33649| 5.590855|
|     1|   5490|5.4940114|
|     1| 171495|5.3887906|
|     1|   5416| 5.360516|
|     1|   5328| 5.360516|
|     1|   3951| 5.360516|
|     1|  78836| 5.340627|
|     1|   5915|5.3398614|
|     1|   6460| 5.303682|
+------+-------+---------+
```

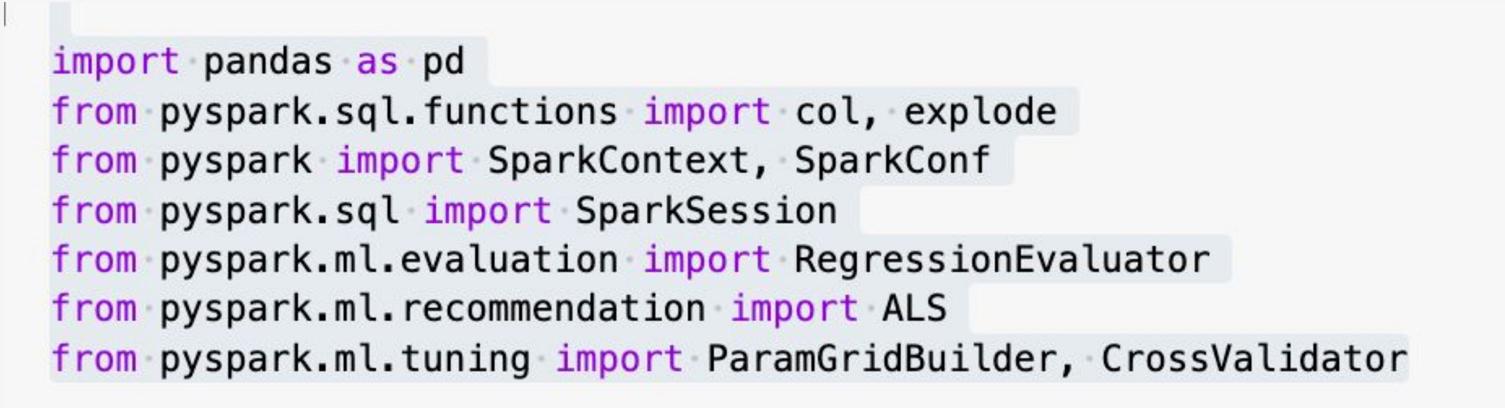
Do the recommendations make sense?

s to teh recommendation r

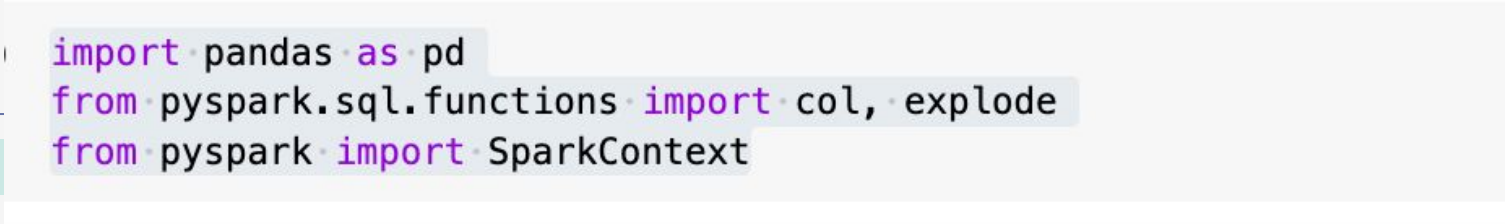
```
nrecommendations.join(movies, on='movieId').fil
```

- **Step 3.5: Run the py file saved at Step 3.4 on GCP**

Packages

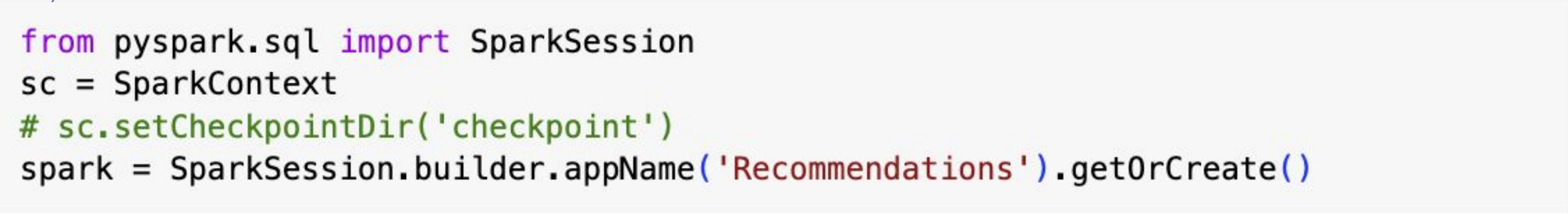


```python
import pandas as pd
from pyspark.sql.functions import col, explode
from pyspark import SparkContext, SparkConf
from pyspark.sql import SparkSession
from pyspark.ml.evaluation import RegressionEvaluator
from pyspark.ml.recommendation import ALS
from pyspark.ml.tuning import ParamGridBuilder, CrossValidator
```



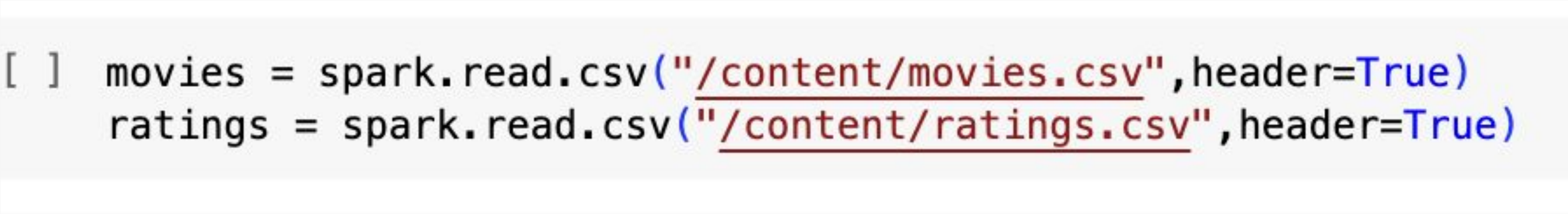
```python
import pandas as pd
from pyspark.sql.functions import col, explode
from pyspark import SparkContext
```

Initiate Spark Session



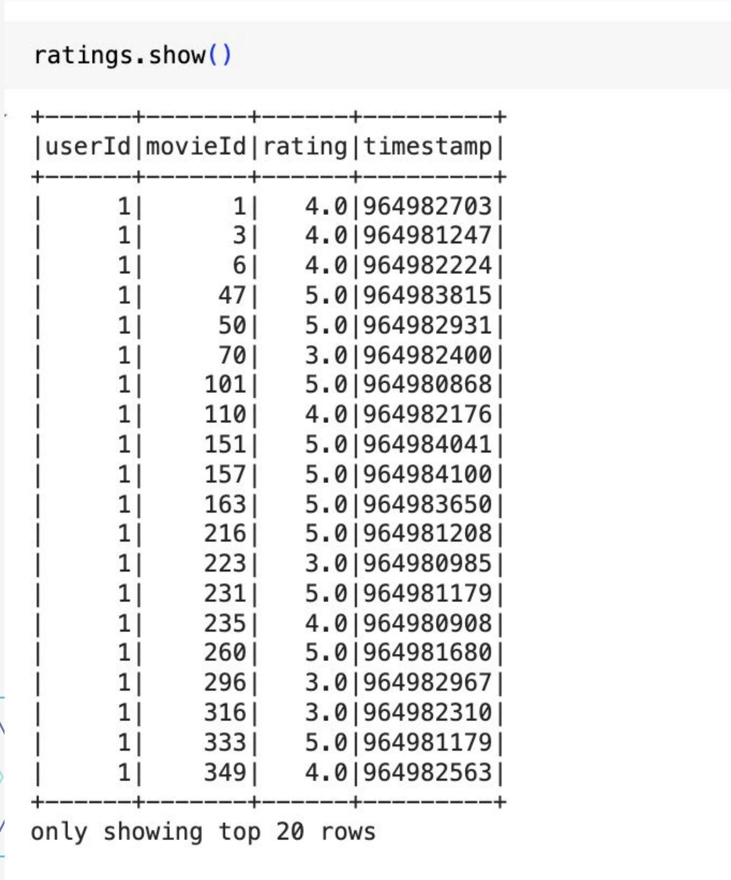
```
from pyspark.sql import SparkSession
sc = SparkContext
# sc.setCheckpointDir('checkpoint')
spark = SparkSession.builder.appName('Recommendations').getOrCreate()
```

Load Data



```
movies = spark.read.csv("/content/movies.csv",header=True)
ratings = spark.read.csv("/content/ratings.csv",header=True)
```

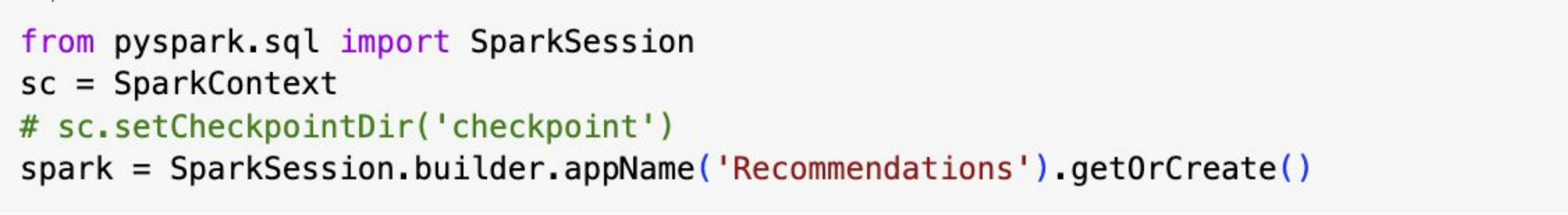
Display and Schema of Ratings Data



```
ratings.show()
```

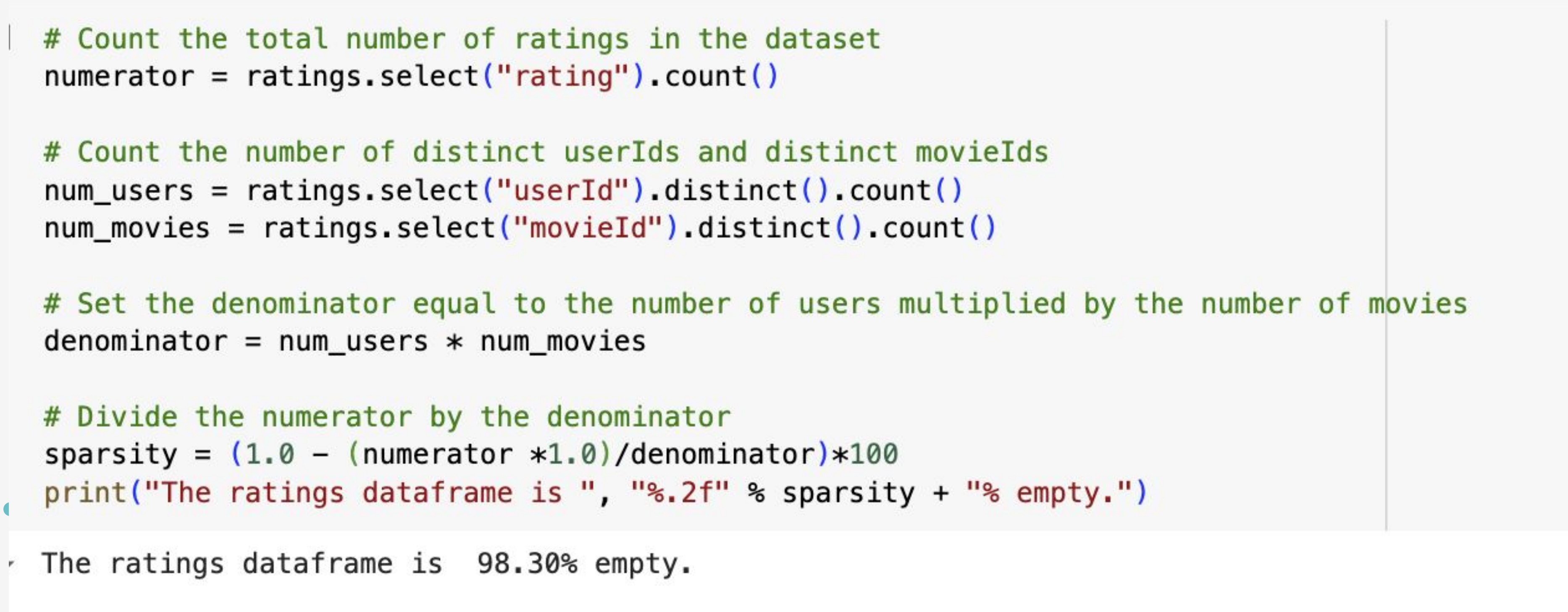
```
+------+-------+------+---------+
|userId|movieId|rating|timestamp|
+------+-------+------+---------+
|     1|      1|   4.0|964982703|
|     1|      3|   4.0|964981247|
|     1|      6|   4.0|964982224|
|     1|     47|   5.0|964983815|
|     1|     50|   5.0|964982931|
|     1|     70|   3.0|964982400|
|     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
```

Data Preprocessing



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

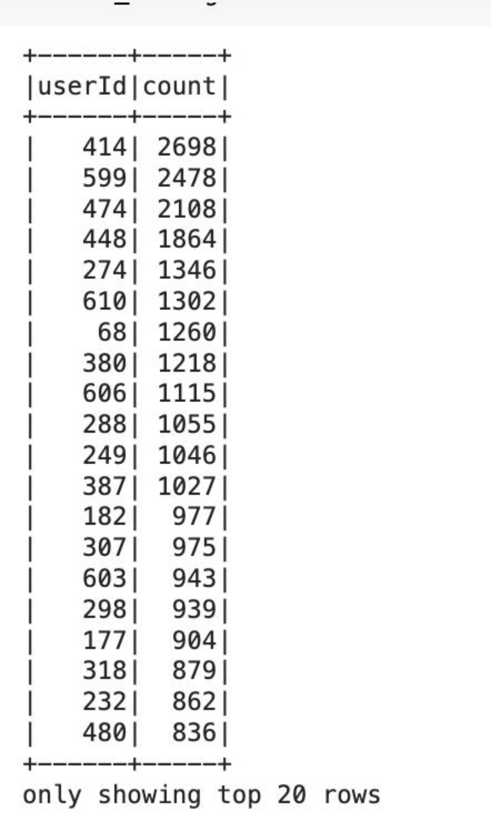
Calculate Sparsity



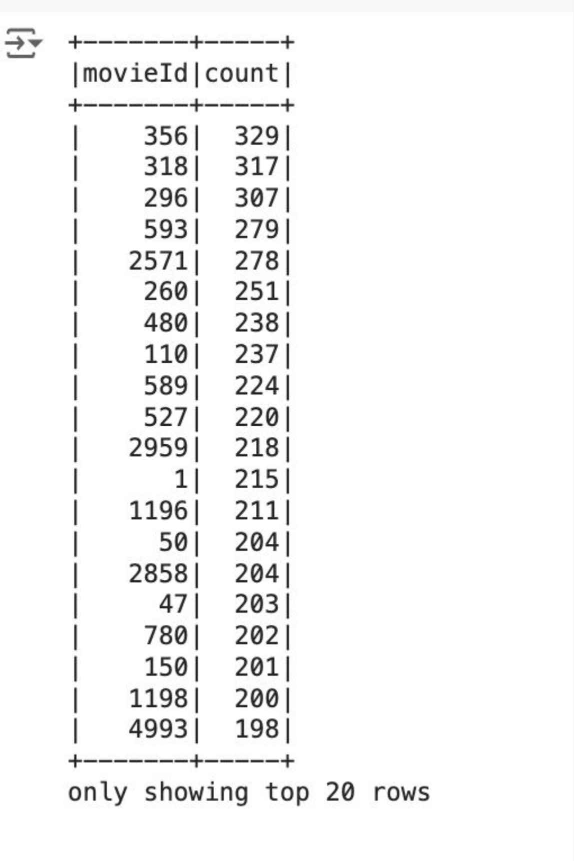
```
# Count the total number of ratings in the dataset
numerator = ratings.select("rating").count()

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

```
The ratings dataframe is  98.30% empty.
```

Interpret Ratings

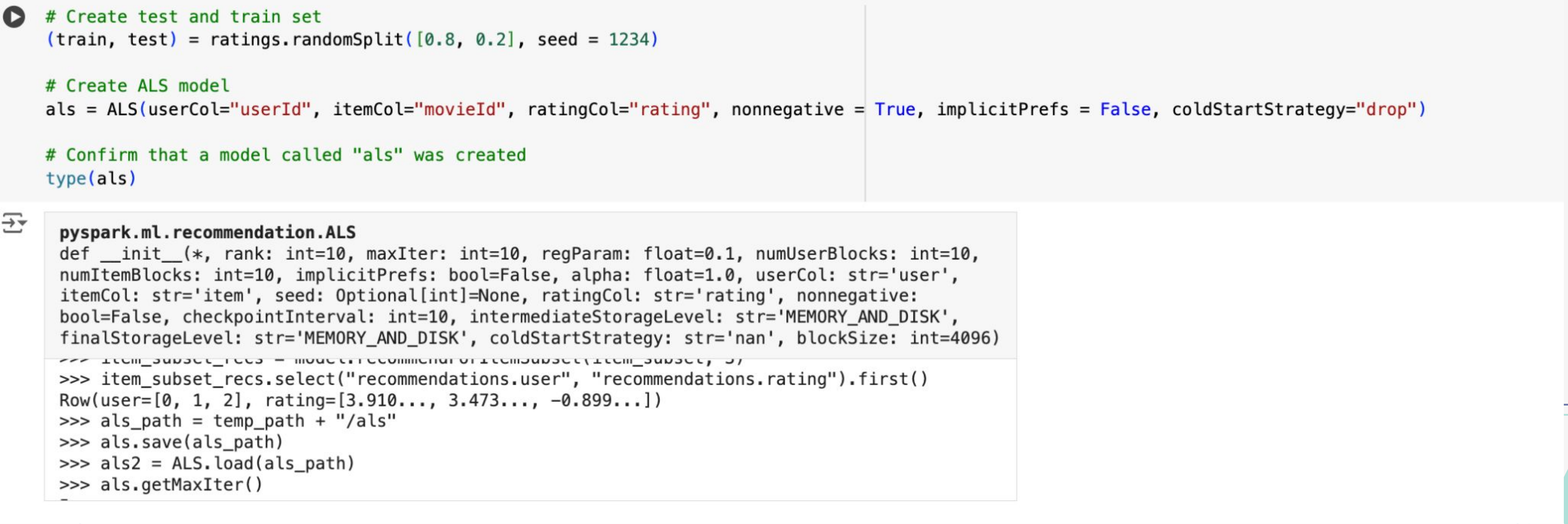


```
+------+-----+
|userId|count|
+------+-----+
|   414| 2698|
|   599| 2478|
|   474| 2108|
|   448| 1864|
|   274| 1346|
|   610| 1302|
|    68| 1260|
|   380| 1218|
|   606| 1115|
|   288| 1055|
|   249| 1046|
|   387| 1027|
|   182|  977|
|   307|  975|
|   603|  943|
|   298|  939|
|   177|  904|
|   318|  879|
|   232|  862|
|   480|  836|
+------+-----+
only showing top 20 rows
```



```
+-------+-----+
|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|
|     50|  204|
|   2858|  204|
|     47|  203|
|    780|  202|
|    150|  201|
|   1198|  200|
|   4993|  198|
+-------+-----+
only showing top 20 rows
```

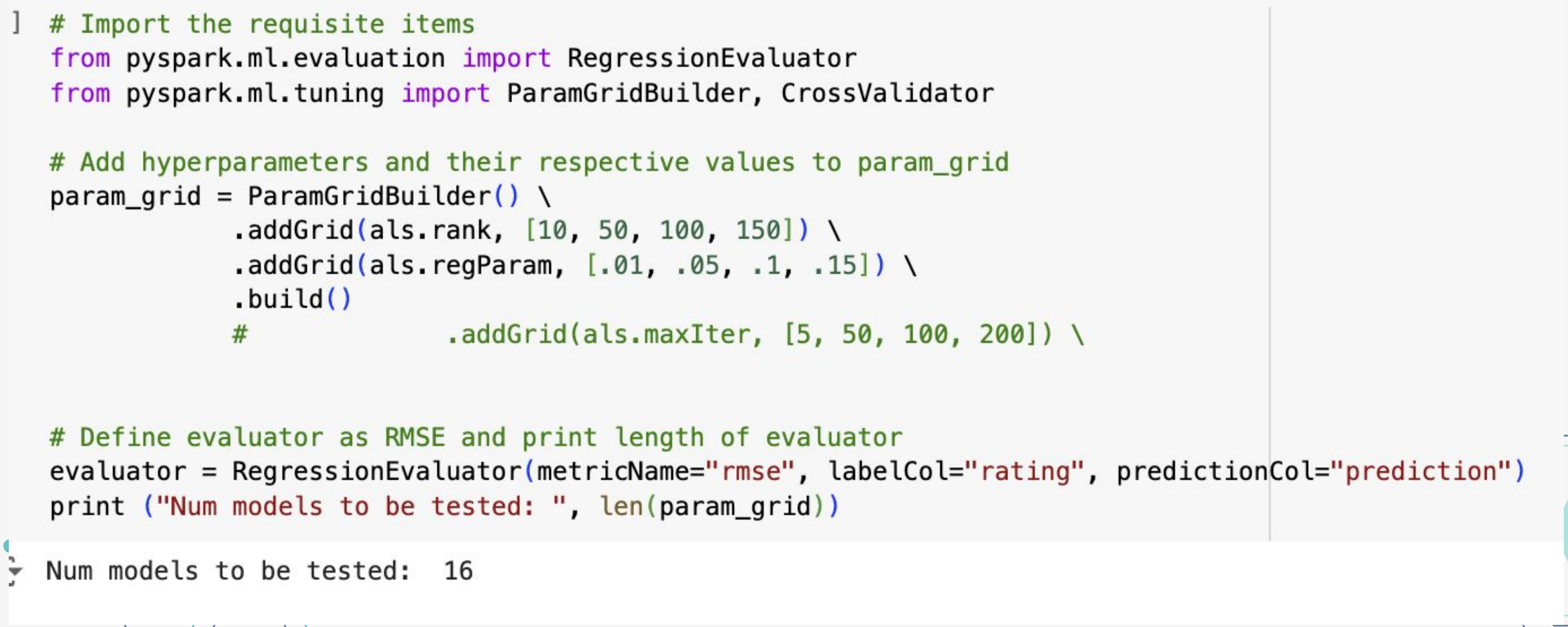
Build ALS Model



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

# 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)
```

```
pyspark.ml.recommendation.ALS
def __init__(*, rank: int=10, maxIter: int=10, regParam: float=0.1, numUserBlocks: int=10,
numItemBlocks: int=10, implicitPrefs: bool=False, alpha: float=1.0, userCol: str='user',
itemCol: str='item', seed: Optional[int]=None, ratingCol: str='rating', nonnegative:
bool=False, checkpointInterval: int=10, intermediateStorageLevel: str='MEMORY_AND_DISK',
finalStorageLevel: str='MEMORY_AND_DISK', coldStartStrategy: str='nan', blockSize: int=4096)
>>> item_subset_recs = model.recommendForItemSubset(item_subset, 3)
>>> item_subset_recs.select("recommendations.user", "recommendations.rating").first()
Row(user=[0, 1, 2], rating=[3.910..., 3.473..., -0.899...])
>>> als_path = temp_path + "/als"
>>> als.save(als_path)
>>> als2 = ALS.load(als_path)
>>> als.getMaxIter()
```

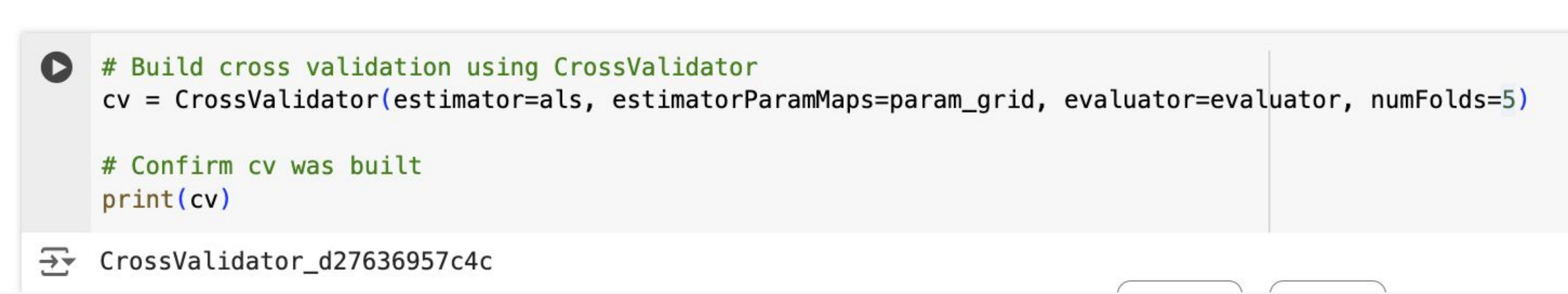
Tune ALS Model



```python
# 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))
```

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

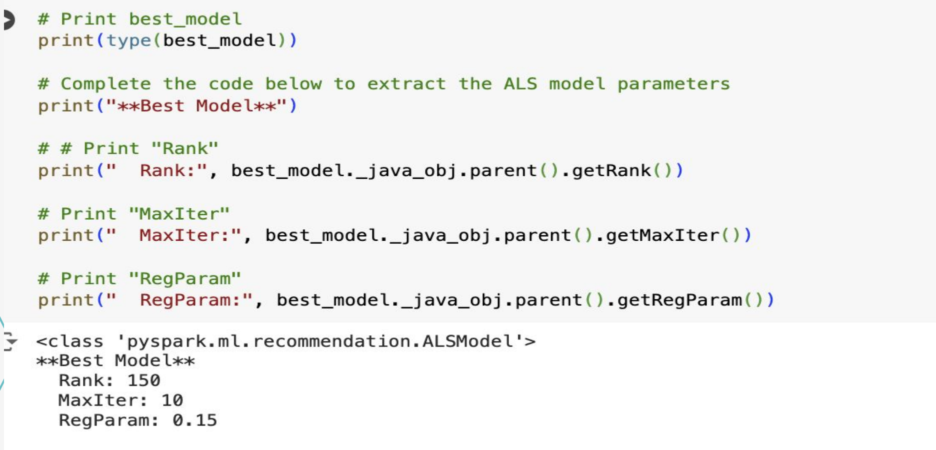
Cross-Validation



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

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

```
CrossValidator_d27636957c4c
```

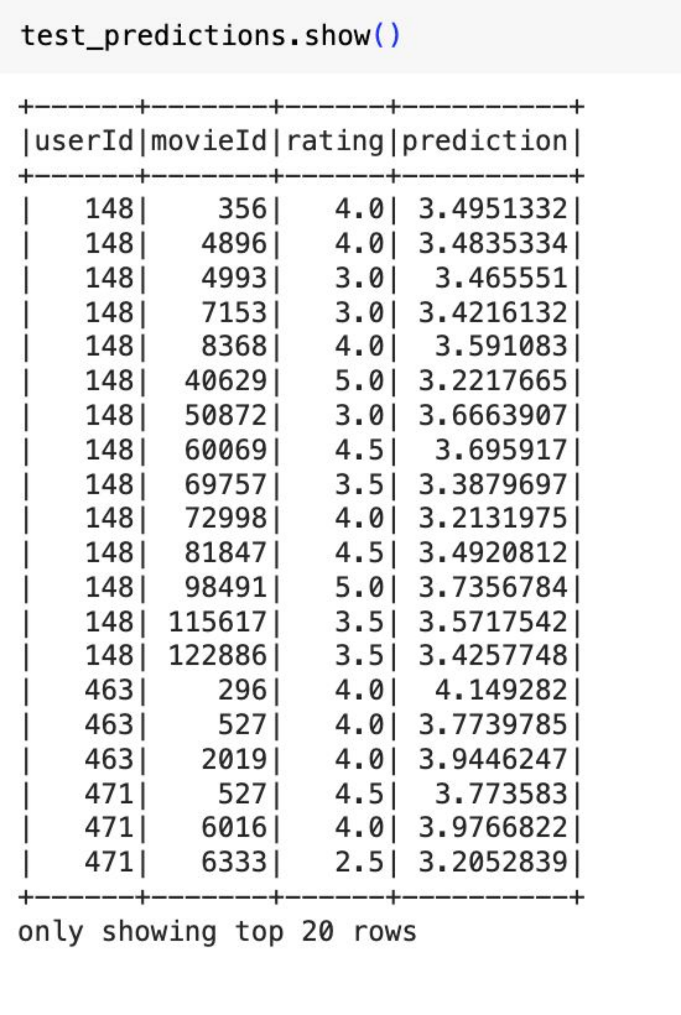
Train and Evaluate Model



```
# 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())
```

```
<class 'pyspark.ml.recommendation.ALSModel'>
**Best Model**
  Rank: 150
  MaxIter: 10
  RegParam: 0.15
```

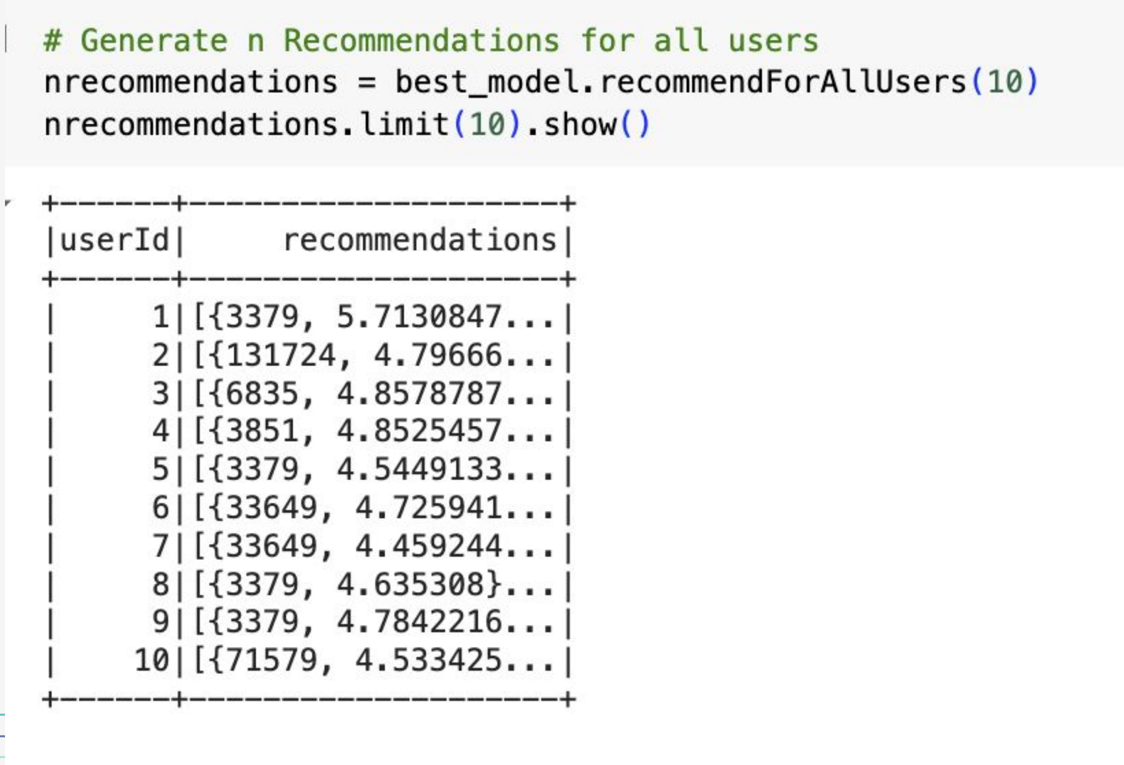
Make Predictions



```
test_predictions.show()
```

```
+------+-------+------+----------+
|userId|movieId|rating|prediction|
+------+-------+------+----------+
|   148|    356|   4.0| 3.4951332|
|   148|   4896|   4.0| 3.4835334|
|   148|   4993|   3.0|  3.465551|
|   148|   7153|   3.0| 3.4216132|
|   148|   8368|   4.0|  3.591083|
|   148|  40629|   5.0| 3.2217665|
|   148|  50872|   3.0| 3.6663907|
|   148|  60069|   4.5|  3.695917|
|   148|  69757|   3.5| 3.3879697|
|   148|  72998|   4.0| 3.2131975|
|   148|  81847|   4.5| 3.4920812|
|   148|  98491|   5.0| 3.7356784|
|   148| 115617|   3.5| 3.5717542|
|   148| 122886|   3.5| 3.4257748|
|   463|    296|   4.0|  4.149282|
|   463|    527|   4.0| 3.7739785|
|   463|   2019|   4.0| 3.9446247|
|   471|    527|   4.5|  3.773583|
|   471|   6016|   4.0| 3.9766822|
|   471|   6333|   2.5| 3.2052839|
+------+-------+------+----------+
only showing top 20 rows
```

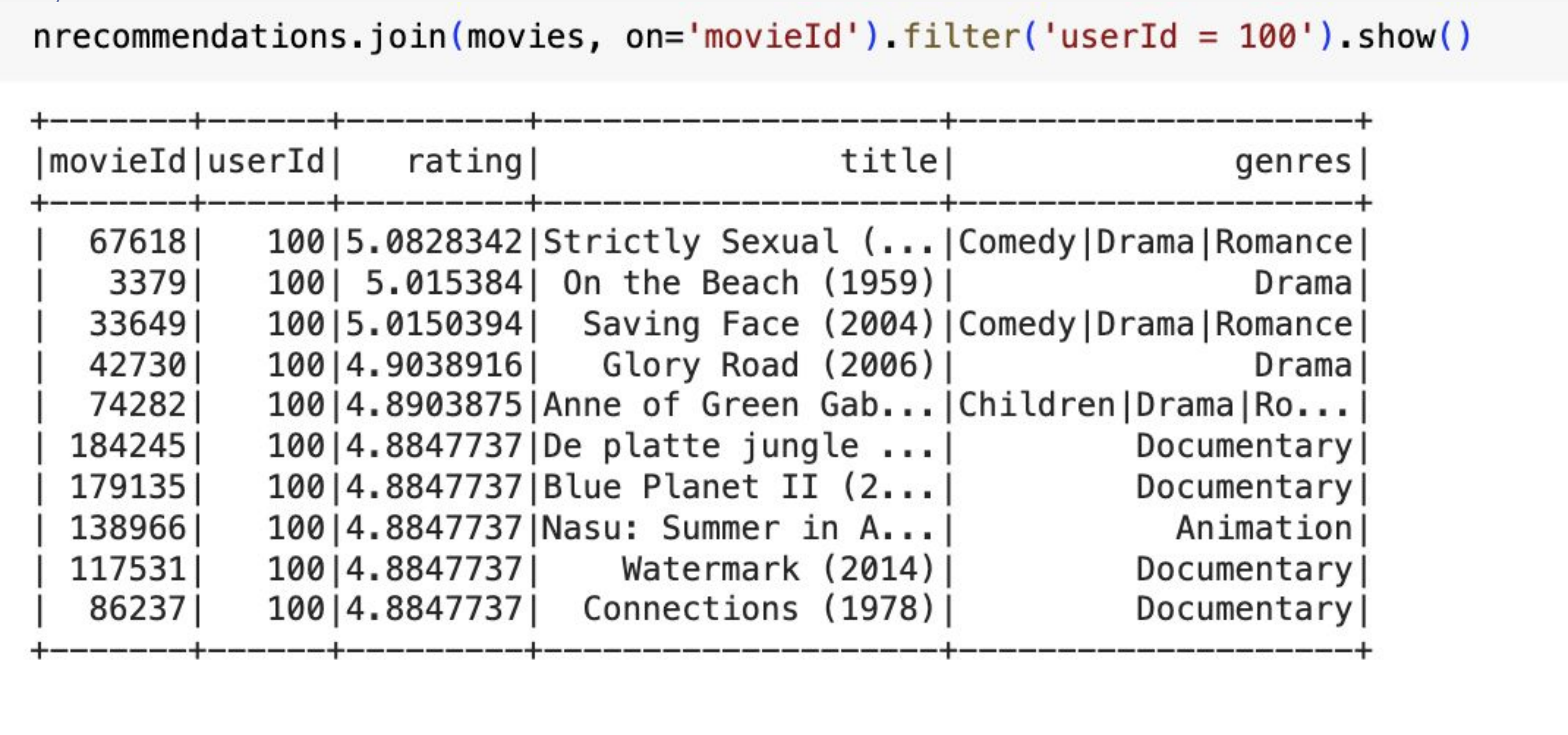
Generate Recommendations



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

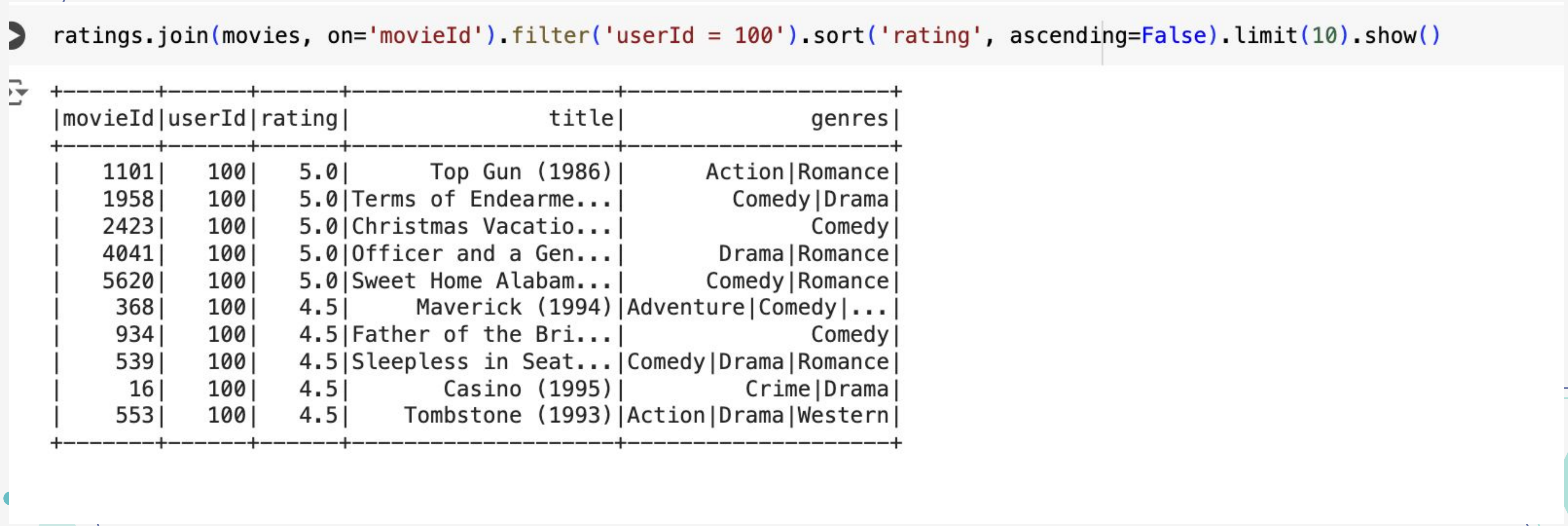
```
+------+--------------------+
|userId|     recommendations|
+------+--------------------+
|     1|[{3379, 5.7130847...|
|     2|[{131724, 4.79666...|
|     3|[{6835, 4.8578787...|
|     4|[{3851, 4.8525457...|
|     5|[{3379, 4.5449133...|
|     6|[{33649, 4.725941...|
|     7|[{33649, 4.459244...|
|     8|[{3379, 4.635308}...|
|     9|[{3379, 4.7842216...|
|    10|[{71579, 4.533425...|
+------+--------------------+
```

Merge with Movies Data for Interpretability



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

```
+-------+------+---------+--------------------+--------------------+
|movieId|userId|   rating|               title|              genres|
+-------+------+---------+--------------------+--------------------+
|  67618|   100|5.0828342|Strictly Sexual (...|Comedy|Drama|Romance|
|   3379|   100| 5.015384| On the Beach (1959)|               Drama|
|  33649|   100|5.0150394|  Saving Face (2004)|Comedy|Drama|Romance|
|  42730|   100|4.9038916|   Glory Road (2006)|               Drama|
|  74282|   100|4.8903875|Anne of Green Gab...|Children|Drama|Ro...|
| 184245|   100|4.8847737|De platte jungle ...|         Documentary|
| 179135|   100|4.8847737|Blue Planet II (2...|         Documentary|
| 138966|   100|4.8847737|Nasu: Summer in A...|           Animation|
| 117531|   100|4.8847737|    Watermark (2014)|         Documentary|
|  86237|   100|4.8847737|  Connections (1978)|         Documentary|
+-------+------+---------+--------------------+--------------------+
```



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

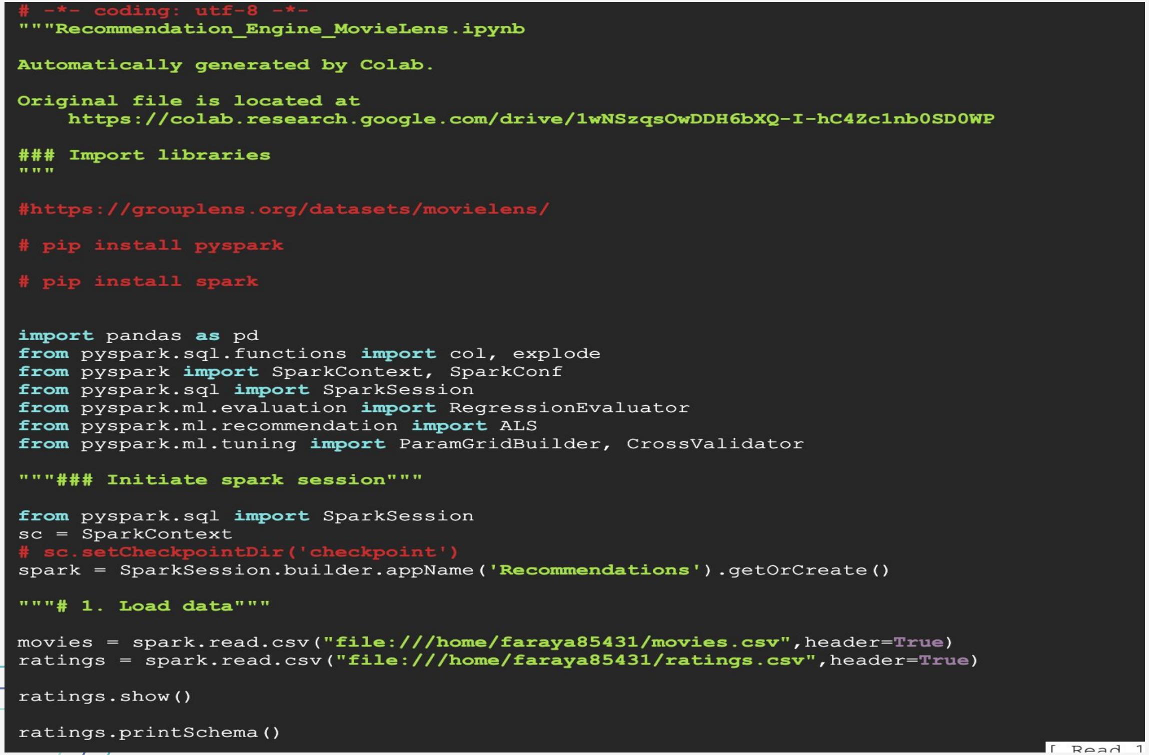
```
+-------+------+------+--------------------+--------------------+
|movieId|userId|rating|               title|              genres|
+-------+------+------+--------------------+--------------------+
|   1101|   100|   5.0|      Top Gun (1986)|      Action|Romance|
|   1958|   100|   5.0|Terms of Endearme...|        Comedy|Drama|
|   2423|   100|   5.0|Christmas Vacatio...|              Comedy|
|   4041|   100|   5.0|Officer and a Gen...|       Drama|Romance|
|   5620|   100|   5.0|Sweet Home Alabam...|      Comedy|Romance|
|    368|   100|   4.5|     Maverick (1994)|Adventure|Comedy|...|
|    934|   100|   4.5|Father of the Bri...|              Comedy|
|    539|   100|   4.5|Sleepless in Seat...|Comedy|Drama|Romance|
|     16|   100|   4.5|       Casino (1995)|         Crime|Drama|
|    553|   100|   4.5|    Tombstone (1993)|Action|Drama|Western|
+-------+------+------+--------------------+--------------------+
```

04

Test

Process to test the project

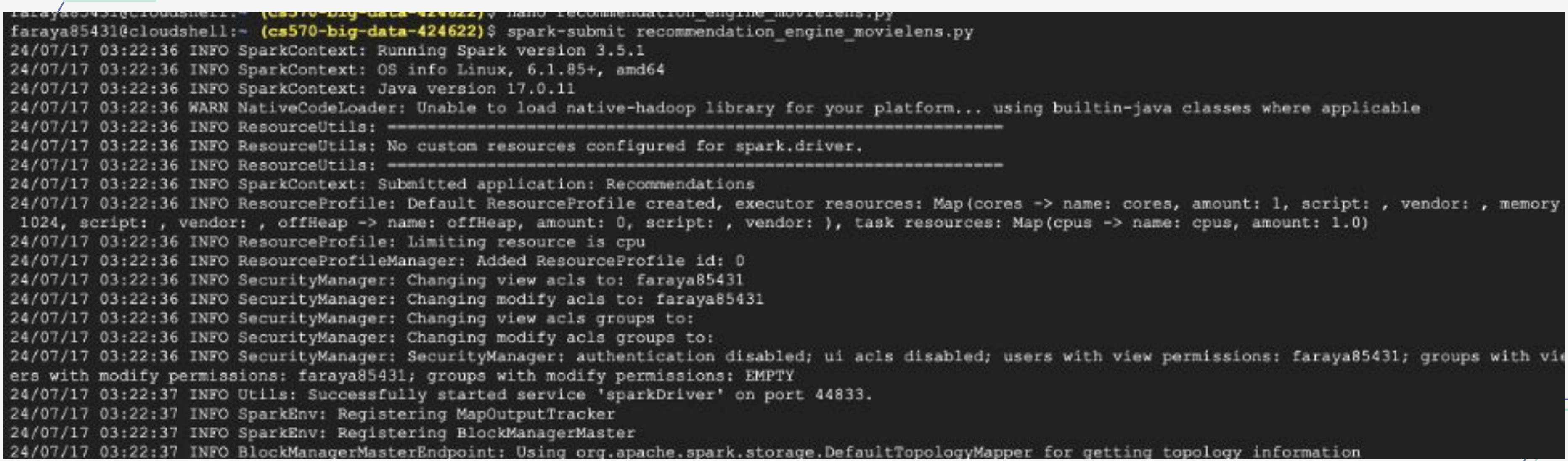
Open GCP and upload your the recommendation_Engine_MovieLens.py file



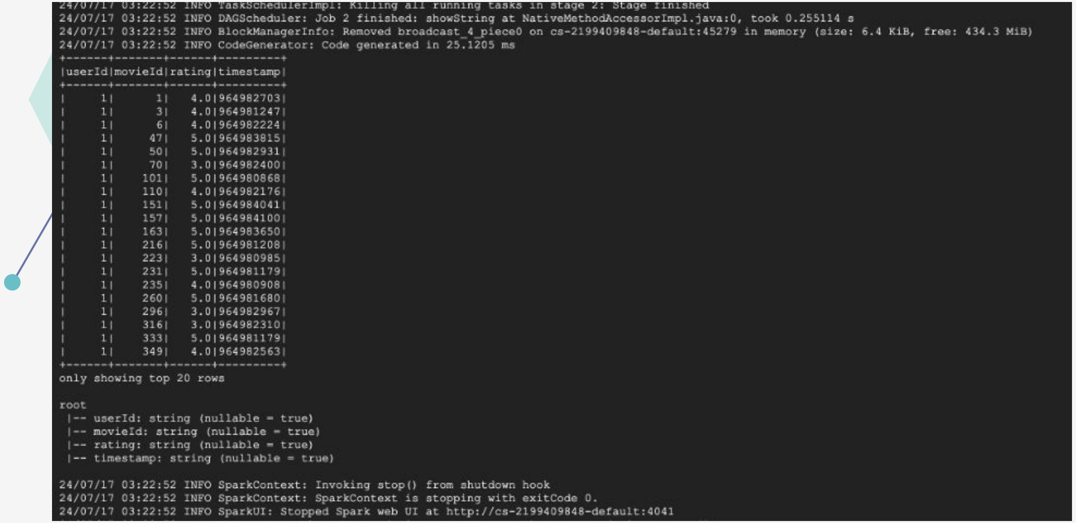
```
# -*- coding: utf-8 -*-
"""Recommendation_Engine_MovieLens.ipynb

Automatically generated by Colab.

Original file is located at
    https://colab.research.google.com/drive/1wNSzqsOwDDH6bXQ-I-hC4Zc1nb0SD0WP

### Import libraries
"""

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

# pip install pyspark

# pip install spark


import pandas as pd
from pyspark.sql.functions import col, explode
from pyspark import SparkContext, SparkConf
from pyspark.sql import SparkSession
from pyspark.ml.evaluation import RegressionEvaluator
from pyspark.ml.recommendation import ALS
from pyspark.ml.tuning import ParamGridBuilder, CrossValidator

"""### Initiate spark session"""

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

"""# 1. Load data"""

movies = spark.read.csv("file:///home/faraya85431/movies.csv",header=True)
ratings = spark.read.csv("file:///home/faraya85431/ratings.csv",header=True)

ratings.show()

ratings.printSchema()
```

[Read 1

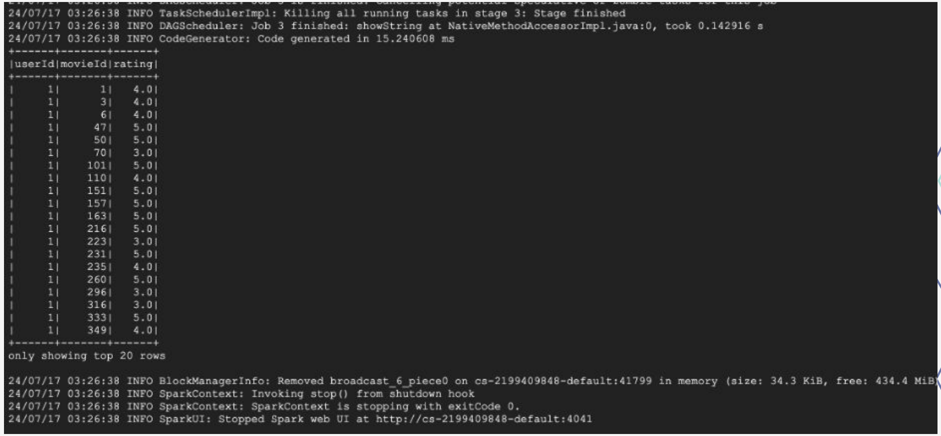
Run the py file



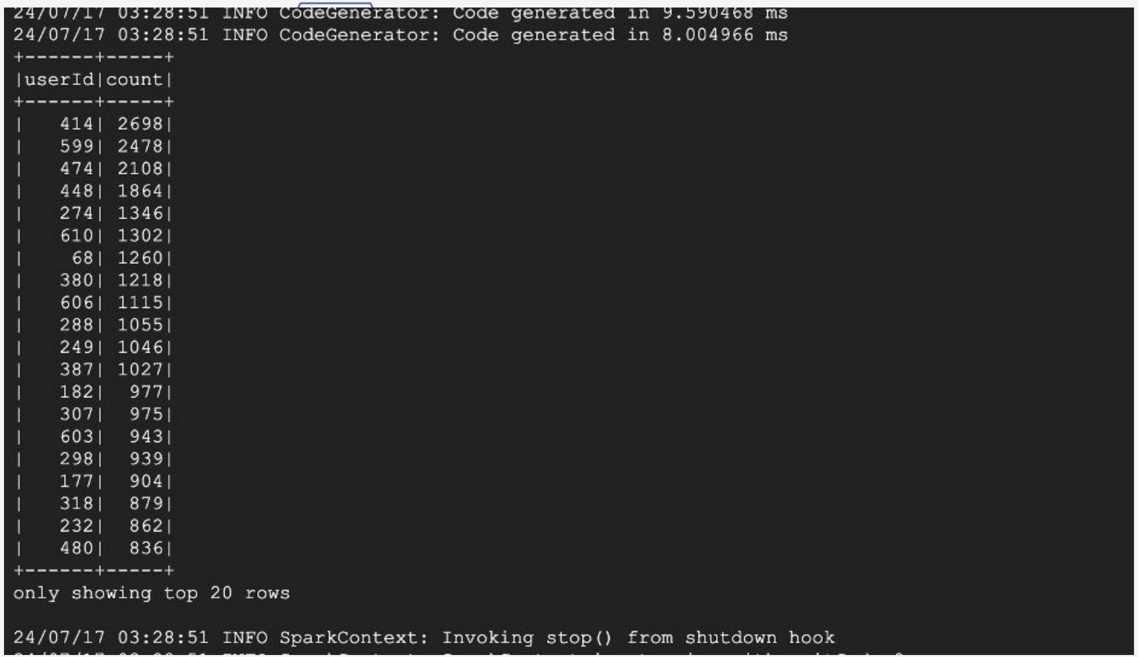
```
faraya85431@cloudshell:~ (cs570-big-data-424622)$ nano recommendation_engine_movielens.py
faraya85431@cloudshell:~ (cs570-big-data-424622)$ spark-submit recommendation_engine_movielens.py
24/07/17 03:22:36 INFO SparkContext: Running Spark version 3.5.1
24/07/17 03:22:36 INFO SparkContext: OS info Linux, 6.1.85+, amd64
24/07/17 03:22:36 INFO SparkContext: Java version 17.0.11
24/07/17 03:22:36 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
24/07/17 03:22:36 INFO ResourceUtils: ==============================================================
24/07/17 03:22:36 INFO ResourceUtils: No custom resources configured for spark.driver.
24/07/17 03:22:36 INFO ResourceUtils: ==============================================================
24/07/17 03:22:36 INFO SparkContext: Submitted application: Recommendations
24/07/17 03:22:36 INFO ResourceProfile: Default ResourceProfile created, executor resources: Map(cores -> name: cores, amount: 1, script: , vendor: , memory
 1024, script: , vendor: , offHeap -> name: offHeap, amount: 0, script: , vendor: ), task resources: Map(cpus -> name: cpus, amount: 1.0)
24/07/17 03:22:36 INFO ResourceProfile: Limiting resource is cpu
24/07/17 03:22:36 INFO ResourceProfileManager: Added ResourceProfile id: 0
24/07/17 03:22:36 INFO SecurityManager: Changing view acls to: faraya85431
24/07/17 03:22:36 INFO SecurityManager: Changing modify acls to: faraya85431
24/07/17 03:22:36 INFO SecurityManager: Changing view acls groups to:
24/07/17 03:22:36 INFO SecurityManager: Changing modify acls groups to:
24/07/17 03:22:36 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: faraya85431; groups with vie
ers with modify permissions: faraya85431; groups with modify permissions: EMPTY
24/07/17 03:22:37 INFO Utils: Successfully started service 'sparkDriver' on port 44833.
24/07/17 03:22:37 INFO SparkEnv: Registering MapOutputTracker
24/07/17 03:22:37 INFO SparkEnv: Registering BlockManagerMaster
24/07/17 03:22:37 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
```



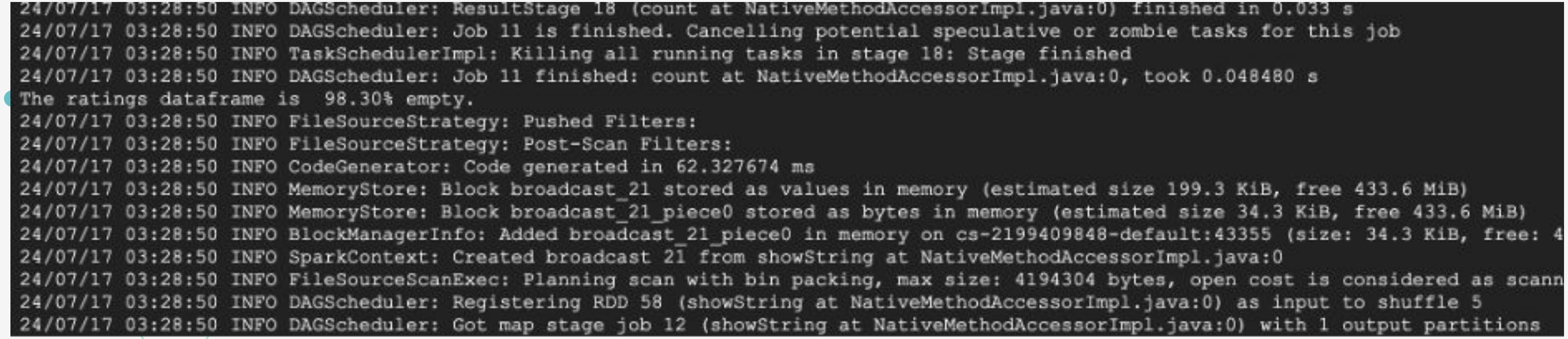
```
24/07/17 03:22:52 INFO TaskSchedulerImpl: Killing all running tasks in stage 2: Stage finished
24/07/17 03:22:52 INFO DAGScheduler: Job 2 finished: showString at NativeMethodAccessorImpl.java:0, took 0.255114 s
24/07/17 03:22:52 INFO BlockManagerInfo: Removed broadcast_4_piece0 on cs-2199409848-default:45279 in memory (size: 6.4 KiB, free: 434.3 MiB)
24/07/17 03:22:52 INFO CodeGenerator: Code generated in 25.1205 ms
+------+-------+------+---------+
|userId|movieId|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

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

24/07/17 03:22:52 INFO SparkContext: Invoking stop() from shutdown hook
24/07/17 03:22:52 INFO SparkContext: SparkContext is stopping with exitCode 0.
24/07/17 03:22:52 INFO SparkUI: Stopped Spark web UI at http://cs-2199409848-default:4041
```



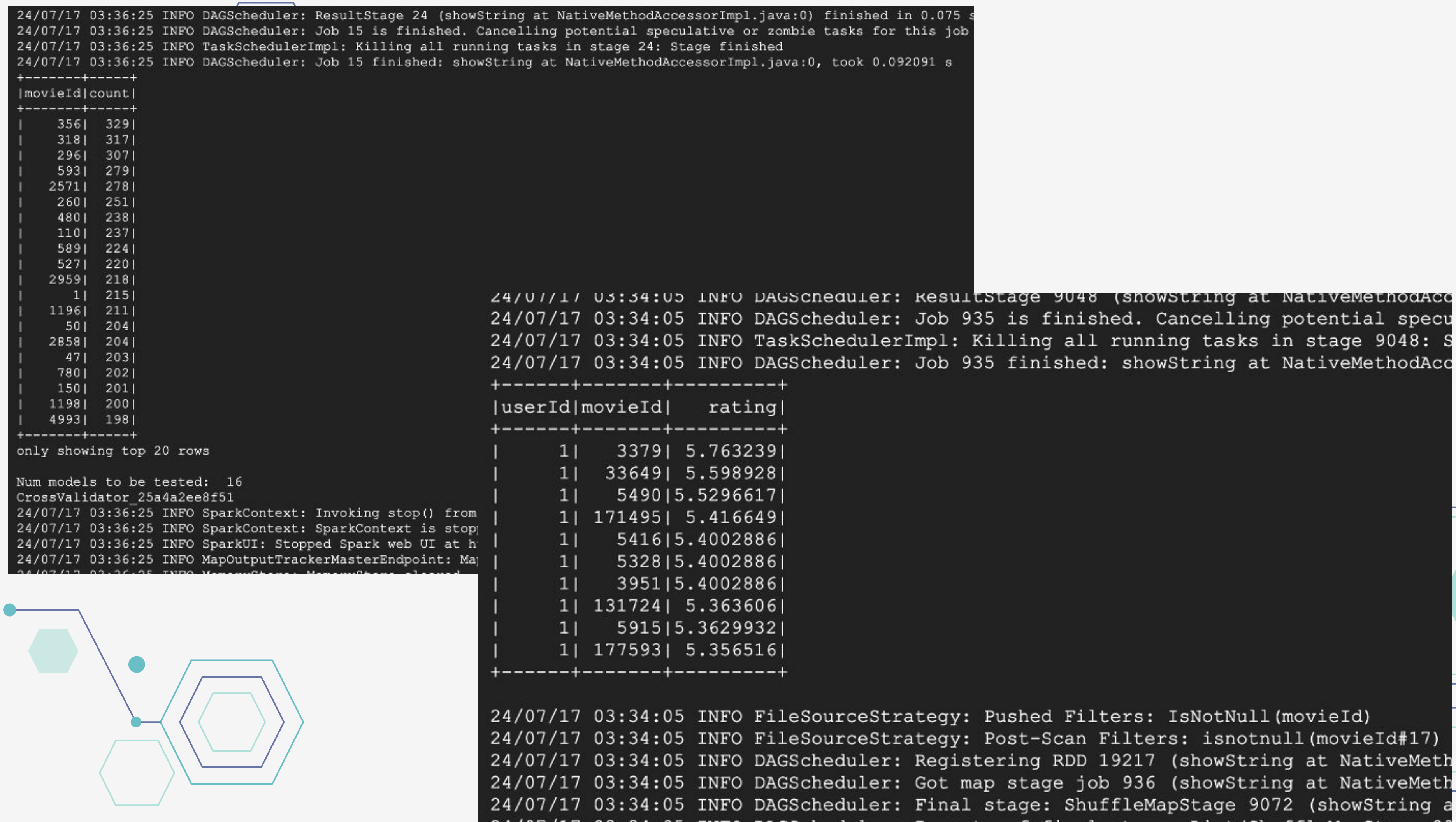
```
24/07/17 03:26:38 INFO DAGScheduler: Job 3 is finished. Cancelling potential speculative or zombie tasks for this job
24/07/17 03:26:38 INFO TaskSchedulerImpl: Killing all running tasks in stage 3: Stage finished
24/07/17 03:26:38 INFO DAGScheduler: Job 3 finished: showString at NativeMethodAccessorImpl.java:0, took 0.142916 s
24/07/17 03:26:38 INFO CodeGenerator: Code generated in 15.240608 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

24/07/17 03:26:38 INFO BlockManagerInfo: Removed broadcast_6_piece0 on cs-2199409848-default:41799 in memory (size: 34.3 KiB, free: 434.4 MiB
24/07/17 03:26:38 INFO SparkContext: Invoking stop() from shutdown hook
24/07/17 03:26:38 INFO SparkContext: SparkContext is stopping with exitCode 0.
24/07/17 03:26:38 INFO SparkUI: Stopped Spark web UI at http://cs-2199409848-default:4041
```



```
24/07/17 03:28:51 INFO CodeGenerator: Code generated in 9.590468 ms
24/07/17 03:28:51 INFO CodeGenerator: Code generated in 8.004966 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

24/07/17 03:28:51 INFO SparkContext: Invoking stop() from shutdown hook
```

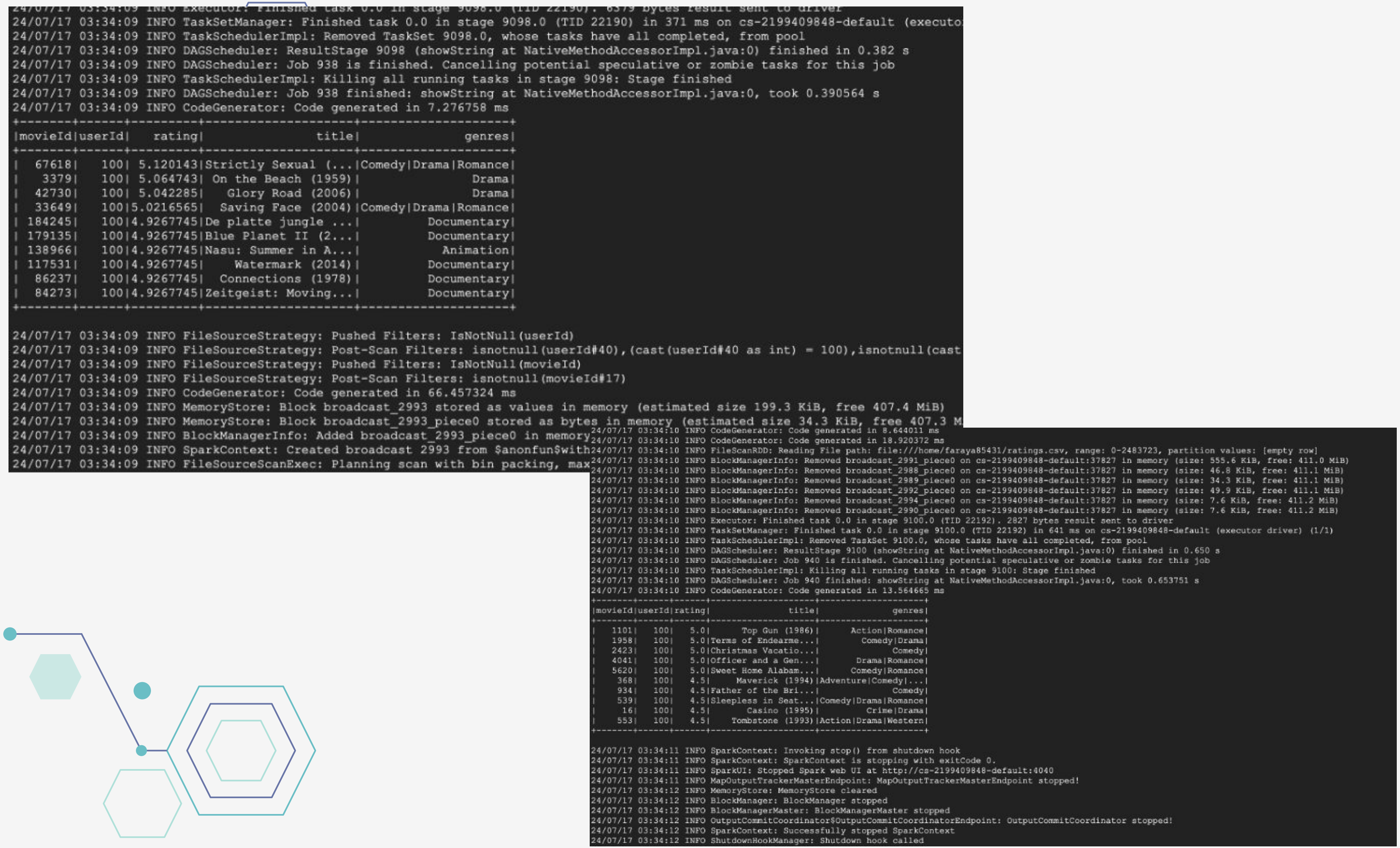


```
24/07/17 03:28:50 INFO DAGScheduler: ResultStage 18 (count at NativeMethodAccessorImpl.java:0) finished in 0.033 s
24/07/17 03:28:50 INFO DAGScheduler: Job 11 is finished. Cancelling potential speculative or zombie tasks for this job
24/07/17 03:28:50 INFO TaskSchedulerImpl: Killing all running tasks in stage 18: Stage finished
24/07/17 03:28:50 INFO DAGScheduler: Job 11 finished: count at NativeMethodAccessorImpl.java:0, took 0.048480 s
The ratings dataframe is  98.30% empty.
24/07/17 03:28:50 INFO FileSourceStrategy: Pushed Filters:
24/07/17 03:28:50 INFO FileSourceStrategy: Post-Scan Filters:
24/07/17 03:28:50 INFO CodeGenerator: Code generated in 62.327674 ms
24/07/17 03:28:50 INFO MemoryStore: Block broadcast_21 stored as values in memory (estimated size 199.3 KiB, free 433.6 MiB)
24/07/17 03:28:50 INFO MemoryStore: Block broadcast_21_piece0 stored as bytes in memory (estimated size 34.3 KiB, free 433.6 MiB)
24/07/17 03:28:50 INFO BlockManagerInfo: Added broadcast_21_piece0 in memory on cs-2199409848-default:43355 (size: 34.3 KiB, free: 4
24/07/17 03:28:50 INFO SparkContext: Created broadcast 21 from showString at NativeMethodAccessorImpl.java:0
24/07/17 03:28:50 INFO FileSourceScanExec: Planning scan with bin packing, max size: 4194304 bytes, open cost is considered as scann
24/07/17 03:28:50 INFO DAGScheduler: Registering RDD 58 (showString at NativeMethodAccessorImpl.java:0) as input to shuffle 5
24/07/17 03:28:50 INFO DAGScheduler: Got map stage job 12 (showString at NativeMethodAccessorImpl.java:0) with 1 output partitions
```



```
24/07/17 03:36:25 INFO DAGScheduler: ResultStage 24 (showString at NativeMethodAccessorImpl.java:0) finished in 0.075 s
24/07/17 03:36:25 INFO DAGScheduler: Job 15 is finished. Cancelling potential speculative or zombie tasks for this job
24/07/17 03:36:25 INFO TaskSchedulerImpl: Killing all running tasks in stage 24: Stage finished
24/07/17 03:36:25 INFO DAGScheduler: Job 15 finished: showString at NativeMethodAccessorImpl.java:0, took 0.092091 s
+-------+-----+
|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|
|     50|  204|
|   2858|  204|
|     47|  203|
|    780|  202|
|    150|  201|
|   1198|  200|
|   4993|  198|
+-------+-----+
only showing top 20 rows

Num models to be tested:  16
CrossValidator_25a4a2ee8f51
24/07/17 03:36:25 INFO SparkContext: Invoking stop() from
24/07/17 03:36:25 INFO SparkContext: SparkContext is stop
24/07/17 03:36:25 INFO SparkUI: Stopped Spark web UI at h
24/07/17 03:36:25 INFO MapOutputTrackerMasterEndpoint: Ma
24/07/17 03:36:25 INFO MemoryStore: MemoryStore cleared
```

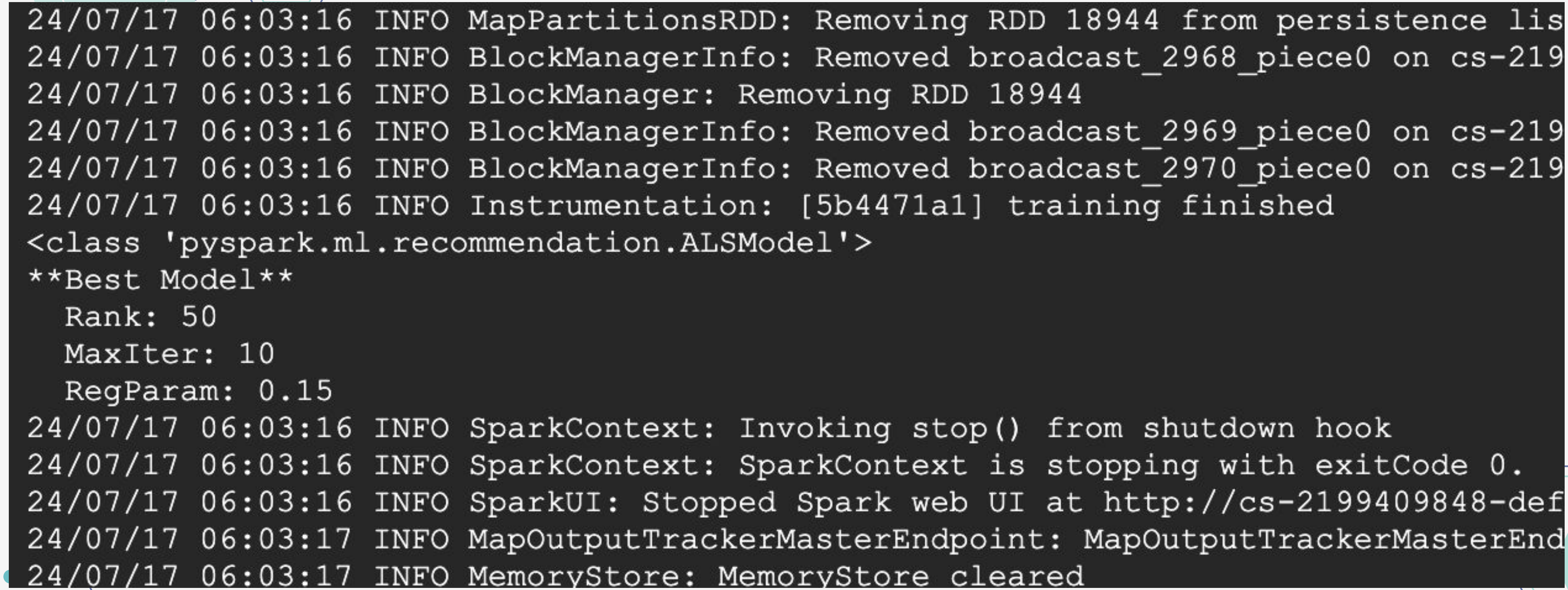
```
24/07/17 03:34:05 INFO DAGScheduler: ResultStage 9048 (showString at NativeMethodAcc
24/07/17 03:34:05 INFO DAGScheduler: Job 935 is finished. Cancelling potential specu
24/07/17 03:34:05 INFO TaskSchedulerImpl: Killing all running tasks in stage 9048: S
24/07/17 03:34:05 INFO DAGScheduler: Job 935 finished: showString at NativeMethodAcc
+------+-------+---------+
|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|
+------+-------+---------+

24/07/17 03:34:05 INFO FileSourceStrategy: Pushed Filters: IsNotNull(movieId)
24/07/17 03:34:05 INFO FileSourceStrategy: Post-Scan Filters: isnotnull(movieId#17)
24/07/17 03:34:05 INFO DAGScheduler: Registering RDD 19217 (showString at NativeMeth
24/07/17 03:34:05 INFO DAGScheduler: Got map stage job 936 (showString at NativeMeth
24/07/17 03:34:05 INFO DAGScheduler: Final stage: ShuffleMapStage 9072 (showString a
```



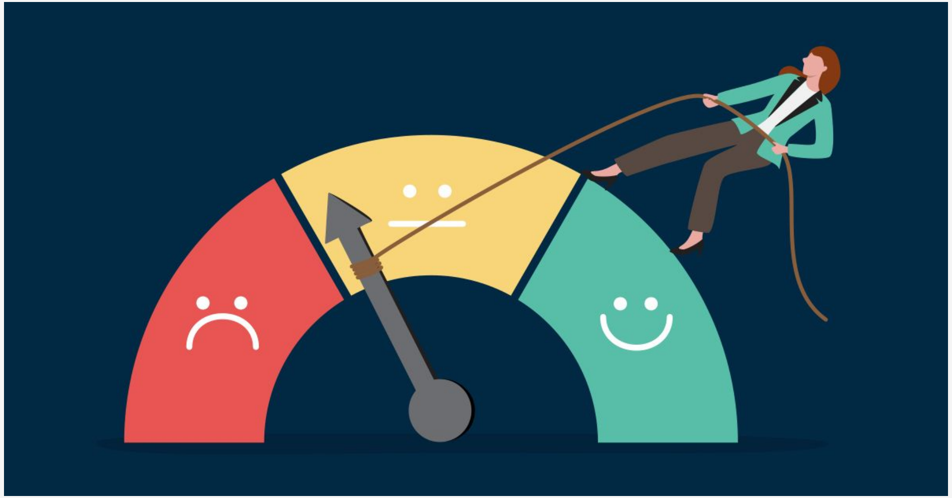
```
24/07/17 03:34:09 INFO Executor: Finished task 0.0 in stage 9098.0 (TID 22190). 6379 bytes result sent to driver
24/07/17 03:34:09 INFO TaskSetManager: Finished task 0.0 in stage 9098.0 (TID 22190) in 371 ms on cs-2199409848-default (executo
24/07/17 03:34:09 INFO TaskSchedulerImpl: Removed TaskSet 9098.0, whose tasks have all completed, from pool
24/07/17 03:34:09 INFO DAGScheduler: ResultStage 9098 (showString at NativeMethodAccessorImpl.java:0) finished in 0.382 s
24/07/17 03:34:09 INFO DAGScheduler: Job 938 is finished. Cancelling potential speculative or zombie tasks for this job
24/07/17 03:34:09 INFO TaskSchedulerImpl: Killing all running tasks in stage 9098: Stage finished
24/07/17 03:34:09 INFO DAGScheduler: Job 938 finished: showString at NativeMethodAccessorImpl.java:0, took 0.390564 s
24/07/17 03:34:09 INFO CodeGenerator: Code generated in 7.276758 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|
| 179135|   100|4.9267745|Blue Planet II (2...|         Documentary|
| 138966|   100|4.9267745|Nasu: Summer in A...|           Animation|
| 117531|   100|4.9267745|     Watermark (2014)|         Documentary|
|  86237|   100|4.9267745|   Connections (1978)|         Documentary|
|  84273|   100|4.9267745|Zeitgeist: Moving...|         Documentary|
+-------+------+---------+--------------------+--------------------+

24/07/17 03:34:09 INFO FileSourceStrategy: Pushed Filters: IsNotNull(userId)
24/07/17 03:34:09 INFO FileSourceStrategy: Post-Scan Filters: isnotnull(userId#40),(cast(userId#40 as int) = 100),isnotnull(cast
24/07/17 03:34:09 INFO FileSourceStrategy: Pushed Filters: IsNotNull(movieId)
24/07/17 03:34:09 INFO FileSourceStrategy: Post-Scan Filters: isnotnull(movieId#17)
24/07/17 03:34:09 INFO CodeGenerator: Code generated in 66.457324 ms
24/07/17 03:34:09 INFO MemoryStore: Block broadcast_2993 stored as values in memory (estimated size 199.3 KiB, free 407.4 MiB)
24/07/17 03:34:09 INFO MemoryStore: Block broadcast_2993_piece0 stored as bytes in memory (estimated size 34.3 KiB, free 407.3 M
24/07/17 03:34:09 INFO BlockManagerInfo: Added broadcast_2993_piece0 in memory
24/07/17 03:34:09 INFO SparkContext: Created broadcast 2993 from $anonfun$with
24/07/17 03:34:09 INFO FileSourceScanExec: Planning scan with bin packing, max
```

```
24/07/17 03:34:10 INFO CodeGenerator: Code generated in 8.644011 ms
24/07/17 03:34:10 INFO CodeGenerator: Code generated in 18.920372 ms
24/07/17 03:34:10 INFO FileScanRDD: Reading File path: file:///home/faraya85431/ratings.csv, range: 0-2483723, partition values: [empty row]
24/07/17 03:34:10 INFO BlockManagerInfo: Removed broadcast_2991_piece0 on cs-2199409848-default:37827 in memory (size: 555.6 KiB, free: 411.0 MiB)
24/07/17 03:34:10 INFO BlockManagerInfo: Removed broadcast_2988_piece0 on cs-2199409848-default:37827 in memory (size: 46.8 KiB, free: 411.1 MiB)
24/07/17 03:34:10 INFO BlockManagerInfo: Removed broadcast_2989_piece0 on cs-2199409848-default:37827 in memory (size: 34.3 KiB, free: 411.1 MiB)
24/07/17 03:34:10 INFO BlockManagerInfo: Removed broadcast_2992_piece0 on cs-2199409848-default:37827 in memory (size: 49.9 KiB, free: 411.1 MiB)
24/07/17 03:34:10 INFO BlockManagerInfo: Removed broadcast_2994_piece0 on cs-2199409848-default:37827 in memory (size: 7.6 KiB, free: 411.2 MiB)
24/07/17 03:34:10 INFO BlockManagerInfo: Removed broadcast_2990_piece0 on cs-2199409848-default:37827 in memory (size: 7.6 KiB, free: 411.2 MiB)
24/07/17 03:34:10 INFO Executor: Finished task 0.0 in stage 9100.0 (TID 22192). 2827 bytes result sent to driver
24/07/17 03:34:10 INFO TaskSetManager: Finished task 0.0 in stage 9100.0 (TID 22192) in 641 ms on cs-2199409848-default (executor driver) (1/1)
24/07/17 03:34:10 INFO TaskSchedulerImpl: Removed TaskSet 9100.0, whose tasks have all completed, from pool
24/07/17 03:34:10 INFO DAGScheduler: ResultStage 9100 (showString at NativeMethodAccessorImpl.java:0) finished in 0.650 s
24/07/17 03:34:10 INFO DAGScheduler: Job 940 is finished. Cancelling potential speculative or zombie tasks for this job
24/07/17 03:34:10 INFO TaskSchedulerImpl: Killing all running tasks in stage 9100: Stage finished
24/07/17 03:34:10 INFO DAGScheduler: Job 940 finished: showString at NativeMethodAccessorImpl.java:0, took 0.653751 s
24/07/17 03:34:10 INFO CodeGenerator: Code generated in 13.564665 ms
+-------+------+------+--------------------+--------------------+
|movieId|userId|rating|               title|              genres|
+-------+------+------+--------------------+--------------------+
|   1101|   100|   5.0|     Top Gun (1986)|     Action|Romance|
|   1958|   100|   5.0|Terms of Endearme...|        Comedy|Drama|
|   2423|   100|   5.0|Christmas Vacatio...|              Comedy|
|   4041|   100|   5.0|Officer and a Gen...|       Drama|Romance|
|   5620|   100|   5.0|Sweet Home Alabam...|      Comedy|Romance|
|    368|   100|   4.5|    Maverick (1994)|Adventure|Comedy|...|
|    934|   100|   4.5|Father of the Bri...|              Comedy|
|    539|   100|   4.5|Sleepless in Seat...|Comedy|Drama|Romance|
|     16|   100|   4.5|      Casino (1995)|         Crime|Drama|
|    553|   100|   4.5|   Tombstone (1993)|Action|Drama|Western|
+-------+------+------+--------------------+--------------------+

24/07/17 03:34:11 INFO SparkContext: Invoking stop() from shutdown hook
24/07/17 03:34:11 INFO SparkContext: SparkContext is stopping with exitCode 0.
24/07/17 03:34:11 INFO SparkUI: Stopped Spark web UI at http://cs-2199409848-default:4040
24/07/17 03:34:11 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
24/07/17 03:34:12 INFO MemoryStore: MemoryStore cleared
24/07/17 03:34:12 INFO BlockManager: BlockManager stopped
24/07/17 03:34:12 INFO BlockManagerMaster: BlockManagerMaster stopped
24/07/17 03:34:12 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
24/07/17 03:34:12 INFO SparkContext: Successfully stopped SparkContext
24/07/17 03:34:12 INFO ShutdownHookManager: Shutdown hook called
```

Result



```
24/07/17 06:03:16 INFO MapPartitionsRDD: Removing RDD 18944 from persistence lis
24/07/17 06:03:16 INFO BlockManagerInfo: Removed broadcast_2968_piece0 on cs-219
24/07/17 06:03:16 INFO BlockManager: Removing RDD 18944
24/07/17 06:03:16 INFO BlockManagerInfo: Removed broadcast_2969_piece0 on cs-219
24/07/17 06:03:16 INFO BlockManagerInfo: Removed broadcast_2970_piece0 on cs-219
24/07/17 06:03:16 INFO Instrumentation: [5b4471a1] training finished
<class 'pyspark.ml.recommendation.ALSModel'>
**Best Model**
  Rank: 50
  MaxIter: 10
  RegParam: 0.15
24/07/17 06:03:16 INFO SparkContext: Invoking stop() from shutdown hook
24/07/17 06:03:16 INFO SparkContext: SparkContext is stopping with exitCode 0.
24/07/17 06:03:16 INFO SparkUI: Stopped Spark web UI at http://cs-2199409848-def
24/07/17 06:03:17 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEnd
24/07/17 06:03:17 INFO MemoryStore: MemoryStore cleared
```



05

Enhancements

Can we get better result?

- Incorporate additional data (e.g., user demographics)
- Use hybrid recommendation models
- Experiment with different machine learning algorithms
- Implement real-time recommendation updates

06

Conclusion

Key takeaways:

- Importance of data preprocessing
- Benefits of hyperparameter tuning
- Real-world applications of recommendation engines

07

References

How to Build a Movie Recommendation System Based on Collaborative Filtering

Movie Recommendation with Collaborative Filtering in Pyspark

GitHub

Thanks!

CREDITS: This presentation template was created by **Slidesgo**, and includes icons by **Flaticon**, and infographics & images by **Freepik**

Please keep this slide for attribution