ACADGILD
Learn. Do. Earn.

CATEGORIES ▾

□ Home / Spark / Spark Use Case – The Daily Show

16
APRIL
2016

Spark Use Case – The Daily Show

In this blog we will be taking a famous Tv show dataset i.e., The Daily show and we will be performing analysis on the guests who came to the show.

Before going ahead we recommend readers to go through our previous blogs on various publicly available datasets.

Youtube Data Analysis

Titanic Data Analysis

Olympic Data Analysis

We have a historical data of the daily show guests from 1999 to 2004. The dataset can be downloaded from here.

Please find the the dataset description below:

Dataset Description:

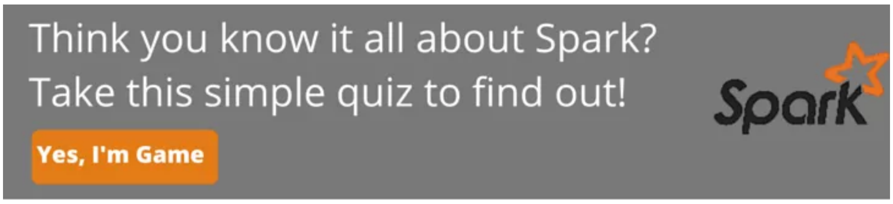
YEAR – The year the episode aired

GoogleKnowlege_Occupation -Their occupation or office, according to Google's Knowledge Graph or, if they're not in there, how Stewart introduced them on the program.

Show – Air date of episode. Not unique, as some shows had more than one guest

Group – A larger group designation for the occupation. For instance, us senators, us presidents, and former presidents are all under "politicians"

Raw_Guest_List – The person or list of people who appeared on the show, according to Wikipedia. The GoogleKnowlege_Occupation only refers to one of them in a given row.



Problem Statement:

Find the top 5 kinds of GoogleKnowlege_Occupation people gusted the show in a particular time period.

Source Code:

```
val file = sc.textFile("/home/kiran/dialy_show_guests")
val split = file.map(line => line.split(","))
val format = new java.text.SimpleDateFormat("MM/dd/yy")
val pair = split.map(line => (line(1),format.parse(line(2))))
val fil = pair.filter(x => {if(x._2.after(format.parse("1/11/99")) && x._2.before(format.parse("6/11/99"))) true else false})
val cnt = fil.map(x => (x._1,1)).reduceByKey(_+_).map(item => item.swap).sortByKey(false).take(5)
```

SEARCH

Search Now

CATEGORIES

- AcadGild
- Android
- Android For Kids
- AngularJS
- Big Data and Hadoop
- Careers
- Cloud computing
- Database
- Digital Marketing
- Front End
- Full Stack
- Hadoop Administration
- IOS
- Java
- Kids
- Linux Administration
- NodeJS
- Others
- Python
- Quiz

Walk through of the above code:

In **line 1** we are creating a new RDD by loading the dataset which is in local file system.

In **line 2** we are splitting the records by using the delimiter as ‘,’ since the data is delimited by ‘,’.

In **line 3** we are declaring the date format by using the java library java.text.SimpleDateFormat. In the dataset the data format is “MM/dd/YY”.

In **line 4** we are creating a pair of GoogleKnowlege_Occupation and Show(date of the show). Here date of the show is taken as a string and we are converting this string to date format using the *parse* method available in java.text.SimpleDateFormat.

In **line 5** we are using the filter method to filter out the records which doesn’t match our requirement. Here we are giving the range of data explicitly in between we need to count the GoogleKnowlege_Occupation people gusted. Here we have given the range as 6 months i.e., from 1/11/99 to 6/11/99.

In **line 6** we will get the data which is in specified range from that we are creating a pair of *GoogleKnowlege_Occupation* and *1* as key value pairs respectively. After that we are performing reduceByKey action on the RDD which will count all the values for each unique key. Then we are swapping the *GoogleKnowlege_Occupation* and its *count*, and sorting the result by *sortByKey* operation with this we will get the sorted records of *GoogleKnowlege_Occupation* and its *count* in descending order. Finally, we are taking the top five from the list.

Output:

(28,actor), (20,actress), (4,comedian), (3,television actress), (2,stand-up comedian)

The same is displayed in the below screen shot.

- R & Machine Learning
- Scala
- Spark
- Uncategorized

GET SOCIAL

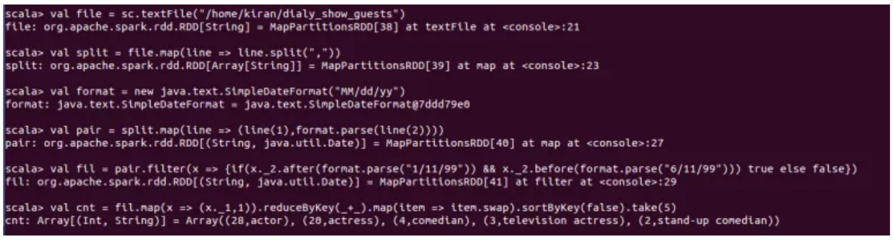
AcadGild
Local Business · Bangal
92,286 likes

3 friends like this

AcadGild shared a li
1 hr

Spark Use Case – U
In this post, we will be perforn
ACADGILD.COM

WHAT’S TRENDING



```
scala> val file = sc.textFile("/home/kiran/dialy_show_guests")
file: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[38] at textFile at <console>:21

scala> val split = file.map(line => line.split(","))
split: org.apache.spark.rdd.RDD[Array[String]] = MapPartitionsRDD[39] at map at <console>:23

scala> val format = new java.text.SimpleDateFormat("MM/dd/yy")
format: java.text.SimpleDateFormat = java.text.SimpleDateFormat@7ddd79e0

scala> val pair = split.map(line => (line(1),format.parse(line(2))))
pair: org.apache.spark.rdd.RDD[(String, java.util.Date)] = MapPartitionsRDD[40] at map at <console>:27

scala> val fil = pair.filter(x => {if(x._2.after(format.parse("1/11/99")) && x._2.before(format.parse("6/11/99"))) true else false})
fil: org.apache.spark.rdd.RDD[(String, java.util.Date)] = MapPartitionsRDD[41] at filter at <console>:29

scala> val cnt = fil.map(x => (x._1,1)).reduceByKey(_+_).map(item => item.swap).sortByKey(false).take(5)
cnt: Array[(Int, String)] = Array((28,actor), (20,actress), (4,comedian), (3,television actress), (2,stand-up comedian))
```

Hope this blog helped you in understanding how to perform analysis data using apache spark in scala with a real time dataset. Keep visiting our site for more updates on Big Data and other technologies.

Spark
Learn SPARK from our Expert Mentors in just 12 weeks and Boost your Career
Enroll Today

Share this:

57 2

Related

Analyzing New York Crime Data Using SparkSQL
April 28, 2016
In "Spark"

Beginner's Guide for Hive
December 18, 2015
In "Big Data and Hadoop"

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

KIRAN KRISHNA

Kiran Krishna Innamuri is a Passionate Big Data enthusiast with 2 + years of experience in Hadoop and Spark Development. He is a passionate Java and scala programmer. AcadGild was founded with the vision of "Learn. Do. Earn". We provide skill development courses based on current industry needs. But 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. We provide courses for Android Programming, Big Data,

Tweets by @acadgild

ACADGILD
@acadgild

#Spark Use Case - Uber #DataAnalysis buff.ly/1NvpXDN

Spark…
In this …
acadg…

8s

ACADGILD
@acadgild

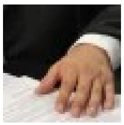
#TechTipTuesday:

Embed View on Twitter

RECENT POSTS

Spark Use Case – Uber Data Analysis
□ May 16, 2016

Why Learning MongoDB Will Boost Your Career
□ May 14, 2016



Job Responsibilities of Hadoop Professionals
□ May 13, 2016

Graphical Exploratory Data Analysis-II
□ May 13, 2016

Front End, Full Stack, AngularJS, NodeJS and Android Programming for children.

□ PREVIOUS ARTICLE

Know the API differences between MRV1 and MRV2

NEXT ARTICLE □

Android vs iOS

RELATED POSTS

Spark Use Case – Uber Data Analysis

May 16, 2016

Integrating SparkSQL with MySQL

May 12, 2016

Spark Use Case – Travel Data Analysis

May 8, 2016

LEAVE A REPLY

COMMENTS *

NAME *

EMAIL *

ARCHIVES

- May 2016
- April 2016
- March 2016
- February 2016
- January 2016
- December 2015
- November 2015
- September 2015
- August 2015
- July 2015
- June 2015
- May 2015
- November 2014
- October 2014
- September 2014
- August 2014

WEBSITE

SUBMIT

□ **NOTIFY ME OF FOLLOW-UP COMMENTS BY EMAIL.**

□ **NOTIFY ME OF NEW POSTS BY EMAIL.**

CATEGORIES

- AcadGild
- Android
- Android For Kids
- AngularJS
- Big Data and Hadoop
- Careers
- Cloud computing
- Database
- Digital Marketing
- Front End
- Full Stack
- Hadoop Administration
- IOS
- Java
- Kids
- Linux Administration

TAGS

ANDORID PROFILING TOOLS

ANDROID APP FOR SPEECH TO TEXT

ANDROID APP FOR TEXT TO SPEECH

ANDROID DEVELOPMENT

ANDROID MEMORY ANALYZER

ANDROID MEMORY MANAGEMENT

BANGALORE SUMMER CAMP

BEST SUMMER CAMPS 2016

BIG DATA DEVELOPEMENT

COMMISSIONING AND DECOMMISSIONING OF DATANODE IN HADOOP

DEPENDENCY INJECTION

LIKE WHAT YOU SEE? SUBSCRIBE TO OUR BLOG

We send only 1 email in a week

Enter your email...

Subscribe

- NodeJS
- Others
- Python
- Quiz
- R & Machine Learning
- Scala
- Spark
- Uncategorized

DIFFERENCE BETWEEN ANDROID VS IOS

FEATURES OF DDMS

FILE FORMATS

FILE FORMATS IN HADOOP

HADOOP

HADOOP ADMINISTRATION

HDFS

HIVE WITH MYSQL

INTRODUCTION TO SPARK

JAVASCRIPT MVC FRAMEWORK

JOB OPPORTUNITIES IN HADOOP

JOB TRENDS IN BIG DATA BLOG

JOB TRENDS IN HADOOP

LINUX

LINUX BASIC

LINUX BASIC COMMANDS

LINUX COMMANDS

MYSQL

MYSQL-CONNECTOR-JAVA-5.1.2.JAR

MYSQL-CONNECTOR-JAVA.JAR

MYSQL WITH HIVE

MYSQL WITH SQOOP

PYTHON

RACK AWARENESS

RECYCLE BIN

RESILIENT DISTRIBUTED DATASET (RDD)

SPARK

SQOOP

SQOOP WITH MYSQL

STYLING A RESPONSIVE WEB PAGE

SUMMER CAMP

SUMMER CAMP 2016

TOP 10 RECORDS IN MAPREDUCE

TRASH CONFIGURATION