Project: Movie Recommendation with MLlib - Collaborative Filtering - RDD-based API

Jisen Fang

Table of contents

Key Technologies

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

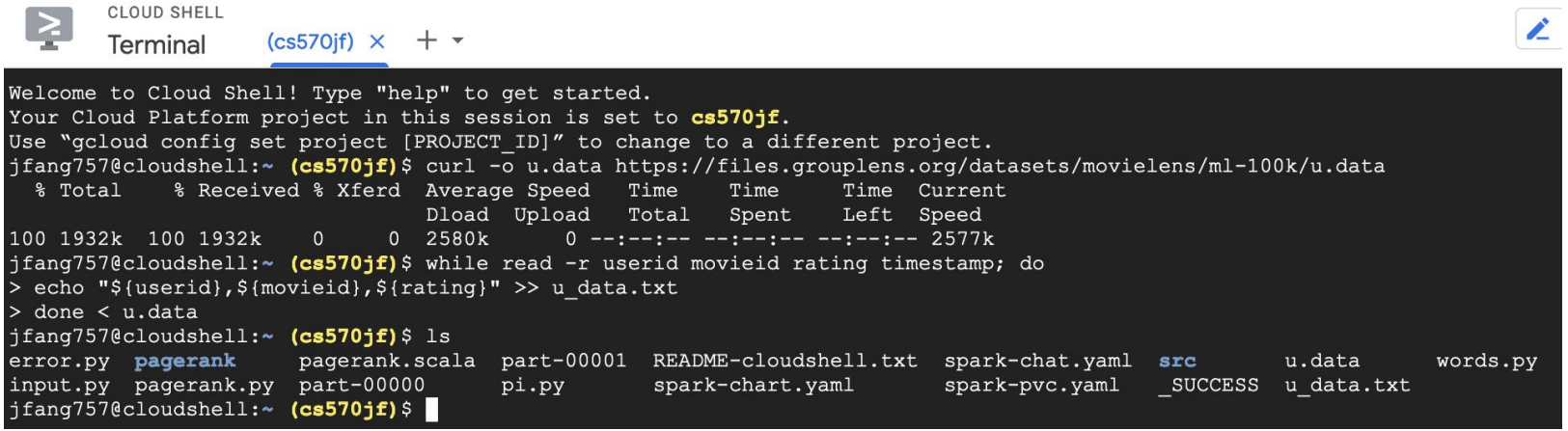
Collaborative filtering is commonly used for recommender systems. These techniques aim to fill in the missing entries of a user-item association matrix.

GCP: Google Cloud Platform is one of the major Cloud Computing Platforms. It consists of a set of physical assets, such as computers and hard disk drives, and virtual resources, such as virtual machines (VMs), that are contained in Google's data centers around the globe.

PySpark: PySpark is the Python API for Apache Spark. Python is an interpreted, object-oriented, high-level programming language along with dynamic typing and dynamic binding.

Create an u_data.txt

Download MoveLens' data and convert into the format of (UserID, MovieID, rating)



Create Bucket and Cluster on GCP

Create a Bucket and Dataproc cluster on GCP, and then move the u_data.txt to the bucket

```
jfang757@cloudshell:~ (cs570jf)$ ls
error.py  movie_recommendation.py  pagerank.py     part-00000  pi.py                  spark-chart.yaml  spark-pvc.yaml  _SUCCESS  u_data.txt
input.py  pagerank                 pagerank.scala  part-00001  README-cloudshell.txt  spark-chat.yaml   src             u.data    words.py
jfang757@cloudshell:~ (cs570jf)$ gsutil cp u_data.txt gs://w9h1/input/
Copying file://u_data.txt [Content-Type=text/plain]...
- [1 files][956.2 KiB/956.2 KiB]
Operation completed over 1 objects/956.2 KiB.
jfang757@cloudshell:~ (cs570jf)$
```

PySpark Code

```
import pyspark
import sys
from pyspark.mllib.recommendation import ALS, Rating

if len(sys.argv) != 3:
  raise Exception("Exactly 1 arguments are required: <inputUri>")
inputUri = sys.argv[1]
outputUri = sys.argv[2]

# Load and parse the data
sc = pyspark.SparkContext()
data = sc.textFile(inputUri)
ratings = data.map(lambda l: l.split(','))\
       .map(lambda l: Rating(int(l[0]), int(l[1]), float(l[2])))
```

```
# Build the recommendation model using Alternating Least Squares
rank = 10
numIterations = 10
model = ALS.train(ratings, rank, numIterations)

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

# Save model
model.save(sc, outputUri)
```

PySpark Execution

In Cloud Shell Terminal, type:

```
gcloud dataproc jobs submit pyspark movie_recommendation.py --cluster=w9h1
--region=us-west2 -- gs://w9h1/input/u_data.txt gs://w9h1/output
```

```
jfang757@cloudshell:~ (cs570jf)$ gcloud dataproc jobs submit pyspark movie_recommendation.py --cluster=w9h1 --region=us-west2 -- gs://w9h1/input/u_data.txt gs://w9h1/output
Job [21b14e5b71ec481f93eada2c10d202df] submitted.
Waiting for job output...
23/07/19 08:24:24 INFO SparkEnv: Registering MapOutputTracker
23/07/19 08:24:24 INFO SparkEnv: Registering BlockManagerMaster
23/07/19 08:24:24 INFO SparkEnv: Registering BlockManagerMasterHeartbeat
23/07/19 08:24:24 INFO SparkEnv: Registering OutputCommitCoordinator
23/07/19 08:24:25 INFO DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at w9h1-m.us-west2-b.c.cs570jf.internal./10.168.0.20:8032
23/07/19 08:24:25 INFO AHSProxy: Connecting to Application History server at w9h1-m.us-west2-b.c.cs570jf.internal./10.168.0.20:10200
23/07/19 08:24:26 INFO Configuration: resource-types.xml not found
23/07/19 08:24:26 INFO ResourceUtils: Unable to find 'resource-types.xml'.
23/07/19 08:24:27 INFO YarnClientImpl: Submitted application application_1689752313363_0006
23/07/19 08:24:28 INFO DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at w9h1-m.us-west2-b.c.cs570jf.internal./10.168.0.20:8030
23/07/19 08:24:30 WARN GhfsStorageStatistics: Detected potential high latency for operation op_get_file_status. latencyMs=502; previousMaxLatencyMs=0; operationCount=1; context=gs:
//dataproc-temp-us-west2-66473074362-tdp6glrs/9aadff71-ebe5-430b-bd12-975294bb5231/spark-job-history
23/07/19 08:24:30 INFO GoogleCloudStorageImpl: Ignoring exception of type GoogleJsonResponseException; verified object already exists with desired state.
23/07/19 08:24:30 WARN GhfsStorageStatistics: Detected potential high latency for operation op_mkdirs. latencyMs=337; previousMaxLatencyMs=0; operationCount=1; context=gs://datapro
c-temp-us-west2-66473074362-tdp6glrs/9aadff71-ebe5-430b-bd12-975294bb5231/spark-job-history
23/07/19 08:24:31 WARN GhfsStorageStatistics: Detected potential high latency for operation op_create. latencyMs=431; previousMaxLatencyMs=0; operationCount=1; context=gs://datapro
c-temp-us-west2-66473074362-tdp6glrs/9aadff71-ebe5-430b-bd12-975294bb5231/spark-job-history/application_1689752313363_0006.inprogress
23/07/19 08:24:33 WARN GhfsStorageStatistics: Detected potential high latency for operation op_glob_status. latencyMs=120; previousMaxLatencyMs=0; operationCount=1; context=path=gs
://w9h1/input/u_data.txt; pattern=org.apache.hadoop.mapred.FileInputFormat$MultiPathFilter@e217a7f
23/07/19 08:24:33 INFO FileInputFormat: Total input files to process : 1
Mean Squared Error = 0.48417075722935254
23/07/19 08:25:05 INFO GoogleCloudStorageFileSystem: Successfully repaired 'gs://w9h1/output/metadata/' directory.
23/07/19 08:25:05 WARN GhfsStorageStatistics: Detected potential high latency for operation op_delete. latencyMs=126; previousMaxLatencyMs=0; operationCount=1; context=gs://w9h1/ou
tput/metadata/_temporary
23/07/19 08:25:05 WARN GhfsStorageStatistics: Detected potential high latency for operation stream_write_close_operations. latencyMs=225; previousMaxLatencyMs=0; operationCount=1;
context=gs://w9h1/output/metadata/_SUCCESS
23/07/19 08:25:13 INFO GoogleCloudStorageFileSystem: Successfully repaired 'gs://w9h1/output/data/user/' directory.
23/07/19 08:25:14 WARN GhfsStorageStatistics: Detected potential high latency for operation stream_write_close_operations. latencyMs=506; previousMaxLatencyMs=225; operationCount=2
; context=gs://w9h1/output/data/user/_SUCCESS
23/07/19 08:25:15 INFO GoogleCloudStorageFileSystem: Successfully repaired 'gs://w9h1/output/data/product/' directory.
Job [21b14e5b71ec481f93eada2c10d202df] finished successfully.
done: true
driverControlFilesUri: gs://dataproc-staging-us-west2-66473074362-ei6hp88p/google-cloud-dataproc-metainfo/9aadff71-ebe5-430b-bd12-975294bb5231/jobs/21b14e5b71ec481f93eada2c10d202df
/
driverOutputResourceUri: gs://dataproc-staging-us-west2-66473074362-ei6hp88p/google-cloud-dataproc-metainfo/9aadff71-ebe5-430b-bd12-975294bb5231/jobs/21b14e5b71ec481f93eada2c10d202
df/driveroutput
jobUuid: d90a8fe1-0713-35dc-b0e7-563a04c162df
placement:
  clusterName: w9h1
  clusterUuid: 9aadff71-ebe5-430b-bd12-975294bb5231
pysparkJob:
```

```
jfang757@cloudshell:~ (cs570jf)$ gcloud dataproc jobs submit pyspark movie_recommendation.py --cluster=w9h1 --region=us-west2 -- gs://w9h1/input/u_data.txt gs://w9h1/output
Job [21b14e5b71ec481f93eada2c10d202df] submitted.
Waiting for job output...
23/07/19 08:24:24 INFO SparkEnv: Registering MapOutputTracker
23/07/19 08:24:24 INFO SparkEnv: Registering BlockManagerMaster
23/07/19 08:24:24 INFO SparkEnv: Registering BlockManagerMasterHeartbeat
23/07/19 08:24:24 INFO SparkEnv: Registering OutputCommitCoordinator
23/07/19 08:24:25 INFO DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at w9h1-m.us-west2-b.c.cs570jf.internal./10.168.0.20:8032
23/07/19 08:24:25 INFO AHSProxy: Connecting to Application History server at w9h1-m.us-west2-b.c.cs570jf.internal./10.168.0.20:10200
23/07/19 08:24:26 INFO Configuration: resource-types.xml not found
23/07/19 08:24:26 INFO ResourceUtils: Unable to find 'resource-types.xml'.
23/07/19 08:24:27 INFO YarnClientImpl: Submitted application application_1689752313363_0006
23/07/19 08:24:28 INFO DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at w9h1-m.us-west2-b.c.cs570jf.internal./10.168.0.20:8030
23/07/19 08:24:30 WARN GhfsStorageStatistics: Detected potential high latency for operation op_get_file_status. latencyMs=502; previousMaxLatencyMs=0; operationCount=1; context=gs:
//dataproc-temp-us-west2-66473074362-tdp6glrs/9aadff71-ebe5-430b-bd12-975294bb5231/spark-job-history
23/07/19 08:24:30 INFO GoogleCloudStorageImpl: Ignoring exception of type GoogleJsonResponseException; verified object already exists with desired state.
23/07/19 08:24:30 WARN GhfsStorageStatistics: Detected potential high latency for operation op_mkdirs. latencyMs=337; previousMaxLatencyMs=0; operationCount=1; context=gs://datapro
c-temp-us-west2-66473074362-tdp6glrs/9aadff71-ebe5-430b-bd12-975294bb5231/spark-job-history
23/07/19 08:24:31 WARN GhfsStorageStatistics: Detected potential high latency for operation op_create. latencyMs=431; previousMaxLatencyMs=0; operationCount=1; context=gs://datapro
c-temp-us-west2-66473074362-tdp6glrs/9aadff71-ebe5-430b-bd12-975294bb5231/spark-job-history/application_1689752313363_0006.inprogress
23/07/19 08:24:33 WARN GhfsStorageStatistics: Detected potential high latency for operation op_glob_status. latencyMs=120; previousMaxLatencyMs=0; operationCount=1; context=path=gs
://w9h1/input/u_data.txt; pattern=org.apache.hadoop.mapred.FileInputFormat$MultiPathFilter@e217a7f
23/07/19 08:24:33 INFO FileInputFormat: Total input files to process : 1
Mean Squared Error = 0.48417075722935254
23/07/19 08:25:05 INFO GoogleCloudStorageFileSystem: Successfully repaired 'gs://w9h1/output/metadata/' directory.
23/07/19 08:25:05 WARN GhfsStorageStatistics: Detected potential high latency for operation op_delete. latencyMs=126; previousMaxLatencyMs=0; operationCount=1; context=gs://w9h1/ou
tput/metadata/_temporary
23/07/19 08:25:05 WARN GhfsStorageStatistics: Detected potential high latency for operation stream_write_close_operations. latencyMs=225; previousMaxLatencyMs=0; operationCount=1;
context=gs://w9h1/output/metadata/_SUCCESS
23/07/19 08:25:13 INFO GoogleCloudStorageFileSystem: Successfully repaired 'gs://w9h1/output/data/user/' directory.
23/07/19 08:25:14 WARN GhfsStorageStatistics: Detected potential high latency for operation stream_write_close_operations. latencyMs=506; previousMaxLatencyMs=225; operationCount=2
; context=gs://w9h1/output/data/user/_SUCCESS
23/07/19 08:25:15 INFO GoogleCloudStorageFileSystem: Successfully repaired 'gs://w9h1/output/data/product/' directory.
Job [21b14e5b71ec481f93eada2c10d202df] finished successfully.
done: true
driverControlFilesUri: gs://dataproc-staging-us-west2-66473074362-ei6hp88p/google-cloud-dataproc-metainfo/9aadff71-ebe5-430b-bd12-975294bb5231/jobs/21b14e5b71ec481f93eada2c10d202df
/
driverOutputResourceUri: gs://dataproc-staging-us-west2-66473074362-ei6hp88p/google-cloud-dataproc-metainfo/9aadff71-ebe5-430b-bd12-975294bb5231/jobs/21b14e5b71ec481f93eada2c10d202
df/driveroutput
jobUuid: d90a8fe1-0713-35dc-b0e7-563a04c162df
placement:
  clusterName: w9h1
  clusterUuid: 9aadff71-ebe5-430b-bd12-975294bb5231
pysparkJob:
```