Wiki - Top10 Frequent Visits

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

In this assignment, we will use Wikipedia traffic statistics data obtained from http://aws.amazon.com/datasets/4182 . To make the analysis feasible (within the short timeframe of the exercise), we only took a small sample.

The data files contain traffic statistics for all pages in a specific hour. In total there are nine input files, each with around 200MB data:

- smart_programmer/wiki/part-00097
- smart_programmer/wiki/part-00098
- smart_programmer/wiki/part-00099
- smart_programmer/wiki/part-00100
- smart_programmer/wiki/part-00101
- smart_programmer/wiki/part-00102
- smart_programmer/wiki/part-00103
- smart_programmer/wiki/part-00104
- smart_programmer/wiki/part-00105

The content of each file is like the follows:

20090507-040000 aa ?page=http://www.stockphotosharing.com/Themes/Images/users_raw/id.txt 3 39267

20090507-040000 aa Main_Page 7 51309

20090507-040000 aa Special:Boardvote 1 11631

20090507-040000 aa Special:Imagelist 1 931

Each line, delimited by a space, contains stats for one page. The schema is:

The <date_time> field specifies a date in the YYYYMMDD format (year, month, day) followed by a hyphen and then the hour in the HHmmSS format (hour, minute, second). There is no information in mmSS. The <project_code> field contains information about the language of the pages. For example, project code “en” indicates an English page. The <page_title> field gives the title of the Wikipedia page.
The <num_hits> field gives the number of page views in the hour-long time slot starting at <data_time>. The <page_size> field gives the size in bytes of the Wikipedia page.

Your task in this assignment is to write a spark program to find the top-10 frequently visited pages. The visit count is summed across the whole dataset in the nine input files.

Note: Pages with different project_code are considered different pages.

The output shall be like:

aa XXX 100000

aa YYY 53123

en ZZZ 23112

…

Note:

The items in each line shall be separated by "/t"

The spark program shall consist of two python code file. One is "driver.py", which is given to you by the smart_programmer system. You can read its code in the assignment page. "driver.py" will create the spark context and load the input data files as intial RDDs, then pass these input RDDs to the function "main" that shall be written by you in "main.py"

To output the result, simply use "print" to print the result in stdout in the main function.

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.                  = [
6.      "smart_programmer/wiki/part-00097",
7.      "smart_programmer/wiki/part-00098",
8.      "smart_programmer/wiki/part-00099",
9.      "smart_programmer/wiki/part-00100",
10.     "smart_programmer/wiki/part-00101",
11.     "smart_programmer/wiki/part-00102",
12.     "smart_programmer/wiki/part-00103",
13.     "smart_programmer/wiki/part-00104",
14.     "smart_programmer/wiki/part-00105"
15. ]
16.
17. if          == "__main__":
18.     sc = SparkContext(appName="wiki-top10")
19.     RDDs = []
20.     for input_file_name in input_file_names:
21.             =    .        ("hdfs://eden/user/driver/"+              )
22.         RDDs.append(rdd)
23.         (RDDs)
24.
```

Discussion:

Discussion is not available for this exercise.

Latest Submission Grade: 100 (submitted on 2015-05-19 19:14)

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

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

==========[Phase3: cloudera_pyspark_cluster]==========
Spent time: 103.884063959 seconds. The standard execution time is 151 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 19:14	100	Diff
2015-05-19 19:06	0	Diff
2015-05-19 18:48	10	Diff
2015-05-19 18:35	10	Diff
2015-05-19 18:27	10	Diff
2015-05-19 15:50	10	Diff
2015-05-19 15:10	10	Diff
2015-05-19 14:53	10	Diff
2015-05-19 14:49	0	Diff
2015-05-19 14:48	0	Diff

Top performers in the class:

Name	Points	Total Trials	Last Submission
持盾卫士	100	13	2015-05-19 19:13
DIS	100	10	2015-05-19 19:14
XiaoMoMo	100	7	2015-05-19 19:20
把我飘准的普通发凡我	100	6	2015-05-19 20:07
莫宇诚	100	5	2015-05-19 20:17
叶菀仪	100	10	2015-05-19 20:18
郑穗展	100	5	2015-05-19 20:42
陈泽宇	100	8	2015-05-19 21:46
刘翔宇	100	4	2015-05-19 21:57
陈章根	100	8	2015-05-19 23:21

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	
BDSE_2015	Spent time: 157.344066858 seconds. The standard execution time is 151 seconds. 求教如何优化?
BDSE_2015	为什么今天开始三道题都出现同样的问题，是爆内存了吗 Output content: Traceback (most recent call last): File "/tmp/smart_programmer/assignment-4158/submission-1164424/./driver.py", line 18, in <module> sc = SparkContext(appName="wiki-top10") File "/opt/cloudera/parcels/CDH-5.4.0-1.cdh5.4.0.p0.27/lib/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/lib/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/lib/spark/python/pyspark/context.py", line 212, in _initialize_context return self._jvm.JavaSparkContext(jconf) File "/opt/cloudera/parcels/CDH-5.4.0-1.cdh5.4.0.p0.27/lib/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/lib/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 calling None.org.apache.spark.api.java.JavaSparkContext. : org.apache.hadoop.ipc.RemoteException(org.apache.hadoop.hdfs.server.namenode.SafeModeException): Cannot create file/user/spark/applicationHistory/local-1432182409830.inprogress. Name node is in safe mode. Resources are low on NN. Please add or free up more resources then turn off safe mode manually. NOTE: If you turn off safe mode before adding resources, the NN will immediately return to safe mode. Use "hdfs dfsadmin -safemode leave" to turn safe mode off.
BDSE_2015	Spent time: 219.591500998 seconds. The standard execution time is 151 seconds. help.....

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