14 JUNE 2016

Spark Use Case – Analyzing MovieLens Dataset

In this blog, we will discuss a use case involving MovieLens dataset and try to analyze how the movies fare on a rating scale of 1 to 5.

We will start our discussion with the data definition by considering a sample of four records.

196	242	3
186	302	3

Column 1: User ID

Column 2: Movie ID

Column 3: Rating

Column 4: Timestamp

You can download the input file from here.

Below is the code that is used to calculate the number of movies that are rated on a scale of 1 to 5.

```
1  from pyspark import SparkConf,SparkContext
2
3  import collections
4
5  my_lines = sc.textFile('/home/acadgild/ml-100k/u.data')
6
7  ratings = lines.map(lambda x : x.split()[2])
8
9  res = ratings.countByValue()
10
11 my_sortedres = collections.OrderedDict(sorted(res.items()))
12
13 for key,value in sortedres.items():
14
15 print ("%s %i" %(key,value))
```

The first two lines of the code import SparkConf ,SparkContext from pyspark libraries that are present in Spark.

SparkContext is the fundamental starting point in Spark that enables us to create RDDs.

Once the collections are imported [illegible] be sorted out.

1/60 Next»

Looking for more videos like this? Click here!

The statement in the screenshot below, loads the data file by creating the RDD through sc.textFile method. The data file is loaded into RDD my_lines and the textFile property breaks every line of text into a value in the RDD.

```
>>> my_lines = sc.textFile('/home/acadgild/ml-100k/u.data')
16/06/10 17:42:25 INFO MemoryStore: ensureFreeSpace(230688) called with curMem=1
792827, maxMem=560497950
16/06/10 17:42:25 INFO MemoryStore: Block broadcast_19 stored as values in memor
y (estimated size 225.3 KB, free 532.6 MB)
```

In the screenshot below, expression X is passed on to split function. The 3rd column is extracted and new RDD is created with the new results.

```
>>> ratings = my_lines.map(lambda x : x.split()[2])
>>>
```

In the screenshot below, the statement countByValue() is executed on ratings RDD to calculate the occurrence of each and every rating starting from 1 to 5 and results are stored in a new Python object *res.*

```
>>> res = ratings.countByValue()
16/06/10 17:45:05 INFO FileInputFormat: Total input paths to process : 1
16/06/10 17:45:05 INFO SparkContext: Starting job: countByValue at <stdin>:1
16/06/10 17:45:05 INFO DAGScheduler: Got job 11 (countByValue at <stdin>:1) with
 1 output partitions
16/06/10 17:45:05 INFO DAGScheduler: Final stage: ResultStage 11(countByValue at
 <stdin>:1)
```

Below is the sample output of the above action applied on ratings RDD.

```
>>> res
defaultdict(<class 'int'>, {'4': 34174, '1': 6110, '5': 21201, '3': 27145, '2':
11370})
>>>
```

The code below creates an ordered Dictionary to sort the results based on the key which is rating.

```
>>> my_sortedres = collections.Ordere
>>>
```

The Python code below iterates th

- AcadGild
- Android App Development
- Big Data Hadoop & Spark
- Big Data Hadoop & Spark – Advanced
- Blockchain
- Careers
- Data Analytics with R, Excel & Tableau
- Data Science and Artificial Intelligence
- Full stack Web Development
- Graphic Design & UX

GET SOCIAL

```
...
1 6110
2 11370
3 27145
4 34174
5 21201
>>>
```

We hope this blog was helpful in understanding the use of Spark with Python. Keep visiting our site www.acadgild.com for more updates on Big data and other technologies.

Share this:

Related

Spark Use Case - Popular Movie Analysis
March 3, 2018
In "Big Data Hadoop & Spark"

MapReduce Custom Partitioner
May 18, 2016
In "Big Data Hadoop & Spark"

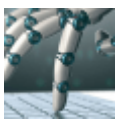
Spark Use Case - Youtube Data Analysis
April 5, 2016
In "Big Data Hadoop & Spark"

SATYAM

Satyam Kumar is a Big Da
in AcadGild with rich exp
technologies like Hadoop
related technologies. He strives to code in

WHAT'S TRENDING

RECENT POSTS



Data Science Glossary- Machine Learning Tools and Terminologies

With Python

what sets us apart is earning opportunities we provide after successful completion of course. We also provide live mentoring and 24x7 support. Our mentors are industry thought leaders in their respective fields.

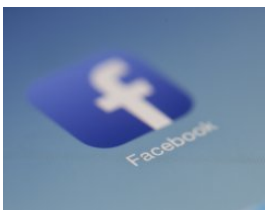
← PREVIOUS ARTICLE

Introduction to Backbone.JS

NEXT ARTICLE →

Custom Input Format in Hadoop

RELATED POSTS



Data Breach- A Threat To Privacy and The Famous Facebook-Cambridge Analytica Incident

April 16, 2018

Partitioning In Hive

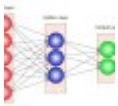
March 6, 2018

Testing your Scripts with PigUnit

March 5, 2018

Data Breaches like Facebook-Cambridge Analytica Scandal

April 27, 2018



Implementation Of Recurrent Neural Network

April 27, 2018

ARCHIVES

- May 2018
- April 2018
- March 2018
- February 2018
- January 2018
- December 2017
- November 2017
- October 2017
- September 2017
- August 2017

hi,

Any sample projects which use Scala instead of python.

Thanks,
Rajesh

LEAVE A REPLY

COMMENTS *

NAME * **EMAIL** * **WEBSITE**

SUBMIT

NOTIFY ME OF FOLLOW-UP COMMENTS BY EMAIL.

NOTIFY ME OF NEW POSTS BY EMAIL.

- March 2017
- February 2017
- January 2017
- December 2016
- November 2016
- October 2016
- September 2016
- August 2016
- July 2016
- June 2016
- May 2016
- April 2016
- March 2016
- February 2016
- January 2016
- December 2015
- November 2015
- August 2015

- August 2014

CATEGORIES

- AcadGild
- Android App Development
- Big Data Hadoop & Spark
- Big Data Hadoop & Spark – Advanced
- Blockchain
- Careers
- Data Analytics with R, Excel & Tableau
- Data Science and Artificial Intelligence
- Full stack Web Development
- Graphic Design & UX

TAGS

ACADGILD

ACADGILD ONLINE COURSES

ANDROID

ANDROID APP

ANDROID APP DEVELOPMENT

ANDROID DEVELOPMENT

ANDROID DEVELOPMENT COURSE

APACHE HIVE

APACHE PIG

APACHE SPARK

ARTIFICIAL INTELLIGENCE

BIG DATA

BIG DATA AND HADOOP

BIG DATA AND HADOOP ON COURSES

BIG DATA

LIKE WHAT YOU SEE? SUBSCRIBE TO OUR BLOG

We send only 1 email in a week

Enter your email...

Subscribe