16 MAY 2016

Spark Use Case – Uber Data Analysis

In this post, we will be performing analysis on the Uber dataset in Apache spark using Scala.

The Uber dataset consists of 4 columns. They are dispatching_base_number, date, active_vehicles and trips. You can download the dataset from the below link:

https://drive.google.com/open?id=0ByJLBTmJojjzS2c2UktqLW5uRG8

Problem Statement:

Find the days on which each basement has more trips.

Source Code:

YES, I WANT TO BOOST MY CAREER & INCREASE MY SALARY!

Your Name (required)

Your Email (required)

Your Contact Number (required)

Your Message

TELL ME HOW

LIKE WHAT YOU SEE? SUBSCRIBE TO OUR BLOG

We send only 1 email in a week

Enter your email...

Subscribe

```
1 val dataset = sc.textFile("/home/kiran/Desktop/uber")
2 val header = dataset.first()
3 val format = new java.text.SimpleDateFormat("MM/dd/yyyy")
4 var days =Array("Sun","Mon","Tue","Wed","Thu","Fri","Sat")
5 val eliminate = dataset.filter(line => line != header)
6 val split = eliminate.map(line => line.split(",")).map { x => (x
7 (0),format.parse(x(1)),x(3)) }
8 val combine = split.map(x => (x._1+" "+days(x._2.getDay),x.
  _3.toInt))
  val arrange = combine.reduceByKey(_+_).map(item => item.
  swap).sortByKey(false).collect.foreach(println)
```

Here's the explanation of the above code:

- In **line 1,** we are loading the dataset in our local system, using the textFile method.
- In **line 2,** we are creating a variable **header,** which holds the first line of the dataset (In this dataset the first line is header line).
- In **line 3,** we are declaring the date format using SimpleDateFormat in Java. Here the date is in the format of MM/dd/YYYY.
- In **line 4,** we are declaring an array, which will hold the days of the week from Sunday to Saturday.
- In **line 5,** we are filtering the header line from the dataset using the filter RDD.

Think you know it all about Spark?
Take this simple quiz to find out!
Yes, I'm Game
Spark

- In **line 6,** we are splitting the dataset using the delimiter as ***coma and*** taking out the three columns; *dispatching_base_number,* which is in the 1st column, date which is in the second column and the *number of trips,* which is in the fourth column. While extracting the columns, we are parsing the date, which is in string format to date format.
- After this step, we will get the records as *B02512, Thu Jan 01 00:00:00 IST 2015, 1132.*

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

- In **line 7,** we are adding the two columns, *dispatching_base_number* and *formatted date.* To get the *day* from the formatted date, we need to use the **getDay** method of java.util.Date package. Here, we will get the day number of the week and pass the day number into the array consisting of the names of the days. Finally, we will get the combination of *dispatching_base_number* and *day* of the week and the number of weeks. These are like keys and values.

- In **line 8,** we are using the reduceByKey RDD to combine all the values for each unique key, where key is the combination of *dispatching_base_number* and *day* of the week. After this, we are swapping the keys and values and then perform sortByKey action on the RDD, which will sort the records by values in the descending order. Finally, we are printing the result using the **collect** action.

Output:

(356789,B02764 Sat)

(326968,B02764 Fri)

(304200,B02764 Thu)

(249896,B02764 Sun)

(241137,B02764 Wed)

- R & Machine Learning
- Scala
- Spark
- Uncategorized

GET SOCIAL

AcadGild
Local Business · Bangal
92,286 likes

Like Page

3 friends like this

AcadGild shared a li
1 hr

This summer, give your child a skill for life. Check out our technology summer camps!

Tell Me More

ACADGILD
Learn. Do. Earn.

CATEGORIES

(125067,B02617 Fri)

(120283,B02682 Sat)

(118254,B02617 Thu)

(114662,B02682 Fri)

WHAT'S TRENDING

(106643,B02682 Thu)

(94887,B02617 Wed)

(94588,B02598 Sat)

(93126,B02598 Fri)

(91722,B02617 Sun)

(90333,B02598 Thu)

(86602,B02617 Tue)

(86252,B02682 Wed)

(82825,B02682 Sun)

(80591,B02617 Mon)

(76905,B02682 Tue)

(74939,B02682 Mon)

(71956,B02598 Wed)

(66477,B02598 Sun)

(63429,B02598 Tue)

(60882,B02598 Mon)

(36737,B02765 Sat)

(34934,B02765 Fri)

(30408,B02765 Thu)

(24340,B02765 Wed)

(22741,B02765 Tue)

(22536,B02765 Sun)

(21974,B02765 Mon)

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, 2

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

(16435,B02512 Fri)

(15809,B02512 Thu)

(15026,B02512 Sat)

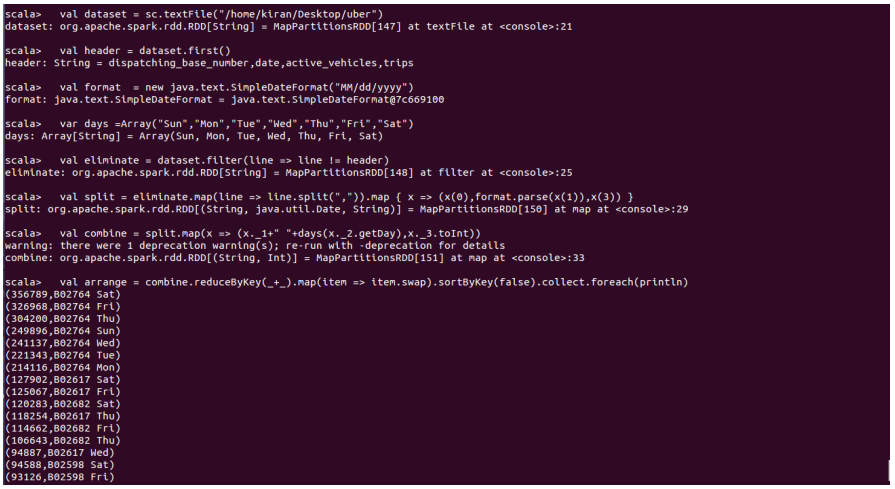
(12691,B02512 Wed)

(12041,B02512 Tue)

(11297,B02512 Mon)

(10487,B02512 Sun)

The same is as displayed in the below screenshot.



```
scala>   val dataset = sc.textFile("/home/kiran/Desktop/uber")
dataset: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[147] at textFile at <console>:21

scala>   val header = dataset.first()
header: String = dispatching_base_number,date,active_vehicles,trips

scala>   val format  = new java.text.SimpleDateFormat("MM/dd/yyyy")
format: java.text.SimpleDateFormat = java.text.SimpleDateFormat@7c669100

scala>   var days =Array("Sun","Mon","Tue","Wed","Thu","Fri","Sat")
days: Array[String] = Array(Sun, Mon, Tue, Wed, Thu, Fri, Sat)

scala>   val eliminate = dataset.filter(line => line != header)
eliminate: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[148] at filter at <console>:25

scala>   val split = eliminate.map(line => line.split(",")).map { x => (x(0),format.parse(x(1)),x(3)) }
split: org.apache.spark.rdd.RDD[(String, java.util.Date, String)] = MapPartitionsRDD[150] at map at <console>:29

scala>   val combine = split.map(x => (x._1+" "+days(x._2.getDay),x._3.toInt))
warning: there were 1 deprecation warning(s); re-run with -deprecation for details
combine: org.apache.spark.rdd.RDD[(String, Int)] = MapPartitionsRDD[151] at map at <console>:33

scala>   val arrange = combine.reduceByKey(_+_).map(item => item.swap).sortByKey(false).collect.foreach(println)
(356789,B02764 Sat)
(326968,B02764 Fri)
(304200,B02764 Thu)
(249896,B02764 Sun)
(241137,B02764 Wed)
(221343,B02764 Tue)
(214116,B02764 Mon)
(127902,B02617 Sat)
(125067,B02617 Fri)
(120283,B02682 Sat)
(118254,B02617 Thu)
(114662,B02682 Fri)
(106643,B02682 Thu)
(94887,B02617 Wed)
(94588,B02598 Sat)
(93126,B02598 Fri)
```

Hope this post has been helpful in understanding how to perform analysis in Spark using the Uber dataset. In case of any queries, feel free to comment below or write to us at support@acadguild.com.

Keep visiting our site www.acadgild.com for more updates on Big Data,Spark and other technologies.

Share this:

26

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

Related

Spark Use Case - Olympics Data Analysis
April 7, 2016
In "Spark"

Spark Use Case – The Daily Show
April 16, 2016
In "Spark"

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

PREVIOUS ARTICLE

Why Learning MongoDB Will Boost Your Career

RELATED POSTS

Integrating SparkSQL with

Spark Use Case – Travel Data

Analyzing New York Crime Data

MySQL	Analysis	Using SparkSQL
May 12, 2016	May 8, 2016	April 28, 2016

LEAVE A REPLY

COMMENTS *

NAME * EMAIL *

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

TAGS

ANDORID PROFILING TOOLS

ANDROID APP FOR SPEECH TO TEXT

ANDROID APP FOR TEXT TO SPEECH

ANDROID DEVELOPMENT

LIKE WHAT YOU SEE? SUBSCRIBE TO OUR BLOG

We send only 1 email in a week

Enter your email...

Subscribe

- Cloud computing
- Database
- Digital Marketing
- Front End
- Full Stack
- Hadoop Administration
- IOS
- Java
- Kids
- Linux Administration
- NodeJS
- Others
- Python
- Quiz
- R & Machine Learning
- Scala
- Spark
- Uncategorized

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

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

© Copyright 2016. ACADGILD.