Access Log - independent IPs

Course: 大数据软件工程

Exercise type: PySpark

Soft Due: 2015-05-27 18:00

Hard Due: 2015-05-27 18:00

Note: hard due has passed. You can no longer submit answer.

Author: 王欣明(TEACHER)

Email: wangxm35@mail.sysu.edu.cn

你已经接近或者达到满分。在完成这道题后，如果愿意的话，请你在下面**Feedback**那里评价一下这道题。你的反馈将同时用邮件发送给作者。

Description:

0 Questions and 0 Answers | Ask New Question

Write a Spark program to analyze access_log and find the URL with top-20 independent IP visits.

The data format is as follows:

10.223.157.186 - - [15/Jul/2009:14:58:59 -0700] "GET / HTTP/1.1" 403 202
10.223.157.186 - - [15/Jul/2009:14:58:59 -0700] "GET /favicon.ico HTTP/1.1" 404 209
10.223.157.186 - - [15/Jul/2009:15:50:35 -0700] "GET / HTTP/1.1" 200 9157
10.223.157.186 - - [15/Jul/2009:15:50:35 -0700] "GET /assets/js/lowpro.js HTTP/1.1" 200 10469
10.223.157.186 - - [15/Jul/2009:15:50:35 -0700] "GET /assets/css/reset.css HTTP/1.1" 200 1014
10.223.157.186 - - [15/Jul/2009:15:50:35 -0700] "GET /assets/css/960.css HTTP/1.1" 200 6206
10.223.157.186 - - [15/Jul/2009:15:50:35 -0700] "GET /assets/css/the-associates.css HTTP/1.1" 200 15779
10.223.157.186 - - [15/Jul/2009:15:50:35 -0700] "GET /assets/js/the-associates.js HTTP/1.1" 200 4492
10.223.157.186 - - [15/Jul/2009:15:50:35 -0700] "GET /assets/js/lightbox.js HTTP/1.1" 200 25960

URL is defined as the whole string quoted by " in each line.

The output format shall be:

URL IP-visit-count

URL IP-visit-count

....

Note:

- Items in each line are separated by "\t"
- The list is ordered by ip visit count, from high to low.

Hint:

Hint is not available for this exercise.

Your answer:

Already pass (hard) due date: May 27, 2015, 6 p.m. You can no longer submit your code to smart_programmer by yourself.

We don't accept late submission in principle. But if you really have a convincing reason, send the answer code to your TA by email and ask her/him to submit it for you. Don't copy the standard answer. It will be considered as plagirism.

If you want to study the standard answer, you can add it as an optional assignment.

Add this exercise as an optional assignment

Hard due has passed. The standard answer is unlocked:

driver.py | [*]main.py

```
1. import
2. from pyspark import SparkContext
3. from      import
4.
5. if          == "__main__":
6.     sc = SparkContext(appName="Co-occurrence")
7.         =    .        ("hdfs://eden/user/driver/smart_programmer/access_log/access_log")
8.     main(rdd)
9.
```

Discussion:

Discussion is not available for this exercise.

Latest Submission Grade: 100 (submitted on 2015-05-19 22:35)

```
==================================[Phase1: check_plagirism_cloudera]==================================
Pass.

==================================[Phase2: cloudera_prepare]==========================================
Pass syntax checking. You got 10 points.

==================================[Phase3: cloudera_pyspark_cluster]==================================
Spent time: 14.9319150448 seconds. The standard execution time is 45 seconds.
Your spark program runs faster than the standard answer. How do you manage to achieve this? You get 100% c

Correctness points: 90
```

Your submission record:

Submission Date	Grade	Diff
2015-05-19 22:35	100	Diff
2015-05-19 22:23	10	Diff
2015-05-19 22:03	10	Diff
2015-05-19 21:32	10	Diff
2015-05-19 21:26	10	Diff
2015-05-19 21:14	0	Diff
2015-05-19 20:34	10	Diff
2015-05-19 20:01	10	Diff

Top performers in the class:

Name	Points	Total Trials	Last Submission
持盾卫士	100	10	2015-05-19 22:09
完结撒花	100	14	2015-05-19 22:16
Pan	100	5	2015-05-19 22:30
DIS	100	8	2015-05-19 22:35
邓宇恒	100	8	2015-05-19 22:37
test	100	4	2015-05-19 22:50
刘翔宇	100	2	2015-05-19 22:50
聂照昌	100	3	2015-05-20 01:36
T	100	18	2015-05-20 08:11

陈泽宇	100	5	2015-05-20 08:35

Feedback: (这里不是填答案的地方！)

Warning: Don't put your answer into the feedback, as they will be seen by all other students. If you have doubts on your answer, send an email to the author of this assignment instead.

为减轻服务器负担，在评论中插入大图像时建议用链接，小图像的时候建议用base64编码直接嵌入，即打开html编辑页面插入像<img src="">一样的标签 (图片转base64编码的工具)。注意，评论的大小不能超过100k字节，如果超出就会显示Invalid form data。

难度(选填): ----------

题意表达是否清楚(选填): ----------

对学习的帮助(选填): ----------

你对本题的评价(选填):

Font Family | Font Size | Paragraph

Path: p

Upload

Feedbacks from students:

Class	Comments
BDSE_2015	GET / HTTP/1.1 3521 为什么跟答案不一样 GET / HTTP/1.1 3519
BDSE_2015	答案和标准答案一样 提示 get 100% of the correctness points 为何还是 10 分 ==========[Phase2: cloudera_prepare]========== Pass syntax checking. You got 10 points. ==========[Phase3: cloudera_pyspark_cluster]========== Spent time: 16.0572550297 seconds. The standard execution time is 45 seconds. Your spark program runs faster than the standard answer. How do you manage to achieve this? You get 100% of the correctness points. Correctness points: 0

```
Your result:
GET /assets/css/combined.css HTTP/1.1   7247
GET /assets/js/javascript_combined.js HTTP/1.1  6872
GET /assets/img/home-logo.png HTTP/1.1  6130
GET /assets/img/banner/ten-years-banner-white.jpg HTTP/1.1
        6083
GET /assets/img/banner/ten-years-banner-grey.jpg HTTP/1.1
        6041
GET /assets/img/banner/ten-years-banner.png HTTP/1.1    59
76
GET /favicon.ico HTTP/1.1       5665
GET /images/filmmediablock/290/Harpoon_2d.JPG HTTP/1.1  52
22
GET /assets/img/banner/ten-years-banner-black.jpg HTTP/1.1
        3533
GET / HTTP/1.1  3521
GET /assets/css/ie.css HTTP/1.1 3217
GET /assets/img/search-button.gif HTTP/1.1      2881
GET /assets/img/play_icon.png HTTP/1.1  2527
GET /assets/img/x.gif HTTP/1.1  2263
GET /assets/img/release-schedule-logo.png HTTP/1.1      16
46
GET /images/frontpagepics/0000/0028/Trucker2.jpg HTTP/1.1
        1597
GET /release-schedule/ HTTP/1.1 1476
GET /release-schedule HTTP/1.1  1346
GET /images/frontpagepics/0000/0024/The_Greatestcrop.jpg H
TTP/1.1 1274
GET /robots.txt HTTP/1.1        1239

is different from the standard answer:
GET /assets/css/combined.css HTTP/1.1   7247
GET /assets/js/javascript_combined.js HTTP/1.1  6872
GET /assets/img/home-logo.png HTTP/1.1  6130
GET /assets/img/banner/ten-years-banner-white.jpg HTTP/1.1
        6083
GET /assets/img/banner/ten-years-banner-grey.jpg HTTP/1.1
        6041
GET /assets/img/banner/ten-years-banner.png HTTP/1.1    59
76
GET /favicon.ico HTTP/1.1       5665
GET /images/filmmediablock/290/Harpoon_2d.JPG HTTP/1.1  52
22
GET /assets/img/banner/ten-years-banner-black.jpg HTTP/1.1
        3533
GET / HTTP/1.1  3519
GET /assets/css/ie.css HTTP/1.1 3217
GET /assets/img/search-button.gif HTTP/1.1      2881
GET /assets/img/play_icon.png HTTP/1.1  2527
GET /assets/img/x.gif HTTP/1.1  2263
GET /assets/img/release-schedule-logo.png HTTP/1.1      16
46
GET /images/frontpagepics/0000/0028/Trucker2.jpg HTTP/1.1
        1597
GET /release-schedule/ HTTP/1.1 1476
GET /release-schedule HTTP/1.1  1346
GET /images/frontpagepics/0000/0024/The_Greatestcrop.jpg H
TTP/1.1 1274
GET /robots.txt HTTP/1.1        1239
.
```

BDSE_2015

报错是这样的，不知道是怎么回事呢

```
Output content:
Traceback (most recent call last):
  File "/tmp/smart_programmer/assignment-4160/submission-1
164392/./driver.py", line 6, in <module>
    sc = SparkContext(appName="Co-occurrence")
```

```
  File "/opt/cloudera/parcels/CDH-5.4.0-1.cdh5.4.0.p0.27/l
ib/spark/python/pyspark/context.py", line 111, in __init__
    conf, jsc, profiler_cls)
  File "/opt/cloudera/parcels/CDH-5.4.0-1.cdh5.4.0.p0.27/l
ib/spark/python/pyspark/context.py", line 159, in _do_init
    self._jsc = jsc or self._initialize_context(self._conf
._jconf)
  File "/opt/cloudera/parcels/CDH-5.4.0-1.cdh5.4.0.p0.27/l
ib/spark/python/pyspark/context.py", line 212, in _initial
ize_context
    return self._jvm.JavaSparkContext(jconf)
  File "/opt/cloudera/parcels/CDH-5.4.0-1.cdh5.4.0.p0.27/l
ib/spark/python/lib/py4j-0.8.2.1-src.zip/py4j/java_gateway
.py", line 701, in __call__
  File "/opt/cloudera/parcels/CDH-5.4.0-1.cdh5.4.0.p0.27/l
ib/spark/python/lib/py4j-0.8.2.1-src.zip/py4j/protocol.py"
, line 300, in get_return_value
py4j.protocol.Py4JJavaError: An error occurred while calli
ng None.org.apache.spark.api.java.JavaSparkContext.
: org.apache.hadoop.ipc.RemoteException(org.apache.hadoop.
hdfs.server.namenode.SafeModeException): Cannot create fil
e/user/spark/applicationHistory/local-1432181055910.inprog
ress. Name node is in safe mode.
Resources are low on NN. Please add or free up more resour
ces then turn off safe mode manually. NOTE:  If you turn o
ff safe mode before adding resources, the NN will immediat
ely return to safe mode. Use "hdfs dfsadmin -safemode leav
e" to turn safe mode off.
        at org.apache.hadoop.hdfs.server.namenode.FSNamesy
stem.checkNameNodeSafeMode(FSNamesystem.java:1413)
        at org.apache.hadoop.hdfs.server.namenode.FSNamesy
stem.startFileInt(FSNamesystem.java:2671)
        at org.apache.hadoop.hdfs.server.namenode.FSNamesy
stem.startFile(FSNamesystem.java:2557)
        at org.apache.hadoop.hdfs.server.namenode.NameNode
RpcServer.create(NameNodeRpcServer.java:570)
        at org.apache.hadoop.hdfs.server.namenode.Authoriz
ationProviderProxyClientProtocol.create(AuthorizationProvi
derProxyClientProtocol.java:110)
        at org.apache.hadoop.hdfs.protocolPB.ClientNamenod
eProtocolServerSideTranslatorPB.create(ClientNamenodeProto
colServerSideTranslatorPB.java:395)
        at org.apache.hadoop.hdfs.protocol.proto.ClientNam
enodeProtocolProtos$ClientNamenodeProtocol$2.callBlockingM
ethod(ClientNamenodeProtocolProtos.java)
        at org.apache.hadoop.ipc.ProtobufRpcEngine$Server$
ProtoBufRpcInvoker.call(ProtobufRpcEngine.java:619)
        at org.apache.hadoop.ipc.RPC$Server.call(RPC.java:
1060)
        at org.apache.hadoop.ipc.Server$Handler$1.run(Serv
er.java:2044)
        at org.apache.hadoop.ipc.Server$Handler$1.run(Serv
er.java:2040)
        at java.security.AccessController.doPrivileged(Nat
ive Method)
        at javax.security.auth.Subject.doAs(Subject.java:4
22)
        at org.apache.hadoop.security.UserGroupInformation
.doAs(UserGroupInformation.java:1671)
        at org.apache.hadoop.ipc.Server$Handler.run(Server
.java:2038)
```

BDSE_2015	[WARNING] Feedback contains the answer. Don't do that again as this will be seen by other student!
BDSE_2015	我不管提交什么代码都不能运行的样子QAQ以为是我的问题 交了别人过了的题目还是报错QAQ : org.apache.hadoop.ipc.RemoteException(org.apache.hadoop. hdfs.server.namenode.SafeModeException): Cannot create fil e/user/spark/applicationHistory/local-1432141382068.inprog ress. Name node is in safe mode.

	Resources are low on NN. Please add or free up more resour ces then turn off safe mode manually. NOTE: If you turn o ff safe mode before adding resources, the NN will immediat ely return to safe mode. Use "hdfs dfsadmin -safemode leav e" to turn safe mode off.
BDSE_2015	好像不是只有我一个这样是么 An error occurred while calling None.org.apache.spark.api. java.JavaSparkContext. : org.apache.hadoop.ipc.RemoteException(org.apache.hadoop. hdfs.server.namenode.SafeModeException): Cannot create fil e/user/spark/applicationHistory/local-1432171100768.inprog ress. Name node is in safe mode. Resources are low on NN. Please add or free up more resour ces then turn off safe mode manually.
BDSE_2015	

BDSE_2015	空间满了？。。。
	
	==========
	Failure: error code:1
	
	Output content:
	Traceback (most recent call last):
	  File "/tmp/smart_programmer/assignment-4160/submission-1
	165018/./driver.py", line 6, in <module>
	    sc = SparkContext(appName="Co-occurrence")
	  File "/opt/cloudera/parcels/CDH-5.4.0-1.cdh5.4.0.p0.27/l
	ib/spark/python/pyspark/context.py", line 111, in __init__
	    conf, jsc, profiler_cls)
	  File "/opt/cloudera/parcels/CDH-5.4.0-1.cdh5.4.0.p0.27/l
	ib/spark/python/pyspark/context.py", line 159, in _do_init
	    self._jsc = jsc or self._initialize_context(self._conf
	._jconf)
	  File "/opt/cloudera/parcels/CDH-5.4.0-1.cdh5.4.0.p0.27/l
	ib/spark/python/pyspark/context.py", line 212, in _initial
	ize_context
	    return self._jvm.JavaSparkContext(jconf)
	  File "/opt/cloudera/parcels/CDH-5.4.0-1.cdh5.4.0.p0.27/l
	ib/spark/python/lib/py4j-0.8.2.1-src.zip/py4j/java_gateway
	.py", line 701, in __call__
	  File "/opt/cloudera/parcels/CDH-5.4.0-1.cdh5.4.0.p0.27/l
	ib/spark/python/lib/py4j-0.8.2.1-src.zip/py4j/protocol.py"
	, line 300, in get_return_value
	py4j.protocol.Py4JJavaError: An error occurred while calli
	ng None.org.apache.spark.api.java.JavaSparkContext.
	: org.apache.hadoop.ipc.RemoteException(org.apache.hadoop.
	hdfs.server.namenode.SafeModeException): Cannot create fil
	e/user/spark/applicationHistory/local-1432202272570.inprog
	ress. Name node is in safe mode.
	Resources are low on NN. Please add or free up more resour
	ces then turn off safe mode manually. NOTE: If you turn o
	ff safe mode before adding resources, the NN will immediat
	ely return to safe mode. Use "hdfs dfsadmin -safemode leav
	e" to turn safe mode off.

Please contact **Dr. Wang (roadlit at gmail.com)** to report bugs.

Thanks to:

- 吴浩坚同学 and 梁展瑞同学
- Django, Gunicorn, TinyMCE, Sandbox, Nginx
- Valgrind, Google Gode Style SOClone

Recent messages:

- [06/01 14:50] From 王欣明: 最近一段时间经常发生有些同学的作业卡队列的现象。如果发现请及时通知我重启服务队列。
- [06/01 14:49] From 王欣明: 本周是大数据软件工程的期末课程项目，题目会在周二晚上6点开放，同学们可以自己在宿舍做，一周时间内完成。
- [05/31 00:38] From 谢议尊: test

Send message to an user with real name or nick name:

Name:

Message:

Send Message