

Analyzing Wikipedia Articles



Quan Vu

Be advised: A lot of assumptions about the data were made to simplify these queries.

Question 1 Result

Which English wikipedia article got the most traffic on October 20?

- Most traffic/popular page is Main_Page with 5,961,008 count of views.
- It took about 5 mins to complete the job, even though, the log says 10.

```
2020-10-31 00:16:08,045 Stage-2 map = 0%, reduce = 0%
2020-10-31 00:16:13,162 Stage-2 map = 50%, reduce = 0%, Cumulative CPU 3.35 sec
2020-10-31 00:16:20,281 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 14.27 sec
2020-10-31 00:16:21,297 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 15.58 sec
MapReduce Total cumulative CPU time: 15 seconds 580 msec
Ended Job = job_1604125540383_0005
MapReduce Jobs Launched:
Stage-Stage-1: Map: 18 Reduce: 19 Cumulative CPU: 594.0 sec HDFS Read: 4616038853 HDFS Write: 270377994 SUCCESS
Stage-Stage-2: Map: 2 Reduce: 1 Cumulative CPU: 15.58 sec HDFS Read: 270393618 HDFS Write: 414 SUCCESS
Total MapReduce CPU Time Spent: 10 minutes 9 seconds 580 msec
OK
+-----+-----+
| newtable.page_title | total_views |
+-----+-----+
| Main_Page           | 5961008     |
| Special:Search       | 1476831     |
| -                    | 544714      |
| Jeffrey_Toobin       | 321459      |
| C._Rajagopalachari   | 210558      |
| The_Haunting_of_Bly_Manor | 185139      |
| Robert_Redford       | 178779      |
| Jeff_Bridges         | 159163      |
| Bible                | 151484      |
| Chicago_Seven        | 149966      |
+-----+-----+
10 rows selected (158.252 seconds)
0: jdbc:hive2://>
```

Question 1

Assumptions:

- October 20, 2020
- English wikipedia articles have domain_codes en and en.m

Steps:

1. Combined 24 files of October 20, 2020 into one big one, then populate the table. (Push to HDFS)
2. One query with 1 nested query:
 - a. Nested query that gets en and en.m elements and their sum of count_views (no duplicates)
 - b. Outer query to combine similar article names from both en and en.m by summing count_views one more time.

Question 2 Result

What English wikipedia article has the largest fraction of its readers follow an internal link to another wikipedia article?

- The English Wikipedia article that has largest fraction is Lists_of_deaths_by_year (112,328/147,800)

```
2020-11-02 21:16:27,446 Stage-4 map = 0%, reduce = 0%
2020-11-02 21:16:31,530 Stage-4 map = 100%, reduce = 0%, Cumulative CPU 1.18 sec
2020-11-02 21:16:35,586 Stage-4 map = 100%, reduce = 100%, Cumulative CPU 2.29 sec
MapReduce Total cumulative CPU time: 2 seconds 290 msec
Ended Job = job_1604346441296_0081
MapReduce Jobs Launched:
Stage-Stage-1: Map: 6 Reduce: 6 Cumulative CPU: 200.06 sec HDFS Read: 1422448402 HDFS Write: 78930852 SUCCESS
Stage-Stage-5: Map: 6 Reduce: 1 Cumulative CPU: 68.27 sec HDFS Read: 1422406506 HDFS Write: 59015 SUCCESS
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 8.62 sec HDFS Read: 78939346 HDFS Write: 39646 SUCCESS
Stage-Stage-7: Map: 1 Cumulative CPU: 1.8 sec HDFS Read: 66455 HDFS Write: 55510 SUCCESS
Stage-Stage-4: Map: 1 Reduce: 1 Cumulative CPU: 2.29 sec HDFS Read: 63332 HDFS Write: 665 SUCCESS
Total MapReduce CPU Time Spent: 4 minutes 41 seconds 40 msec
OK
```

| referrer | requested | fraction |
|-------------------------------|--------------------------------------|----------|
| Lists_of_deaths_by_year | Deaths_in_2020 | 0.76 |
| Elizabeth_Gillies | Michael_Corcoran_(musician) | 0.74 |
| Mr._Miyagi | Pat_Morita | 0.72 |
| Payback_(2020) | Clash_of_Champions_(2020) | 0.71 |
| I'm_Thinking_of_Ending_Things | I'm_Thinking_of_Ending_Things_(film) | 0.65 |
| Jane_C._Ginsburg | James_Steven_Ginsburg | 0.63 |
| Annie_Murphy | Menno_Versteeg | 0.62 |
| UEFA_Nations_League | 2020-21_UEFA_Nations_League | 0.6 |
| Carole_Baskin | Disappearance_of_Don_Lewis | 0.58 |
| Christina_El_Moussa | Flip_or_Flop | 0.58 |

```
10 rows selected (130.906 seconds)
0: jdbc:hive2://>
```

Question 2

Assumptions:

- Largest fraction means the highest number of clicks for an internal link inside an article divided by the total number of clicks of all internal links in that same article.
- Both original and its internal links are English wikipedia articles.
- Clickstream of September 2020.
- Capping number of popular referrers to 1000.

Steps:

1. Populate table with tabs separated elements (using clickstream data file in HDFS).
2. One query with 2 nested queries:
 - a. First nested to get the total clicks of the referrer. Second nested to get the clicks of the most popular referrer-requested pair.
 - b. Outer query to do the fraction.

Question 3 Result

What series of wikipedia articles, starting with Hotel California, keeps the largest fraction of its readers clicking on internal links?

- The series of English wikipedia articles: Hotel California->Eagles Album->The Long Run Album. (15%)

```
2020-11-02 22:10:33,247 Stage-2 map = 0%, reduce = 0%
2020-11-02 22:10:37,318 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 1.12 sec
2020-11-02 22:10:41,381 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 2.3 sec
MapReduce Total cumulative CPU time: 2 seconds 300 msec
Ended Job = job_1604346441296_0108
MapReduce Jobs Launched:
Stage-Stage-3: Map: 6 Reduce: 1 Cumulative CPU: 45.35 sec HDFS Read: 1422407188 HDFS Write: 3043 SUCCESS
Stage-Stage-6: Map: 6 Reduce: 6 Cumulative CPU: 55.81 sec HDFS Read: 1422452644 HDFS Write: 596 SUCCESS
Stage-Stage-7: Map: 1 Reduce: 1 Cumulative CPU: 2.02 sec HDFS Read: 8125 HDFS Write: 116 SUCCESS
Stage-Stage-11: Map: 1 Cumulative CPU: 1.32 sec HDFS Read: 10100 HDFS Write: 3821 SUCCESS
Stage-Stage-5: Map: 1 Reduce: 1 Cumulative CPU: 3.03 sec HDFS Read: 11291 HDFS Write: 148 SUCCESS
Stage-Stage-8: Map: 6 Cumulative CPU: 46.63 sec HDFS Read: 1422424390 HDFS Write: 4743 SUCCESS
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 2.3 sec HDFS Read: 14372 HDFS Write: 933 SUCCESS
Total MapReduce CPU Time Spent: 2 minutes 36 seconds 460 msec
OK
```

| referrer | requested1 | requested2 | final_fraction |
|------------------|---------------------------------|---------------------------------|----------------|
| Hotel_California | Hotel_California_(Eagles_album) | Hotel_California | 0.15 |
| Hotel_California | Hotel_California_(Eagles_album) | The_Long_Run_(album) | 0.15 |
| Hotel_California | Hotel_California_(Eagles_album) | Their_Greatest_Hits_(1971-1975) | 0.07 |
| Hotel_California | Hotel_California_(Eagles_album) | Eagles_(band) | 0.06 |
| Hotel_California | Hotel_California_(Eagles_album) | The_Beverly_Hills_Hotel | 0.04 |
| Hotel_California | Hotel_California_(Eagles_album) | Randy_Meisner | 0.03 |
| Hotel_California | Hotel_California_(Eagles_album) | Life_in_the_Fast_Lane | 0.03 |
| Hotel_California | Hotel_California_(Eagles_album) | Joe_Walsh | 0.03 |
| Hotel_California | Hotel_California_(Eagles_album) | New_Kid_in_Town | 0.03 |
| Hotel_California | Hotel_California_(Eagles_album) | Don_Felder | 0.03 |

```
10 rows selected (141.801 seconds)
0: jdbc:hive2://>
```

Question 3

Assumptions:

- Similar assumptions from question 2.
- Assume that the fraction number represents clicks of series starting from Hotel California.

Steps:

1. Using the same clickstream table and one query (with 3 nested queries).
 - a. First two nested queries are similar to question 2 (gives highest fraction).
 - b. Outer query (still nested) gets the highest clicks of the referrer-requested pair.
 - c. Outermost query does the fraction.
2. Caveat of using nested queries -- accessing the same file multiple times where Hive is acyclic.

Question 4 Results

Find an example of an English wikipedia article that is relatively more popular in the UK, US, AU.

```
Total MapReduce CPU Time Spent: 1 minutes 25 seconds 610 msec
OK
```

| us_page_title | total_views |
|---------------------------------------------|-------------|
| Main_Page | 879962 |
| Special:Search | 191985 |
| - | 73362 |
| Jeffrey_Toobin | 72949 |
| Kyler_Murray | 61041 |
| Dancing_with_the_Stars_(American_season_29) | 52381 |
| Jeff_Bridges | 49390 |
| The_Haunting_of_Bly_Manor | 43637 |
| Sisters_at_Heart | 43306 |
| Robert_Redford | 41019 |

```
10 rows selected (59.352 seconds)
0: jdbc:hive2://>
```

```
Total MapReduce CPU Time Spent: 1 minutes 57 seconds 980 msec
OK
```

| uk_page_title | total_views |
|---------------------------|-------------|
| Main_Page | 1106769 |
| Special:Search | 293458 |
| - | 106036 |
| Jeffrey_Toobin | 53631 |
| Petr_Čech | 41539 |
| The_Haunting_of_Bly_Manor | 38822 |
| Three_Red_Banners | 32563 |
| Axel_Tuanzebe | 30805 |
| Chicago_Seven | 29130 |
| Gillian_Taylforth | 27182 |

```
10 rows selected (63.664 seconds)
0: jdbc:hive2://>
```

```
Total MapReduce CPU Time Spent: 1 minutes 49 seconds 280 msec
OK
```

| au_page_title | total_views |
|----------------------------|-------------|
| Main_Page | 1016271 |
| Special:Search | 260006 |
| - | 95976 |
| Jeffrey_Toobin | 42548 |
| F5_Networks | 42427 |
| Robert_Redford | 34416 |
| Jeff_Bridges | 30253 |
| Bible | 26585 |
| Murder_of_Robert_McCartney | 22348 |
| The_Haunting_of_Bly_Manor | 21676 |

```
10 rows selected (57.544 seconds)
0: jdbc:hive2://>
```


Question 4

Assumptions:

- Using Internet Rush Hours (7-11 PM UTC):
 - **UK** -- same time, **US (West-Daylight)** -- 2-6 AM UTC, **AU (West)** -- 11 AM-3PM UTC
- October 20, 2020 page views data.
- Number of views is a combination of all countries over the world so our total views aren't accurate.

Steps:

- Created 3 different tables to contain Internet rush hour for en/en.m articles for US, UK, AU.
 - Data loaded into tables are based on UTC time.
- Using the same query as question 1 on all 3 tables.

Question 5 Result

Analyze how many users will see the average vandalized wikipedia page before the offending edit is reversed.

- Average vandalized wikipedia page for October 20, 2020: Enrique Iglesias

```
2020-11-05 12:35:58,413 Stage-2 map = 0%, reduce = 0%
2020-11-05 12:36:01,466 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 1.01 sec
2020-11-05 12:36:05,533 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 2.25 sec
MapReduce Total cumulative CPU time: 2 seconds 250 msec
Ended Job = job_1604596945021_0026
MapReduce Jobs Launched:
Stage-Stage-1: Map: 4 Reduce: 4 Cumulative CPU: 39.0 sec HDFS Read: 880804709 HDFS Write: 457 SUCCESS
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 2.25 sec HDFS Read: 9222 HDFS Write: 120 SUCCESS
Total MapReduce CPU Time Spent: 41 seconds 250 msec
OK
+-----+-----+
| page_title | users |
+-----+-----+
| Enrique_Iglesias | 791 |
+-----+-----+
1 row selected (40.278 seconds)
0: jdbc:hive2://> |
```

Question 5

Assumptions:

- Assuming an En wikipedia page got vandalized when a revision was reverted.
- Down size the data searching to just average vandalized page on October 20, 2020.
- The number of users is a rough estimated number.

Steps:

1. Create a revision table of 70 fields
2. A query to calculate the average revision counts for a page in October 20, 2020

```
Total MapReduce CPU Time Spent: 1 minutes 13 seconds 590 msec
OK
+-----+
| average_page_revision |
+-----+
| 9202.5                  |
+-----+
1 row selected (31.454 seconds)
0: jdbc:hive2://>
```

Question 5

3. A query to find the page title that has the matching revision count to the average count that we found.

```
Total MapReduce CPU Time Spent: 1 minutes 2 seconds 360 msec
OK
+-----+-----+
| page_title | page_revision_count |
+-----+-----+
| DMacks     | 9203                 |
| Enrique_Iglesias | 9203                 |
+-----+-----+
2 rows selected (25.337 seconds)
0: jdbc:hive2://> |
```

4. A query to find all the revisions created for the page and pick the creation timestamp to match page view cutoff. (14076 seconds ~ 4 hours)

```
Total MapReduce CPU Time Spent: 1 minutes 10 seconds 160 msec
OK
+-----+-----+-----+
| page_title | event_timestamp | revision_seconds_to_identity_revert |
+-----+-----+-----+
| Enrique_Iglesias | 2020-10-20 10:24:16.0 | 15188 |
| Enrique_Iglesias | 2020-10-20 10:35:00.0 | 14544 |
| Enrique_Iglesias | 2020-10-20 10:42:48.0 | 14076 |
| Enrique_Iglesias | 2020-10-20 10:53:09.0 | 13455 |
| Enrique_Iglesias | 2020-10-20 10:53:46.0 | 13418 |
+-----+-----+-----+
5 rows selected (31.688 seconds)
0: jdbc:hive2://> |
```

[pageviews-20201020-100000.gz](#)
[pageviews-20201020-110000.gz](#)
[pageviews-20201020-120000.gz](#)
[pageviews-20201020-130000.gz](#)
[pageviews-20201020-140000.gz](#)

[20-Oct-2020 10:51](#)
20-Oct-2020 11:58
20-Oct-2020 12:50
20-Oct-2020 13:55
20-Oct-2020 14:56

5. Final query (similar to question 1 & 4) to look for the article and its total users/views

Question 6 Result (MapReduce)

How many countries edited the enwiki articles in October 2020?

- 134/195 countries with editors for enwiki
- Geoeditor monthly data of October 2020

```
CPU time spent (ms)=1050
Physical memory (bytes) snapshot=596271104
Virtual memory (bytes) snapshot=5123514368
Total committed heap usage (bytes)=560463872
Peak Map Physical memory (bytes)=361136128
Peak Map Virtual memory (bytes)=2559053824
Peak Reduce Physical memory (bytes)=235134976
Peak Reduce Virtual memory (bytes)=2564460544
Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0
File Input Format Counters
  Bytes Read=102875
File Output Format Counters
  Bytes Written=11
quanvu@QUAN-VU:~/hadoop-3.2.1$ hdfs dfs -head '/user/quanvu/output/part-r-00000'
2020-11-05 21:48:05,318 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localhostTrusted = false, remoteHostTrusted = false
enwiki 134
```

GitHub Link

<https://github.com/QuanAVu/Project1-WikipediaAnalysis>