ID2221 Data Intensive Computing

Lab1 Apache Spark

Group1: Yizhan Wu (yizhanw@kth.se), Yage Hao (yage@kth.se)

1 Introduction

In this assignment, we use Apache Spark to explore the page views of the Wikimedia project. Dataset: the page view statistics generated between 0-1am on Jan 1, 2016. The schema looks as follows:

- Project code
- Page title
- Page hits
- Page size

2 How to run the code

Requirements:

- Java SDK 8
- Scala
- Apache Spark
- Jupyter Notebook

You have two ways running the codes:

- Launch LAB1.ipynb in Jupyter Notebook under spark-kernel.
- Or use the following command to run codes in LAB1.scala:

```
>> spark-shell -i LAB1.scala
```

3 Results

Task 1 - Spark

1. Retrieve the first 15 records and print out the result.

```
Log(aa,271_a.C,1,4675)
Log(aa,Category:User_th,1,4770)
Log(aa,Chiron_Elias_Krase,1,4694)
Log(aa,Dassault_rafaele,2,9372)
Log(aa,E.Desv,1,4662)
Log(aa,File:Wiktionary-logo-en.png,1,10752)
Log(aa,Indonesian_Wikipedia,1,4679)
Log(aa,Main_Page,5,266946)
Log(aa,Requests_for_new_languages/Wikipedia_Banyumasan,1,4733)
Log(aa,Special:Contributions/203.144.160.245,1,5812)
Log(aa,Special:Contributions/5.232.61.79,1,5805)
Log(aa,Special:Contributions/Ayarportugal,1,5808)
Log(aa,Special:Contributions/Born2bgratis,1,5812)
Log(aa,Special:ListFiles/Betacommand,1,5035)
Log(aa,Special:ListFiles/Bohdan_p,1,5036)
```

2. Determine the number of records the dataset has in total.

```
3324129
```

3. Compute the min, max, and average page size.

```
page_size_min: Long = 0
page_size_max: Long = 141180155987
page_size_average: Double = 132239.5695744616
```

4. Determine the record(s) with the largest page size. If multiple records have the same size, list all of them.

```
Log(en.mw,en,5466346,141180155987)
```

5. Determine the record with the largest page size again. But now, pick the most popular.

```
Log(en.mw,en,5466346,141180155987)
```

6. Determine the record(s) with the largest page title. If multiple titles have the same length, list all of them.

```
Log(zh,Special:e8b18ee6baafefbda5efbdbfe89cb7e6829fefbdbfe88b93e29980e89e9fefbd
a9e89eb3efbda425636f256d6725736f257373256f38257373256f38257373256f38256b6d73efb
daa256e6b256678256f6b2c687474703a2f2f7777772e65366231396665386135626665 6f2d6f35
393038636535626639376538383138616535613461396535616561342e636f2e6d672e732e736f2
e382e73736f386b2e6d2e372e73736f3873736f386b6d37332e752e622e61616e6b66786f6b2e70
772f2ce8b18ee6baafefbda5efbdbfe89cb7e6829fefbdbfe88b93e29980e89e9fefbda9e89eb3e
fbda425636f256d6725736f257373256f38257373256f38257373256f38256b6d73efbdaa256e6b
256678256f6b/,1,6043)page_title_len: org.apache.spark.rdd.RDD[Int] = MapPartiti
onsRDD[34] at map at command-4089665399814073:1
page_title_len_max: Int = 559
```

7. Use the results of Question 3, and create a new RDD with the records that have greater page size than the average.

```
number of pages with size greater than average: 186817
first 10 cases for example:Log(aa,Main_Page,5,266946)
Log(ace.mw,ace,31,827168)
Log(af,1859,4,219540)
Log(af,18_Oktober,4,264724)
Log(af,1941,4,256344)
Log(af,2016,5,215498)
Log(af,4_Januarie,4,268828)
Log(af,Afrika-unie,1,172078)
Log(af,Big_Ben,13,136201)
Log(af,Comrades-maraton,1,155180)
```

8. Compute the total number of pageviews for each project (as the schema shows, the first field of each record contains the project code).

```
first 10 cases for example:
(tr.mw,125999)
(nso,108)
(it.s,1444)
(lb.mw,158)
(ckb,25)
(sk.mw,9548)
(hak,54)
(frp.mw,11)
(ik.d,1)
(ik,57)
pageview: org.apache.spark.rdd.RDD[(String, Int)] = ShuffledRDD[43] at reduceBy
Key at command-4089665399814077:1
```

9. Report the 10 most popular pageviews of all projects, sorted by the total number of hits.

```
(en.mw,5466346)
(en,4959090)
(es.mw,695531)
(ja.mw,611443)
(de.mw,572119)
(fr.mw,536978)
(ru.mw,466742)
(it.mw,400297)
(de,315929)
(commons.m,285796)
```

10. Determine the number of page titles that start with the article “The”. How many of those page titles are not part of the English project (Pages that are part of the English project have “en” as the first field)?

```
9128
```

11. Determine the percentage of pages that have only received a single page view in this one hour of log data.

```
onepageview: Float = 2558332.0
totalpage: Float = 3324129.0
percentage: Float = 0.76962477
```

12. Determine the number of unique terms appearing in the page titles. Note that in page titles, terms are delimited by “ ” instead of a white space. You can use any number of normalization steps (e.g., lowercasing, removal of non-alphanumeric characters).

```
1688528
```

13. Determine the most frequently occurring page title term in this dataset.

```
of
```

Task 2 - Spark SQL

3. Compute the min, max, and average page size.

```
+------------+---------+------------------+
|   max(size)|min(size)|         avg(size)|
+------------+---------+------------------+
|141180155987|        0|132239.56957446598|
+------------+---------+------------------+
```

5. Determine the record with the largest page size again. But now, pick the most popular.

```
+-----+-----+-------+------------+----+
| code|title|   hits|        size|rank|
+-----+-----+-------+------------+----+
|en.mw|   en|5466346|141180155987|   1|
+-----+-----+-------+------------+----+
```

7. Use the results of Question 3, and create a new RDD with the records that have greater page size than the average.

code	title	hits	size
aa	Main_Page	5	266946
ace.mw	ace	31	827168
af	1859	4	219540
af	18_Oktober	4	264724
af	1941	4	256344
af	2016	5	215498
af	4_Januarie	4	268828
af	Afrika-unie	1	172078
af	Big_Ben	13	136201
af	Comrades-maraton	1	155180

only showing top 10 rows

12. Determine the number of unique terms appearing in the page titles. Note that in page titles, terms are delimited by “ ” instead of a white space. You can use any number of normalization steps (e.g., lowercasing, removal of non-alphanumeric characters).

1688528

13. Determine the most frequently occurring page title term in this dataset.

[of,194407]